Modeling of Bio-inspired Pattern Recognition: from V1 Models to Sensorimotor Representations

Konrad Christoph Gadzicki

Kumulative Dissertation zur Erlangung des Doktorgrades der Ingenieurwissenschaften
– Dr. Ing. –
Universität Bremen – Fachbereich 3 Mathematik und Informatik

Modeling of Bio-inspired Pattern Recognition: from V1 Models to Sensorimotor Representations

Konrad Christoph Gadzicki

Date of Submission: April 27, 2021
Date of Defense: June 1, 2021

Reviewers
Prof. Dr. Kerstin Schill, University of Bremen, Germany
Prof. Dr.-Ing. Tanja Schultz, University of Bremen, Germany

Acknowledgments

I would like to thank Sinaida, my wife, who had to endure some while I have finished this thesis. I would like to thank Kerstin and Christoph who have supported me through all the years.

Abstract

The human visual system served as inspiration to many approaches in computer vision. The ability of the human visual system to spot, track and recognize objects in still and moving pictures is yet unmatched by computer systems. Pattern recognition is one of the major fields of application in computer vision. This thesis sheds light on three different perspectives how biologically inspired pattern recognition can be realized.

The first is a direct transfer of properties of neurons in the primary visual cortex (V1) to computational models. To set the ground for models predicting responses of V1 neurons, these properties and known effects will be introduced briefly. The individual components contributing to a human visual system (HVS) model will be considered next. Such models are suited for low level pattern recognition, e.g., gratings or primitive shapes. It will be shown how such a model can predict the perception of streak distortions and their subjective assessment by human observers. Streak distortion are generated by offset printing machines almost inevitably. The purpose of this model is the evaluation of offset printing machines with regard to the occurrence of this class of distortions. Furthermore, the computations of the frequency distribution as well as the auto- and cross-correlation functions in a neurobiologically plausible way via the gain control function attributed to neurons in the visual pathway will also be presented.

The second perspective is given by neural networks. This is a connectionist approach of realizing computational models with simple but highly connected units. Neural networks, inspired by the structure of the brain, represent a very simplified view on the brain. For pattern recognition they have shown to be perform extremely well, especially in the form of convolutional neural networks (CNN). They are suited for complex recognition tasks, e.g., object, scene or activity recognition. The approaches covered in this thesis are multimodal convolutional neural networks for human activity recognition (HAR). The focus will be on the comparison of the fusion strategies, namely early and late fusion. The human activity recognition model is used for automatic annotation of data recorded from humans. It serves as part of a pipeline which aims at transferring human knowledge about everyday activities to a robot.

The third perspective is covered by processing of sensorimotor representations. Those are pairs of sensory information and motor actions. In the vision domain saccadic eye movements are an example for sensorimotor processing, but it can be applied to all modalities of human sensory perception. The computational models covered in this thesis are applied to visual perception and localization in a spatial environment. The inference mechanism to guide the sensorimotor process is based on information gain, i.e., on selecting actions to minimize the uncertainty regarding the current belief state of the environment. Furthermore the clustering

of sensorimotor features for the generation of hierarchical knowledge-base, which is utilized by the inference process, will be covered as well. Active perception is closely related to sensorimotor processing. Actions are selected and executed which are believed to maximize information gain through sensory input. These principles can be applied to complex systems like spacecrafts. Navigation approaches based on active perception can be used on transfer orbits through the solar system but also for localization and mapping in the proximity of small solar system bodies, e.g., asteroids.

Zusammenfassung

Das menschliche Sehsystem diente als Inspiration für viele Ansätze des maschinellen Sehens. Die Fähigkeit des menschlichen Sehsystems, Objekte in unbewegten und bewegten Bildern zu erkennen, zu verfolgen und wiederzuerkennen, wird von Computersystemen bisher nicht erreicht. Die Mustererkennung ist eines der Hauptanwendungsgebiete des maschinellen Sehens. Diese Arbeit beleuchtet drei verschiedene Perspektiven, wie biologisch inspirierte Mustererkennung realisiert werden kann.

Die erste ist eine direkte Übertragung von Eigenschaften von Neuronen im primären visuellen Cortex (V1) auf Computermodelle. Um die Grundlage für Modelle zu schaffen, die Reaktionen von V1-Neuronen vorhersagen, werden diese Eigenschaften und bekannten Effekte kurz erläutert. Die einzelnen Komponenten, die zu einem Modell des menschlichen visuellen Systems (HVS) beitragen, werden als nächstes dargestellt. Solche Modelle eignen sich für Mustererkennung auf low-level Ebene, z.B. Gittermuster oder primitive Formen. Es wird gezeigt, wie ein solches Modell die Wahrnehmung von Streifenfehlern und deren subjektive Bewertung durch menschliche Beobachter vorhersagen kann. Die Anwendung für dieses Modell ist die Bewertung von Offsetdruckmaschinen, die diese Klasse von Fehlern zwangsläufig mit einer bestimmten Stärke erzeugen. Weiterhin soll gezeigt werden, wie eine Häufigkeitsverteilung, Auto- und Kreuzkorrelationsfunktionen auf neurobiologisch plausible Weise berechnet werden können, indem Gain Control Funktionen verwendet werden, welche Neurone im visuellen System zugeschrieben werden.

Die zweite Perspektive wird durch neuronale Netze gegeben. Hierbei handelt es sich um einen konnektionistischen Ansatz zur Realisierung von Computermodellen mit einfachen, aber hochgradig verbundenen Einheiten. Neuronale Netze, inspiriert von der Struktur des Gehirns, stellen eine sehr vereinfachte Sicht auf das Gehirn dar. Für die Mustererkennung haben sie sich als äußerst leistungsfähig erwiesen, insbesondere in Form von neuronalen Netzen mit Faltung (CNN). Sie eignen sich für komplexe Erkennungsaufgaben, z.B. Objekt-, Szenen- oder Aktivitätserkennung. Die in dieser Arbeit behandelten Ansätze sind multimodale CNNs für menschliche Aktivitätenerkennung (HAR). Der Schwerpunkt liegt dabei auf dem Vergleich von Fusionsstrategien, nämlich der frühen und der späten Fusion. Als Anwendung wird das Modell zur menschlichen Aktivitätenerkennung verwendet, um von Menschen aufgenommene Daten automatisch zu annotieren. Es dient als Teil einer Pipeline, die darauf abzielt, menschliches Wissen über Alltagstätigkeiten auf einen Roboter zu übertragen.

Die dritte Perspektive wird durch die Verarbeitung von sensomotorischen Repräsentationen abgedeckt. Diese sind Paare von sensorischen Informationen und motorischen Aktionen. Im Bereich des Sehens sind sakkadische Augenbewegungen ein Beispiel für die sensomotorische Verarbeitung, aber sie kann auf alle Modalitäten der menschlichen Sinneswahrnehmung ange-

wendet werden. Die in dieser Arbeit behandelten Berechnungsmodelle werden auf die visuelle Wahrnehmung, sowie Lokalisierung in einer räumlichen Umgebung angewendet. Der Inferenzmechanismus zur Steuerung des sensomotorischen Prozesses basiert auf Informationsgewinn, d.h. auf der Auswahl von Aktionen zur Minimierung der Unsicherheit bezüglich des aktuellen Glaubenszustands der Umgebung. Weiterhin wird das Clustering von sensomotorischen Merkmalen zur Generierung einer hierarchischen Wissensbasis, die vom Inferenzprozess genutzt wird, behandelt. Die aktive Wahrnehmung ist eng mit der sensomotorischen Verarbeitung verbunden. Es werden Aktionen ausgewählt und ausgeführt, von denen angenommen wird, dass sie den höchsten Informationsgewinn durch sensorischen Input liefern. Diese Prinzipien lassen sich auf komplexe Systeme wie Raumschiffe anwenden. Auf aktiver Wahrnehmung basierende Navigationsansätze können auf Transferorbits durch das Sonnensystem, aber auch zur Lokalisierung und Kartierung in der Nähe von Kleinkörpern, z.B. Asteroiden, eingesetzt werden.

Contents

List of Figures

List of Acronyms

ANN Artificial Neural Network

CIE Commission internationale de l'éclairage (International Commission on Illumination)

CFS Contrast Sensitivity Function

CNN Convolutional Neural Network

DoG Difference of Gaussians

DNN Deep Neural Networks

DS Dempster-Shafer (Theory)

EASE Everyday Activity Science and Engineering

HAR Human Activity Recognition

HRI Human-Robot-Interaction

HVS Human Visual System

i2D intrinsic 2-Dimensional (Feature)

I3D Inflated 3D (CNN)

IBIG Inference by Information Gain

iFV Improved Fisher-Vector

ITU International Telecommunication Union

MSE Mean Squared Error

NEEM Narrative-enabled Episodic Memories

NN Neural Networks

PCA Principal Component Analysis

PSNR Peak Signal-to-Noise Ratio

ReLU Rectified Linear Unit

RNN Recurrent Neural Network

RoG Ratio of Gaussians

SGD Stochastic Gradient Descent

SIFT Scale-Invariant Feature Transform

SLAM Simultaneous Localization and Mapping

SMC Sensorimotor Contingency

SMF Sensorimotor Feature

SOM Self-Organizing Map

STIP Space-Time Interest Points

SURF Speeded Up Robust Features

SVM Support-Vector Machine

TCN Temporal Convolutional Network

V1 Primary Visual Cortex

VDMA Verband Deutscher Maschinen- und Anlagenbau (Mechanical and Plant Engineering Association)

VLAD Vector of Locally Aggregated Descriptors

VR Virtual Reality

CHAPTER 1

Introduction

Vision is the primary sense for humans which provides rich information about our environment. It enables us to perceive, navigate within and interact with it. A significant amount of information can be obtained through the visual system alone, enabling us to carry out most tasks solely with this information. While we move through our environment we can spot obstacles, recognize objects of interest and perceive their location, if they are in motion or not, and the direction and speed they are moving in, so that we can avoid or interact with them.

Of course our other senses also contribute significantly in our daily life, providing complimentary information to the visual input, e.g. a sound can shift the focus of our attention or can make the task of recognizing an object easier. The interaction between all our senses enables us to perceive our world to the fullest extent and ultimately survive in it for some millions of years already.

This work covers computational models of the human visual system, both as a stand-alone model and part of a multi-sensory system. Furthermore, the incorporation of high-level cognitive principles through information theory is covered as well.

The human brain is a extraordinary complex organ which dedicates a large amount of its volume to the processing of sensory information. The visual cortex located in the occipital lobe at the back of your head is responsible for the processing the visual component. Its first part, the primary visual cortex (V1), serves (among other functionalities) as a feature extractor, recognizing primitive patterns like edges, corners or gratings (Oram and Perrett, 1994). Low level vision can be represented by V1 models, relating the strengths of response with certain patterns. Questions whether a specific pattern might be visible to humans can be answered with the help of such models.

Complex visual tasks like object or activity recognition require much more sophisticated models. The classical approach from machine learning is to use hand-crafted features and map them through learning algorithms to the solution. While these approaches dominated in the past they have been surpasses by deep learning methods in the last decade (Krizhevsky, Sutskever, and Hinton, 2012). In the image domain convolutional neural networks have established themselves as the state-of-the-art approach to solve a large variety of problems. They were successfully applied to other domains like audio (Chandrakala and Jayalakshmi, 2019), radar (Lang et al., 2020) or volumetric data (Qi et al., 2016) as well.

While computational models of the visual system and convolutional neural networks are bottom-up approaches, perception can also be modeled by incorporating top-down approaches,

using higher level cognitive principles. Human perception of its environment is performed through saccadic exploration, directing the fovea to locations of interest and forming a neural image through repeated fixations. This behavior is an example of being guided by higher level cognitive processes and can be modeled by information theoretic approaches.

1.1 Outline and Contribution

This cumulative dissertation consists of three major parts, all under the umbrella of bio-inspired pattern recognition and its applications. Each part sheds a different light on the question of how biological principles can be utilized in computer-science problems.

- Chapter 2 "Human Visual System Models" stays close to the biological foundations of the human visual system (HVS) and the properties of neurons. We have developed a model of the HVS according to the state-of-the-art visual models and have applied it to detection and prediction of subjective assessment of streak distortions in printings produced by offset printing machines. This is a direct application of a model to a problem which is of practical interest to the printing industry. The evaluation of a printing machine among other tests involves a test for streak distortions. The automated testing procedures currently in place show subpar results which might be improved with our system. HVS models are commonly used to approximate the activity of the primary visual cortex (V1), usually focusing on foveal perception. Peripheral vision and models of V2 have recently been investigated by exploiting statistical information about a scene. In this chapter we also show how a simple ensemble of neurons with properties known from neurobiology is able to estimate a frequency distribution and could bridge the gap between statistical information used in models and a plausible neurological mechanism to compute them.

- Chapter 3 "Activity Recognition with Convolutional Neural Networks" introduces neural networks and their most important features, where the focus is on activity recognition using convolutional neural networks. The functioning of the brain itself is the inspiration for neural networks with simple units of limited processing power being able to solve complex problems in a collectivist approach by forming a large, highly connected network. We have investigated the design of deep neural networks regarding the optimal depth of a particular network architecture for skeleton-based activity recognition. Furthermore, we have implemented a multimodal convolutional neural network for activity recognition which sheds light on the question which fusion strategy is the best approach. Such a system can contribute to a pipeline which enables the transfer of human activity data to robots.

- Chapter 4 "Sensorimotor Perception and Navigation" features our work on sensorimotor visual perception and navigation which uses information theoretic approaches. These approaches are inspired by the way saccadic eye-movement guided by high-level cognitive processes, featuring a high-level view on visual pattern recognition. We summarize our work on clustering of sensorimotor features (SMF) and its integration into an agent system. This chapter concludes with the transfer of information theoretic and cognitive approaches to autonomous navigation of a spacecraft.

Chapter 5 "Conclusion and Outlook" summarizes all contributions and gives an outlook on future research opportunities. A list of publications by the author is given afterwards. The accumulated publications can be found in the appendix.

CHAPTER 2

Human Visual System Models

Computational models of the human visual system (HVS) are close to the biological foundation known from neurobiology. They generally aim at predicting the perception of patterns as measured in psychophysical experiments from human subjects.

In this chapter the author summarizes the most important properties of the human visual system (Section 2.1) which are reflected in the standard model of HVS (Section 2.2). We used the HVS model for prediction of the perceived severity of streak distortions in printings as summarized in Section 2.3. Section 2.4 covers the estimation of a frequency distribution with a neural representation. The means to achieve this stem from the properties known from V1 simple cells.

2.1 The Human Visual System

The retina forms the first layer of the visual system. It holds the photoreceptors, but also a thin layer of neural cells which already process the signal from the receptors. These neural cells in retina are the same cells as the brain making them a part of it. The fovea is the spot on the retina with the highest visual acuity (Palmer, 1999, pp. 28–34). It has a diameter of approx. 0.3 millimeters and covers approx. 1° of our visual field. In comparison to the entire field of view of approx. 140° this is very narrow but still this spot is responsible for the perception of all the details in our environment (Tschulakow et al., 2018).

In vision science the concept of the receptive field relates the area of the retina where light stimuli have been detected to the corresponding neural responses (Hartline, 1938). This general concept of describing a neuron has been extended to other sensory neurons as well (Wandell, 1995, pp. 128–135).

The classical receptive field of a ganglion cell is described by a center-surround structure where the center area and the area surrounding it are stimulated oppositely by light. Ganglion cells can be divided into two classes according to this structure (Wandell, 1995, pp. 128–135):

- on-center, off-surround cell where the center area is excitatory and the surround inhibitory to the cell's response,
- and off-center, on-surround which act the other way round.

This is achieved through lateral inhibition (Hartline, Wagner, and Ratliff, 1956) which influences the neighboring photoreceptor's responses (Palmer, 1999, p. 147).

The response of neurons in the visual pathway usually depends on contrast and not on absolute light intensities. Luminance ranges across six orders of magnitude in our daily life between evening and bright daylight, while the common neuron can encode only two to three orders of magnitude making it difficult for biological systems to respond to absolute intensities in such a huge range (Wandell, 1995, p. 117). In general contrast can be defined as the ratio between the intensity change and the average intensity. The average intensity can be calculated across an image or image patch or the area of the receptive field in the case of a neuron (Wandell, 1995, pp. 146–148).

Neurons are typically responding differently depending on the spatial frequency of a pattern. The contrast sensitivity function summarizes the response of a neuron at different spatial frequencies and further provides information about the neuron's receptive field. The concept of contrast sensitivity is not only applied to ganglion cells, but to all neurons in the visual pathway, primarily to those in the early areas where neurons respond to primitive shapes, e.g., bars or edges. Depending on the size of a pattern on the retina, a neuron will respond differently, making it a function of spatial frequency (Wandell, 1995, pp. 135–137). Apart from individual neurons the overall visual function can be described by the contrast sensitivity function (CSF) (Campbell and Robson, 1968). The CSF can be seen as an envelope for a number of different contrast sensitivity functions of parallel channels in the visual system (Robson, 1993).

The primary visual cortex, i.e., Visual Cortex 1 or short V1, receives the neural connections from the retina through the lateral geniculate nucleus (LGN) (Palmer, 1999, pp. 148–151). Hubel and Wiesel have received the Nobel Price for the measurement of receptive fields of neurons in V1, categorized as simple and complex cells (Hubel and Wiesel, 1959; Hubel and Wiesel, 1962; Hubel and Wiesel, 1968). These cells have an oriented receptive field, thus responding better to patterns of a specific orientation. The patterns which simplex and complex cells respond to can be described as bars and edges. The receptive fields are built respectively from adjacent excitatory and inhibitory regions which are longer in one direction than in the other, forming a main axis which defines their preferred orientation (Wandell, 1995, pp. 135–137).

Simple cells are usually selective to patterns with a specific phase (Palmer, 1999, pp. 151–153). For instance, inverting the intensities but keeping the contrast will not trigger a response from a simple cell. Complex cells on the other hand are generally invariant to the polarity of a contrast. They show similar selectivity for orientation and contrast as simple cells, but respond to a pattern and its inverted version equally strong as shown for contrast reversing patterns (De Valois, Albrecht, and Thorell, 1982). Similarly a sinusoidal pattern drifting over the receptive field of a simple cell will trigger a clear peak where the phase of the signal matches the preferred selectivity with regard to polarity and orientation whereas a complex cell will respond on a near constant level (De Valois, Albrecht, and Thorell, 1982).

Figure 2.1 shows the responses of neurons in area V1 as measured by Albrecht and Hamilton, 1982b. The neural response curves show saturation based on the spatial frequency of the pattern. This effect was explained by a mechanism called gain control where the pooled signal for neighboring neurons suppresses the response of a neuron (DeAngelis et al., 1992; Geisler and Albrecht, 1992; Heeger, 1992). A model for this relation is given in Heeger, 1992.

Masking describes the interaction of two or more patterns which are presented simultaneously (Wandell, 1995, pp. 219–221). One pattern serves as the test pattern for which the detection threshold elevation is to be measured. This is the difference between the threshold for the masking pattern alone versus the test and masking pattern together. Test pattern and mask are added, and the resulting sum of the patterns' signals is presented to the observer. When test and mask cannot be discriminated from the mask alone even though the test pattern alone is visible, then we can say that the mask is actually masking the test pattern. But there can be also the case that the test + mask can be discriminated from the mask alone while the test pattern alone is not visible. This case where the mask actually helps to

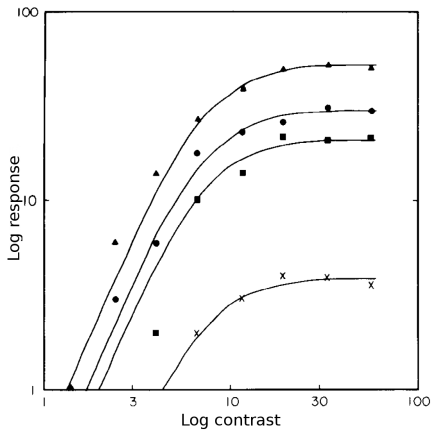


Figure 2.1: Contrast response of a neuron in V1 (adapted from Albrecht and Hamilton, 1982b)

detect the test pattern is called facilitation (Wandell, 1995, pp. 219–221).

The effects seen in contrast normalization and masking generally led to the belief that internal representation is organized in a multi-channel fashion (Ginsburg, 1986; Wiesel and Hubel, 1963; Hubel and Wiesel, 1977).

2.2 Models of V1

A computational HVS model is generally designed in a way that the properties and effects described above can be replicated. The scientific community has arrived at the stage where a standard model can be formulated. It features

- a local contrast stage where absolute luminance values are abstracted from,
- a multi-channel decomposition into spatial frequency scales (fine to coarse images) and over orientation channels,
- an application of a contrast sensitivity function, and
- a pooling stage accounting for masking and contrast normalization effects (Nadenau et al., 2000).

2.2.1 Contrast Sensitivity

The contrast sensitivity function is a central element in a model of human vision. The CSF can be measured for individual detection mechanisms in the visual system, but also globally for the visual system, serving as an envelope for various CSFs (Robson, 1993). From this view the contrast sensitivity can be applied in two ways. The first is to apply it globally and the second way is to tune the individual filters to represent the CSF through the set of filters.

Global tuning is generally achieved with a filter function in the frequency domain (Daly, 1993; Watson and Solomon, 1997). The filter function is applied to frequency domain representation of the input signal, adjusting the amplitudes of spatial frequencies. Since the CSF is a function of spatial frequency and estimated from psychophysical experiments, the design of the filter function must be based on certain assumptions about viewing distance and size of the image in physical dimensions. If there is an experimental setup for the model in development, it can usually provide these pieces of information (e.g., size of the display, viewing distance of the observer).

Tuning of individual filters can be approached in a similar way by taking the values of the CSF function in the points corresponding to the center frequencies of the individual filters. If there is data available from psychophysical experiments, the individual filter channel can be tuned by fitting the filter parameters to the experimental data.

2.2.2 Contrast Computation

Often the contrast is known beforehand, especially if it is a parameter in an experimental setup as is the case in psycho-physical vision experiments. Here the target luminance is calculated from the target contrast and can be reversed for processing of images resulting from this procedure. An example can be found in Watson and Ahumada Jr., 2005 who computed the display luminance according to

$$L(l) = L_0 \left[1 + \frac{c}{127}(l - 128)\right] \tag{2.1}$$

with L being the display luminance, L_0 the mean luminance, l the gray-level of the pixel and c the contrast. In such a case the contrast can be computed by rearranging Equation 2.1.

To compute the contrast from an arbitrary image's luminance function, there are several ways depending on whether global or local contrasts are required. If contrast is to be computed on a global level, Weber contrast (Wandell, 1995, pp. 147-148), Michelson contrast (Michelson, 1927) or King-Smith & Kulikowski contrast (King-Smith and Kulikowski, 1975) can be computed directly.

For local contrast it becomes more difficult since the contrast has to account for the local image information which can be complex in natural images (Peli, 1997). The "Difference of Gaussians" (DOG) is the classical function for the description of the sensitivity of retinal ganglion cells (Enroth-Cugell and Robson, 1966; Enroth-Cugell, Robson, et al., 1983; Rodieck, 1965). It uses two Gaussian functions with different cut-off frequencies, so that the lower frequency filter is subtracted from the higher one, resulting in a linear band-pass output (Eq. 2.2). The 2-dimensional function is defined as

$$g_{i+1}^{DoG}(x,y) = l(x,y) * \left(\frac{1}{2\pi\sigma_i^2} \exp\left(\frac{x^2+y^2}{2\sigma_i^2}\right) - \alpha \frac{1}{2\pi\sigma_{i+1}^2} \exp\left(\frac{x^2+y^2}{2\sigma_{i+1}^2}\right) \right), \tag{2.2}$$

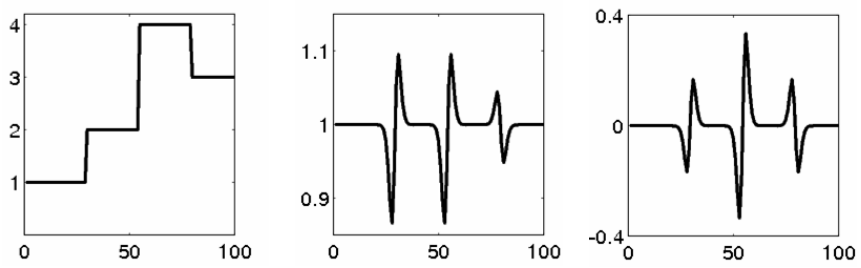
with $l(x,y)$ denoting the luminance function at a position in the image, $*$ being a convolution, α a gain factor, and σ_i and σ_{i+1} the variances of the Gaussian windows (Gadzicki and Zetzsche, 2013). This notation stems from the pyramid representation where $i+1$ denotes the lower frequency channel.

Another operator for calculation of contrast is the "Ration of Gaussian" (RoG) operator (Zetzsche and Hauske, 1989a). It is a divisive operation of two low-pass inputs with different cut-off frequencies resulting in a non-linear band-pass output (Eq. 2.3).

$$g_{i+1}^{RoG}(x,y) = \frac{l(x,y) * \frac{1}{2\pi\sigma_i^2} \exp\left(-\left(\frac{x^2+y^2}{2\sigma_i^2}\right)\right)}{\left(l(x,y) * \frac{1}{2\pi\sigma_{i+1}^2} \exp\left(-\left(\frac{x^2+y^2}{2\sigma_{i+1}^2}\right)\right)\right) + \tau} \tag{2.3}$$

with the same notation as in 2.2 and τ as a constant for avoiding division by zero.

Figure 2.2 illustrates the response of the operators. The RoG operator has the advantage to transform the relative increase in luminance directly (first and second edge in Figure 2.2a) and transform it into contrast of equal levels (Figure 2.2b). The DoG operator is linear and outputs different contrast at these positions (Figure 2.2c), requiring the luminance signal to be represented on a logarithmic scale.



(a) (b) (c)

Figure 2.2: Response of a Ratio of Gaussian (RoG) operator to luminance step edges. (a) Input. (b) The RoG output represents luminance contrast. (c) Linear Difference of Gaussians (DoG) response shown for comparison (Source: Zetzsche and Hauske, 1989a).

The RoG is similar to the local band-limited contrast as suggested by Peli, 1990. It is defined as

$$g^{PELI}(x,y) = \frac{a(x,y)}{b(x,y)} \tag{2.4}$$

with $a(x,y)$ being a band-limited filter (a radically symmetric band-pass filter) output of the image and $b(x,y)$ a low-pass filter output.

2.2.3 Multi-Channel Decomposition

The multi-channel decomposition involves a scale-space representation where features are extracted from high and low spatial frequencies separately.

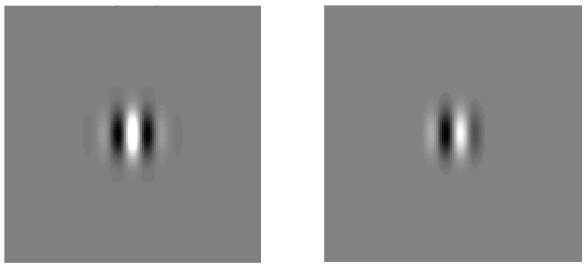
2.2.3.1 Pyramid Representation

The image pyramid is a memory efficient way to represent multi-resolution images and is commonly used in computer vision. In this scheme an image is filtered with an appropriate low-pass and sub-sampled usually by a factor of 2 in every direction (Wandell, 1995, pp. 262–267). The resulting image has $\frac{1}{4}$ size of the original image, but still contains all the necessary information with regard to $\frac{1}{2}$ Nyquist frequency[1] of the original image. Any filters which operate within this limit can be applied to the low-passed filtered sub-sampled channel without any loss of information. This procedure of low-pass filtering and sub-sampling can be repeated multiple times, resulting in a pyramid-like representation. The low-pass filter is usually a Gaussian low-pass, hence titled Gaussian pyramid. Burt and Adelson, 1983 introduced this scheme, but actually aimed to create a difference image for every scale. The low-pass filtered image was subtracted from the original image for every scale, resulting in a high-pass filtered image. This version of the pyramid scheme is known as Laplacian pyramid.

2.2.3.2 Scale and Orientation Selective Filters

The image pyramid provides the base frame for the application of spatial-frequency and orientation selective filters. The Gabor function (Gabor, 1946) is generally considered as a good fit to the selectivity of simple cells (Marcelja, 1980; Kulikowski, Marčelja, and Bishop, 1982; Daugman, 1984). Hawken and Parker, 1987 suggested that the DoG-S function (the Gaussian for the surround is split into two Gaussians separated spatially by $+S$ and $-S$

[1] The Nyquist frequency is the highest alias-free frequency resulting from a discrete sampling process of a (continuous) signal. It is $\frac{1}{2}$ of the sampling rate.



(a) (b)

Figure 2.3: Gabor filter outputs for an orientation of 0°: (a) even and (b) odd symmetric parts.

from the center Gaussian) is a better fit, but the differences are marginal. The 2-dimensional Gabor function (Eq. 2.5) in the spatial domain consists of a complex sinusoidal function (Eq. 2.6) weighted by a Gaussian function (Eq. 2.7)

$$h(x,y) = g(x,y)s(x,y) \tag{2.5}$$

where

$$s(x,y) = \exp(-j2\pi u_0 x) \tag{2.6}$$

$$g(x,y) = \frac{1}{2\pi\sigma_x\sigma_y}\exp\left(-\pi\left(\frac{x^2}{2\sigma_x^2}+\frac{y^2}{2\sigma_y^2}\right)\right), \tag{2.7}$$

with σ_x, σ_y being the variances of the 2D Gaussian function and u_0 the wavelength of the complex sinusoidal function. The orientation can be modeled by a rotation of the coordinate system, thus rotating the desired orientation onto the x-axis as the function itself has an orientation of 0° (Gadzicki and Zetzsche, 2013). The outputs of a Gabor filter are shown in Figure 2.3. The complex sinusoidal function contains two symmetries, the cosine, even symmetric part (Figure 2.3b) and the sine, odd symmetric part (Figure 2.3a)

The Gabor function provides a kernel for both, even and odd parts, which can be convolved with an image. Rather than operating in the spatial domain, often a filter approach in the frequency domain is taken. Here the Gabor filter function can be modeled directly by shifting the Gaussian function to the location u, v which represent the center frequency resulting in the filter function (Eq. 2.8)

$$H(u,v) = \exp\left(-\pi\frac{(u-u_0)^2}{(2\sigma_u)^2}+\pi\frac{v^2}{(2\sigma_v)^2}\right) \tag{2.8}$$

with u_0 being the center frequency and σ_u, σ_v defining the bandwidths of the filter. The orientation is again modeled by a rotation of the coordinate system (Gadzicki and Zetzsche, 2013). The parameters are given in polar notation by a center frequency and an orientation angle.

Field, 1987 argues that the Gabor function is only optimal in terms of the spread of uncertainty in space and spatial frequency in the Cartesian coordinate system. He suggests a Log-Gabor function (Eq. 2.9) which is proposed to be better suited for a representation in polar notation, minimizing the spread and overlap between individual filters. It is given by

$$H_{log}(\rho,\phi) = j^k \exp\left(-\frac{(\log\frac{\rho}{\rho_0})^2}{2(\log\sigma_\rho)^2}\right)\left[-\exp\left(\frac{(\phi-\phi_0)^2}{2\sigma_\phi^2}\right)-(-1^k)\exp\left(\frac{(\phi-\phi_0-\pi)^2}{2\sigma_\phi^2}\right)\right] \tag{2.9}$$

with ρ being the radial frequency, ϕ the angle, ρ_0 the center frequency and ϕ_0 the orientation angle. σ_ρ and σ_ϕ define the frequency and orientation bandwidth of the filter.

Another popular decomposition is the steerable pyramid (Simoncelli, W. T. Freeman, et al., 1992; Simoncelli and W. T. Freeman, 1995). This transformation is a multi-scale, multi-orientation image decomposition which is translation-invariant and rotation-invariant, addressing a drawback of orthogonal wavelet transforms which are not translation-invariant. It has been used in a number of HVS models (Teo and Heeger, 1994; Heeger and Teo, 1995).

Earlier models also used polar separable filters, e.g. the fan filter (Daly, 1993). It consists of a radial and an orientation component, using the cosine function in polar notation.

2.2.4 Normalization and Masking Effects

2.2.4.1 Gain Control

The inhibition from neighboring units can be modeled by pooling over scales, orientations and space, and using the pooled response as a divisive term in the Naka-Rushton function (Albrecht and Hamilton, 1982a) (Eq. 2.10). Pooling over space is performed with a low-pass filter, e.g., a Gaussian filter, given as

$$\bar{r}_{s,o}(x,y) = \frac{r_{s,o}(x,y)^p}{c^q + \sum_{\bar{s}=s-1}^{s+1} \sum_{\bar{o}=1}^{6} (r_{\bar{s},\bar{o}} * k_{\bar{s},\bar{o}})(x,y)^q}. \tag{2.10}$$

The response r of a filter at scale s and orientation o is normalized by the responses the neighboring scales $\bar{s}$ and all orientation $\bar{o}$. k is the low-pass kernel used for spatial pooling within a channel. In addition, c is the point where saturation sets in and q is the exponent.

Such a divisive contrast gain control is a common feature of visual models (Watson and Solomon, 1997; Teo and Heeger, 1994; J. M. Foley, 1994; Heeger, 1992; Legge and J. Foley, 1980).

2.2.4.2 Masking

Masking effects are modeled by adding a second pathway to the model, so that mask and mask+signal are processed separately and compared to each other (Daly, 1993; Watson and Solomon, 1997). For the comparison of these two pathways a norm like the Minkowski distance (Eq. 2.11) has been used in a variety of models (Quick, 1974; Graham, 1977; Graham and Robson, 1987; Watson, 1979), but this is more because of practical reasons than biological insights. The Minkowski distance is given as

$$d = \left(\sum |\bar{r}_{\text{signal+background}} - \bar{r}_{\text{background}}|^p \right)^{1/p} \tag{2.11}$$

Here p is an exponent which is set to 2 in many models, turning it into the Euclidean norm as in our model. In our model the norm was applied per pixel, but in other settings, e.g., prediction of detection thresholds, it is applied globally and summed over space (Watson and Ahumada Jr., 2005).

2.2.5 Model Overview

The overview of our model, used in Gadzicki, 2009; Gadzicki and Zetzsche, 2013, is shown in Figure 2.4 and 2.5.

On the top level the model consists of two pathways accounting for masking effects by processing the background and signal+background separately and computing the difference in the final stage (Figure 2.4). Each pathway consists of a HVS model shown in Figure 2.5.

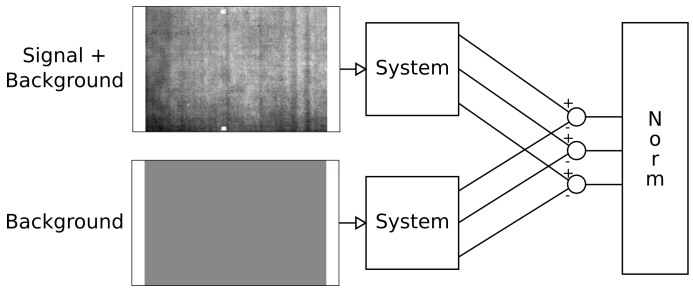


Figure 2.4: Model overview. The system (shown in detail in Figure 2.5) generates a multi-channel representation of both signal and signal+background. The distance between the two representations is assumed to be the perceived strength of the signal, and is computed by the difference norm shown at the right (Adapted from Gadzicki and Zetzsche, 2013).

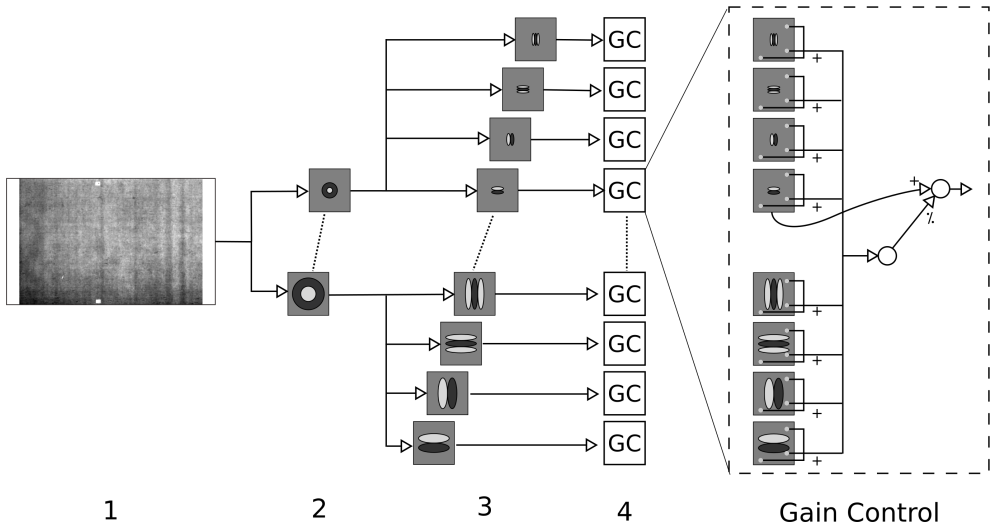


Figure 2.5: Overview over the system (as used for both pathways in Figure 2.4). From the input (1), the contrast is computed by non-linear ROG filters (2) and passed to a set of frequency- and orientation-selective linear filters (3). The outputs are then passed through gain control mechanisms (4). One of these is shown in detail on the right-hand side. Hence each channel is normalized by spatial pooling over the other channels (Adapted from Gadzicki and Zetzsche, 2013).

We use a Gaussian pyramid for decomposition into spatial frequency scales and apply the RoG operator for contrast computation. The scales are further decomposed into orientation channels by Log-Gabor filters (Field, 1987) which are tuned to represent the contrast sensitivity function with the set of filters. The gain control stage pools neighboring responses for normalization.

2.3 Application of a V1 Model to Visual Assessment

2.3.1 Visual Quality Assessment

For technical applications like the transmission of images or video, the relation of image quality and compression level is of major interest. It covers the transmission and display in television, digital media, but also in traditional media like printing. The subjective assessment methods described in 2.3.1.1 deliver the most accurate ratings, but these methods require a lot of time and resources, and might not be feasible for every minor change in a compression algorithm.

Objective methods use an algorithm to determine the image quality. This allows for an automated evaluation, greatly reducing the resources required. Such methods should ideally have a high correlation with the subjective assessments by human subjects.

2.3.1.1 Testing of Subjective Perception

In psychophysical experiments the subjective performance can be generally tested by adjustment or judgment tasks. Subjects in adjustment tasks are given control over the stimulus and are asked to satisfy a given criterion, e.g., adjust the contrast of a pattern to be "just barely detectable", cancel a distortion or match two stimuli. In judgment tasks the subjects are asked to classify a stimulus according to a given criterion, e.g., rate the stimulus on a scale, answer whether a stimulus was present in two-alternative forced-choice or yes-no settings (Pelli and Farell, 1995).

The experiments performed are often aimed at determining a threshold of perception, motivated by the desire to investigate low-level vision, minimizing the effects of cognition. They can be evaluated statistically with the threshold being defined at arbitrary levels of performance. These kind of experiments are well-suited to investigate the perception of simple patterns like bars or sinusoidal gratings (Pelli and Farell, 1995).

For complex patterns as in natural images the threshold approach has limitations. It can still be useful to determine whether any distortion can be perceived at all, but experimenters would to like to judge the effects over a wider range (Nadenau et al., 2000). For compression tasks the association of rate and distortion is crucial for the evaluation of a coding scheme. Here the judgment of overall visual quality is more important.

Standard procedures are for the assessment of visual quality through subjective methods described in *ITU-R BT.500-11* 2002. Even though the International Telecommunication Union (ITU) has developed the methods for television they can be applied to still images as well (Nadenau et al., 2000). Two commonly used methods are Double Stimulus Continuous Quality Scale (DSCQS) and the Double Stimulus Impairment Scale (DSIS) (*ITU-R BT.500-11* 2002).

2.3.1.2 Methods for Measurement of Visual Distortions

The simplest way to assess the distortions introduced with regard to a reference image is to use pixel-based metrics. The metrics are calculated between two images with luminance l_1 and l_2 at positions x, y. Widely used measures are mean squared error (MSE) (Eq. 2.12) or peak signal-to-noise ratio (PSNR) (Eq. 2.13):

$$d_{\texttt{MSE}} = \frac{1}{nm} \sum_{x=1}^{n} \sum_{y=1}^{m} \big(l_1(x,y) - l_2(x,y)\big)^2 \tag{2.12}$$

$$d_{\texttt{PSNR}} = 10 \log \frac{MAX_l^2}{d_{\texttt{MSE}}}. \tag{2.13}$$

MAX is the maximum value of l, e.g., 255 for a 8-bit integer or 1.0 for a float when representing gray-scale or color channel values (Nadenau et al., 2000). Such pixel-based metrics are easy to compute and fast to apply. Unfortunately they have little in common with subjective assessments (Marmolin, 1986; Girod, 1993; Mannos and Sakrison, 1974; Teo and Heeger, 1994; Zhou Wang and Bovik, 2002).

The inadequacy of pixel-based metrics has been realized early and led to the development of more sophisticated methods (Mannos and Sakrison, 1974). There are many approaches which are not based on the human visual system. Feature based scales, e.g., (Miyahara, 1988; Miyahara, Kotani, and Algazi, 1998), apply a simple CSF approximation, weighting errors according to spatial frequencies, and consider particular structural errors. The correlation between these individual features is performed with a principal component analysis (PCA) with the quality metric being a linear combination of principal components.

Methods based on the structural similarity index (SSMI)[2] (Z. Wang, Simoncelli, and Bovik, 2003; Z. Wang, Bovik, Sheikh, et al., 2004; Z. Wang, Bovik, and Simoncelli, 2005) apply a CSF filtering, multi-channel decomposition and error normalization. The SSMI takes local luminance, contrast, and structural composition into account as separate measures, combining them according to their relative importance.

Multi-dimensional impairment scales approaches (Kayargadde and Martens, 1996) analyze the image along multi-dimensional perception axes defined by distortions introduced from blur and noise. The impairment vector spans a space which can be used to assess the quality scale.

X. Zhang and Wandell, 1997 proposed a spatial extension to CIELAB, a uniform color space (*CIE 1976* 1976), which adds filtering of spatial frequencies to CIELABs color separated channels, measuring the quality as the distance ΔE between two images.

The image information fidelity approach (Sheikh and Bovik, 2006) takes inspiration from natural image statistics, and proposes that for the class of natural images the mutual information between two images can be used as a measure of perceptual fidelity. The comparison is calculated over subbands extracted with a wavelet decomposition.

Approaches based on convolutional neural networks have been proposed recently (J. Kim, Nguyen, and Lee, 2019; W. Hou et al., 2015; Y. Li et al., 2015).

2.3.1.3 Image Quality Approaches Based on Human Visual System

Starting with Mannos and Sakrison, 1974 the HVS models for assessment of visual quality have become the norm when performance was considered more important than computational time.

There is a number of image quality models based on the HVS (Zetzsche and Hauske, 1989a; Zetzsche and Hauske, 1989b; Daly, 1993; Lubin, 1993; Lubin, 1995; Teo and Heeger, 1994; Heeger and Teo, 1995; Westen, Lagendijk, and Biemond, 1995; C. C. Taylor et al., 1997) which generally incorporate the most important properties, accounting for luminance invariance, sensitivity to frequencies and orientations, and masking effects (Gadzicki and Zetzsche, 2013). Masking effects have been modeled with a point-wise nonlinearity in earlier models (Daly, 1993; Westen, Lagendijk, and Biemond, 1995) while newer models used divisive inhibition with pooling over channels and spatial positions (Teo and Heeger, 1994; Heeger

[2]The SSMI model has gained wide popularity. Its developers received an Emmy Award in 2015. https://www.emmys.com/news/press-releases/honorees-announced-67th-engineering-emmy-awards

and Teo, 1995). Daly, 1993's Visible Differences Predictor (VDP) has been extended to high dynamic range images (Mantiuk, Myszkowski, and Seidel, 2004).

Larson and Chandler, 2010 argue that the usage of a single strategy for determining the image quality with a HVS is not sufficient. Instead, different strategies for comparing high-quality and low-quality images are suggested.

2.3.2 Technical Applications

HVS models have been used to evaluate the observer's performance depending on the display type (Krupinski et al., 2004). For medical diagnosis, e.g., viewing of mammographic images, it is crucial that the display enables the observer to perform the identification and/or classification task as good as possible. Observer's performance on LCD displays was reported to be superior to CRT displays, and the HVS model was able to predict the performance with high correlation.

2.3.2.1 Prediction of Perceived Distortion of Streak Patterns

> This section summarizes our work published in *"Prediction of the Perceived Quality of Streak Distortions in Offset-Printing with a Psychophysically Motivated Multi-channel Model"* and *"Prediction of the perceived quality of streak distortions in offset-printing with a psychophysically motivated multi-channel model"*.

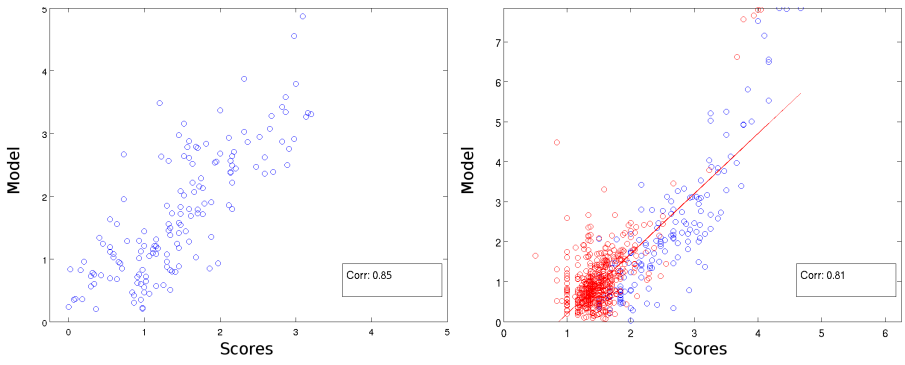
The application in our papers (Gadzicki and Zetzsche, 2012; Gadzicki and Zetzsche, 2013) is the prediction of subjective assessments of streak distortions produced by modern offset printing machines. Visual quality assessment approaches usually rate the entire image, but in our work, we are interested in rating every pixel position.

The goal is to evaluate the printing machine itself according to the distortions it produces. Streak distortions are a typical kind of these distortions, running horizontally across a print, orthogonal to the printing direction. They stem from the printing process due to vibrations of the machine operating at high speeds. These distortions are basically slight shifts of the ink and cannot be avoided entirely, making is necessary to evaluate printing machine regarding how strongly they produce these distortions.

The printing machines are highly sophisticated and expansive, being priced at several millions Euros. Since a faulty or misconfigured machine leads to a significant discount on the price, an evaluation can result in disagreeing options from manufacturer's and customer's side and ultimately in lawsuits. The evaluation task in subjective manner is resource-intensive, being usually carried out by a human expert, motivating the development of automated procedures. The method in practice used the ΔE difference in CIELAB color space measured every 2.5mm in order to determine the distortion strength and classify it according to a threshold (*Handbuch zur technischen Abnahme von Bogenoffset-Rollenoffsetmaschinen* 1996; *Technische Richtlinien Abnahme von Bogenoffsetdruckmaschinen* 2005). The performance of this procedure is rather poor, which prompted the development of our system. The "Arbeitskreis Streifenmessung" consisting of representatives from the major printing machine manufacturers, mechanical and plant engineering association (VDMA), federal and state printing associations and printing-related research institutes accompanied the development.

The model follows the standard multi-channel model as described in 2.2. The parameters of the model have been fit according to assessments collected from human subjects. One group of naive subjects conducted the assessments with streak distortions displayed on-screen, establishing a baseline for the model. Another group of experts conducted assessments on actual printings which were scanned for processing by our model.

In our experiments the subjects performed a judgment task were asked to mark the position of a streak distortion and rate the impairment of the distortion on a scale similar to the



(a) (b)

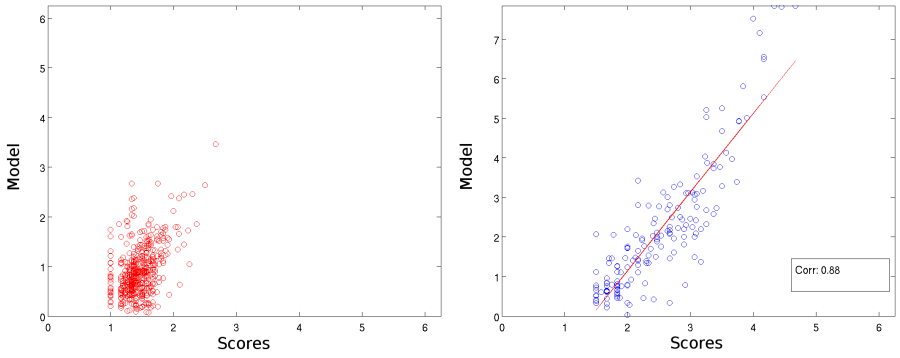
Figure 2.6: Correlation between model predictions and inter-individually averaged assessments for (a) naive observers and (b) experts. The red circles indicate assessments of signals at the threshold of perception, containing at least one minimum grade response (Source: Gadzicki and Zetzsche, 2013).

one defined by ITU. The performance of the subjects has been evaluated by the standard deviations of their responses. Both naive and expert observers were stable in their own assessments and able to reproduce their own responses reliably. As a group the experts were more consistent in their responses.

Correlation Between Model and Assessments We have evaluated the performance of our model by measuring the correlation of the model's predictions with the inter-individually averaged assessments of the observers. Figure 2.6a shows the results for naive observers and Figure 2.6b for the expert observers. The naive observers show a better correlation to the model than the experts even though their deviations of assessments as a group were worse than for the expert observers. This can be explained by the presentation medium as well as by the type of streak distortions presented. The printings had to be scanned, introducing additional noise and contained a significant number of distortions at the threshold of visibility. Such patterns pose a problem to the model, since even though the model could (and actually does) output predictions below the threshold of detection, the assessments given by the observers do not contain information about sub-threshold distortions. We investigated the model behavior at the near-threshold pattern by omitting streaks which received a certain percentage of minimum grades (1, barely visible). The correlation of the model for subsets without such pattern is shown in Figure 2.7. One can see that the distortions at threshold have a high variance with different levels mapped to the minimum grade (Figure 2.7a). Without this subset of distortions, the model's correlation increases from 0.81 to 0.88, meaning that the model is capable of capturing the entire scale of assessments.

2.4 Statistical Operations in Computer Vision

The HVS model described so far are basically models of foveal vision. For peripheral vision recent approaches featuring statistical summary have been proposed (Balas, Nakano, and Rosenholtz, 2009; Rosenholtz, J. Huang, and Ehinger, 2012; Rosenholtz, J. Huang, Raj, et al., 2012; J. Freeman and Simoncelli, 2011), relying on the usage of the auto- and cross-correlation functions. For V2 work on selectivity for visual texture suggests that statistical information plays a key role (Ziemba et al., 2016; Oleskiw and Simoncelli, 2018). Recently models of V2 were proposed which rely on statistical representations (Oleskiw and Simoncelli,



(a) (b)

Figure 2.7: Correlation between model and inter-individually averaged expert's assessments (a) for ratings containing at least one minimum grade response and (b) for ratings not containing the minimum grade. The red circles indicate assessments of signals at the threshold of perception, containing at least one minimum grade response (Source: Gadzicki and Zetzsche, 2013).

2019; Parthasarathy and Simoncelli, 2020; Oleskiw, Lieber, et al., 2020). The relevance of auto- and cross-correlation functions for the perception of patterns has been investigated for some time (Julesz, 1962; Uttal, 1975; Glünder, 1986; Barlow and Berry, 2011). These functions are defined as

$$h(i) = \frac{1}{N} \sum_{k=-N/2+1}^{N/2} e(k) \circ g(i+k), \tag{2.14}$$

where auto-correlation results if $e(k) = g(k)$ and where $\circ$ indicates multiplication.

Representing probability distributions with neurons has been proposed based on population coding (see Pouget, Dayan, and Zemel, 2003 for a review). This is an implicit representation of a distribution where the activation patterns of a set of neurons are interpreted as representing uncertainty. Furthermore the estimation of a probability density with neural networks has been proposed (Reyneri, Colla, and Vannucci, 2011).

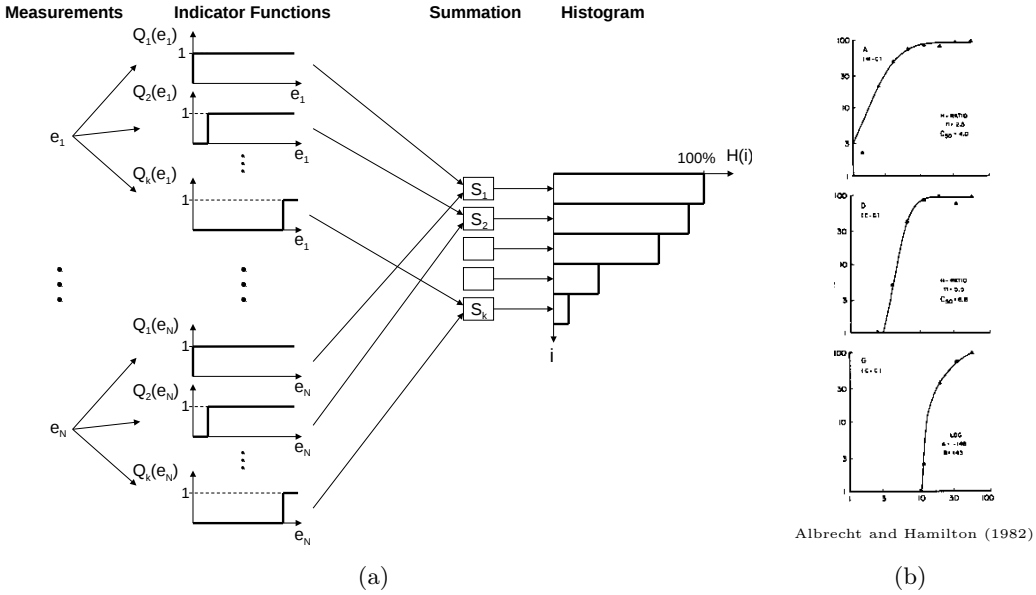
2.4.1 Neural Computation of a Probability Distribution

> This section summarizes our publications *"Statistical Invariants of Spatial Form: From Local AND to Numerosity"* and *"Neural Computation of Statistical Image Properties in Peripheral Vision"*.

Our papers (Zetzsche, Gadzicki, and Kluth, 2013; Zetzsche, Rosenholtz, et al., 2017) propose describe the realization of statistical summary operators with neurons. This is of particular interest for neurobiologically plausible models featuring statistical information since they must rely on functionality attributed to neurons. The auto- and cross-correlation functions are of special interest since they involve multiplication which cannot be realized easily.

The idea behind our work is that a frequency distribution can be represented by the histogram computed by indicator functions assigning values to particular bins. Specific indicator functions can be used to compute reverse cumulative histograms containing the same information as "normal" cumulative histograms (Figure 2.8a). These indicator functions resemble responses of neurons in the visual cortex (Figure 2.8b). Indicator functions have a specific sensitivity and are independent of the input once the threshold has been reached. These properties can also be attributed to neurons in the visual cortex which show a sensitivity to contrast, but can become saturated and lose their sensitivity to contrast. This is achieved

by cortical gain control (contrast normalization) as initially described for cells in the visual cortex (Albrecht and Hamilton, 1982a), but now believed to exist throughout the brain (Carandini and Heeger, 2012).



(a) (b)

Figure 2.8: Computation of the reverse cumulative histogram. (a) shows the set of input variables e_1 to e_n being passed as input to indicator functions. The output of each indicator function is summed in a specific bin of the histogram.
(b) The response functions of three neurons in the visual cortex (Albrecht and Hamilton, 1982a) show similarity with the indicator functions. They have different sensitivities and saturate into a constant output after the transition range (Source: Zetzsche, Gadzicki, and Kluth, 2013).

An approximation of the reverse cumulative histogram can be achieved by replacing the binary indicator functions with the smooth gain control functions. The information about a probability distribution available to the visual cortex is illustrated in Fig. 2.9. The reconstructed histogram (last row in Figure 2.9) can be viewed as low-pass filtered version of the distribution.

The computation of auto- and cross-correlation functions requires multiplication, which can be modeled using neurons (Resnikoff and Wells, 1984; Adelson and Bergen, 1985; Zetzsche and Barth, 1990), but there is no evidence for such structures. However, the multiplication itself is not crucial as the function only needs to output a high value if two features are similar and a low value if they are not. This can be achieved by AND-like neural operations (Zetzsche and Barth, 1990; Zetzsche and Nuding, 2005) via cortical gain control (Albrecht and Hamilton, 1982a; Heeger, 1992) as shown by Zetzsche and Nuding, 2005. Cortical gain control for two different features $s_i(x, y)$ and $s_j(x, y)$ can be written as

$$g_k(x, y) = g(s_i(x, y), s_j(x, y)) := \max\left(0, \frac{s_i + s_j}{(\sqrt{s_i^2 + s_j^2} + \epsilon)\sqrt{2}} - \Theta\right) \tag{2.15}$$

where $k = k(i, j)$, ϵ is a constant which controls the steepness of the response and Θ is a threshold. The resulting auto-correlation functions is an approximation, but the essential features are captured (see Figure 2.10).

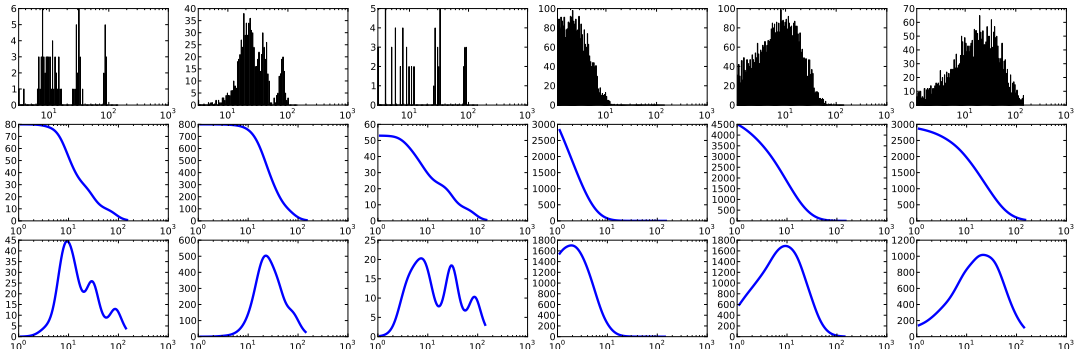


Figure 2.9: Neurobiological computation of a reverse cumulative histogram with the input distribution in the upper row, the corresponding reverse histograms in the middle row and the estimated probability distribution derived from the cumulative distribution in the lower row (Source: Zetzsche, Gadzicki, and Kluth, 2013).

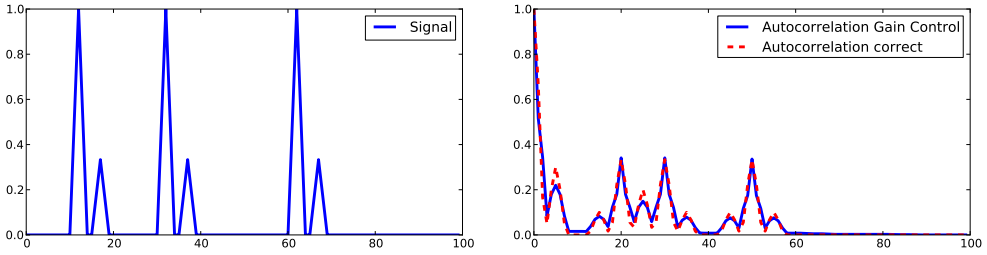


Figure 2.10: (a) shows an input and (b) the corresponding mathematical (red dotted) and neurobiological (blue) auto-correlation function (Source: Zetzsche, Gadzicki, and Kluth, 2013).

2.5 Contribution

In Gadzicki, 2009; Gadzicki and Zetzsche, 2013 we have described the application of a HVS model for the prediction of streak distortions in printings, which, to the best of our knowledge, is new to this domain. We proposed a solution for objective assessment of an important class of distortions since those of the streak type inevitably occur in offset printing machines. Since the measurement procedure in use produces unsatisfactory results, there is a high interest in an objective method which provides better correlation with subjective assessments of expert observers. Our model shows a particularly good correlation to the subjective assessments of both, naive and expert ones over the range of the impairment scale. It performed not as well for the assessments at the threshold of perception levels. As an outlook the model could be modified slightly to predict threshold detection, but the tuning of a modified model would require training data collected from observers for this task.

In Zetzsche, Gadzicki, and Kluth, 2013; Zetzsche, Rosenholtz, et al., 2017 we proposed a neurobiologically plausible computation of a frequency distribution and auto- and cross-correlation functions. We utilized the Naka-Rushton function, which is an established method in vision science to model the gain control of a neuron's response. While statistical operators are commonplace in computer vision, neurobiologically plausible models of human vision must rely on functions that are believed to be computable by neurons or used to model neuron's

responses. Neurobiologically plausible statistical operators are thus of high importance.

CHAPTER 3

Activity Recognition with Convolutional Neural Networks

In this chapter the author covers the connection between the visual system and neural networks in Section 3.1 and the basics of neural networks in Section 3.2. Human activity recognition (HAR) (Section 3.3) sets the context for our contributions on unimodal HAR in Section 3.4 and multimodal HAR with CNNs in Section 3.5. It concludes in Section 3.6 with our contribution which aims at transferring human activity data to a robotic system.

3.1 Feature Invariant Networks

The computational model of HVS described in the previous chapter aims at modeling low-level vision, basically V1. This model allows the general recognition of primitive features, the response to contrast, and interaction between primitive patterns. Moving to recognition of complex patterns requires models that include higher cortical areas.

There are at least two known pathways into which the visual system can be divided, the ventral (form or "what") and the dorsal (motion or "where") (Mishkin, Ungerleider, and Macko, 1983). Even though this separation should not be seen as too strict (Merigan and Maunsell, 1993), it is generally accepted (Wiskott, 2009). The ventral pathway reaches from the retina over the lateral geniculate nucleus (LGN), the visual areas 1 to 4 (V1-4), posterior inferotemporal area (PIT), central inferotemporal area (CIT), anterior inferotemporal area (AIT) to anterior superior temporal polysensory area (STPa) (Wiskott, 2009). The feature selectivity become more and more complex with each area (Figure 3.1) and shift and size invariance increases for the higher areas (Oram and Perrett, 1994).

The neurobiological constrains to computational models for invariant recognition are given by Wiskott, 2009 and include among others a layered structure reflecting the layers of the ventral pathway, a feature hierarchy with increasing feature complexity, an invariance hierarchy regarding shift and learning. Invariant feature networks account for these requirements.

The HMAX model (Serre, Wolf, and Poggio, 2005) is a feature invariant network which stays close to the biological example. It consists of a layer (S1) of Gabor filters which act as simple cells and are organized in orientation and scales. It is followed by a max pooling layer (C1) acting as complex cells, pooling over neighboring scales. The connections of these layers are hard-wired and are interpreted as V1. Training occurs in the following layer (S2) where random patches of various sizes are extracted from layer C1. These are used as prototype vectors for radial basis functions, computing a distance to image patches presented to the

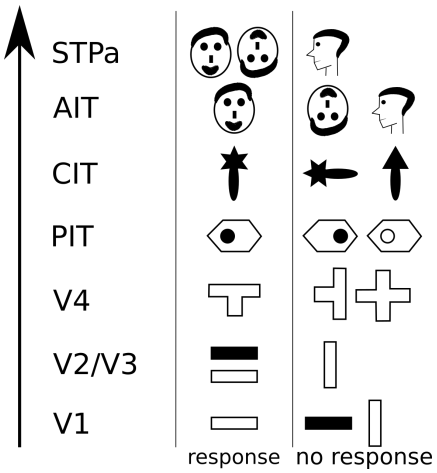


Figure 3.1: Summary of neural responses in the ventral pathway (adapted from Oram and Perrett, 1994.)

model. This layer can be interpreted as V4 or PIT. The final layer (C2) again performs a max pooling over feature maps from S2 over scales and orientations and can be interpreted as inferotemporal cortex (PIT, CIT, or AIT). The resulting feature map from C2 is passed to a Support-Vector Machine (SVM, Cortes and Vapnik, 1995) as a final stage. The HMAX model as described so far is an object classifier, but it was also extended to an action recognition model (Jhuang et al., 2007) by adding a temporal prototype matching stage S3 and a temporal pooling stage C3 on top of layer C2. The features from the last layer were again passed to a SVM.

The HMAX model covers most of Wiskott, 2009's constrains given above. The learning requirement is covered in a rather crude fashion by random sampling and the overall depth of the network is shallow with its four layers. Neural networks account for all these constraints, including training.

3.2 Artificial Neural Networks

Artificial Neural Networks (ANN) are inspired by the human brain and the way neurons interact but they were never meant to serve as realistic models. Instead, they try to imitate some of its behavior in a very basic fashion. Neurons have limited processing power on their own, but are highly connected to other neurons, enabling the entire set of neurons to solve complex tasks. This is the central idea of connectionism or parallel distributed computing (Rumelhart, McClelland, and PDP Research Group, 1986; McClelland, McNaughton, and O'Reilly, 1995) which rose in the context of cognitive science in an attempt to ground the symbolic systems in neural implementation (Goodfellow, Bengio, and Courville, 2016, p. 17).

While artificial neural networks are inspired by the organization of the brain in general, convolutional neural networks (CNN) are more strictly connected to the visual cortex (Laskar, Giraldo, and O. Schwartz, 2018; Grill-Spector et al., 2018). V1 is organized as a spatial map, closely connected to the image structure in the retina (Schiessl and McLoughlin, 2003), which is also the structure of features in CNNs. The simple cells in V1 are described by a convolution (e.g. Gabor filter function (Gabor, 1946; Marcelja, 1980)) with a nonlinear function added which is the design principle of a CNN. The features from the first layer of a CNN are Gabor-like and can be replaced by a Gabor filter bank (Calderón, Roa, and Victorino, 2003; Luan et al., 2018). Complex cells receive their inputs from multiple simple cells, resulting in an

invariance to phase shifts of patterns and small spatial shifts. The pooling operation in CNNs within and over channels attempt to emulate this behavior (Riesenhuber and Poggio, 1999).

3.2.1 Historical Development

This is a brief summary of the historical development (see Goodfellow, Bengio, and Courville, 2016 for a more detailed history).

In the 1950s the Perceptron model was introduced by Rosenblatt, 1958; Rosenblatt, 1962 which serves as the foundation of ANNs up to now. The single-layer Perceptron model was a binary classifier which introduced a weighted sum of its inputs to determine activation and a bias which is a constant input added to the cell. Widrow and Hoff, 1960 suggested a model called ADALINE which features a Least Mean Squares (LMS) algorithm for adjusting the weights. It is a special case of stochastic gradient decent used until now though with slight modifications (Goodfellow, Bengio, and Courville, 2016, p. 15).

The Cognitron (Fukushima, 1975) and Neocognitron (Fukushima, 1980) introduced rectified linear units which add a slight nonlinearity at the output of a neuron's model. The back-propagation algorithm (Rumelhart, Hinton, and Williams, 1986b) and parallel distributed computing (Rumelhart, Hinton, and Williams, 1986a) enabled the design and training of multi-layer networks with hidden units (Goodfellow, Bengio, and Courville, 2016, p. 15). These models turned out to solve nonlinear problems like the XOR problem and later were shown to be universal approximators (Cybenko, 1989; Hornik, 1991).

Convolutional Neural Networks were introduced by Denker et al., 1989 with hand-crafted convolutional kernels and by LeCun, Boser, et al., 1989 with trained weights. The convolution operation enables neural networks to process image data efficiently. The first networks had a relatively shallow structure with only three hidden layers (LeCun, Boser, et al., 1989) or five layers in LeNet5 (LeCun, Haffner, et al., 1999), but outperformed classical approaches at tasks like hand-written digit recognition.

Deep neural networks gained their breakthrough with AlexNet (Krizhevsky, Sutskever, and Hinton, 2012) beating the Imagenet (Deng et al., 2009) challenge by a wide margin. Big data sets like the Imagenet data set with over 1M examples, and better hardware enabling the storage of large networks in memory and train them in reasonable time were the major factors enabling the success of deep neural networks.

3.2.2 Basic Architecture

The major aspects of a parallel distributed processing model like a neural network are defined by Rumelhart, Hinton, and McClelland, 1986. Several processing units with a state of activation and an output function for each unit are connected according to a particular pattern. There is a propagation rule for forwarding of activations, an activation rule for combining the incoming inputs and a learning rule. Lastly it must be situated within an environment.

Feed-forward neural networks are commonly organized in layers of simple neurons. Each unit is modeled as a linear combination of the inputs with an activation function which is applied element-wise (Goodfellow, Bengio, and Courville, 2016, p. 174). Historically this was a threshold function (Rosenblatt, 1958), but current implementations use either a logistic sigmoid function, tanh or rectified linear functions (Jarrett et al., 2009). The Rectified Linear Unit (ReLU) is found in practically all modern approaches due to superior performance (Nair and Hinton, 2010) and observations from neuroscience (Glorot, Bordes, and Bengio, 2011).

The units are organized in layers with an input and output layer and one or more hidden layers in between (see Figure 3.2). The input layer generally just passes on the input data while the activation of output units is tailored toward the task. Most commonly the output unit's activation function is a linear function for Gaussian, a sigmoid function for Bernoulli or softmax for Mulitnoulli output distributions (Goodfellow, Bengio, and Courville, 2016,

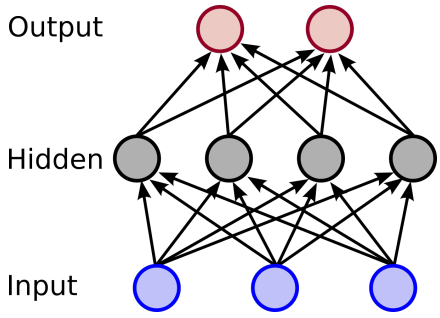


Figure 3.2: Structure of a simple fully-connected feed-forward neural network with one hidden layer.

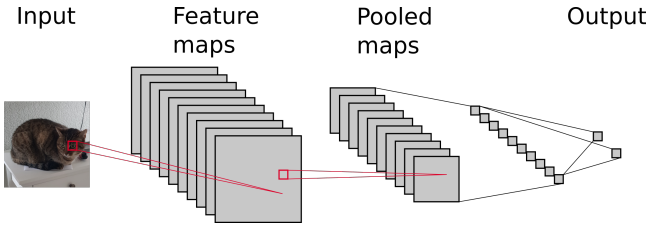


Figure 3.3: Example of a very simple CNN

pp. 181–187). The activation functions also determine the loss functions used for determining the discrepancy between ground truth and prediction in supervised learning.

Hidden layers form the body of the neural network, providing the processing power to solve complex problems. Hidden units use ReLUs or units with functions derived from it (Goodfellow, Bengio, and Courville, 2016, pp. 191–195), e.g. leaky ReLU (Maas, Hannun, and Ng, 2013) or maxout units (Goodfellow, Warde-Farley, et al., 2013). Deep learning networks are characterized by having a significant number of hidden layers, e.g. up to 152 layers in Residual Networks (ResNet) (He, X. Zhang, et al., 2016).

3.2.3 Convolutional Neural Networks

Convolutional neural networks (CNN) use the convolution instead of the simpler linear combination and are suited for processing of uniformly sampled data, e.g., 1D time-series, 2D images or 3D videos. The convolutional kernel is applied to every position of the input in the simplest case and generates a feature map. A convolutional layer has usually several channels, each representing a different feature map generated by a specific kernel. The convolution operation is again followed by a rectifier linear function as an activation function. CNNs also use pooling operations which apply a statistical operation to a region of the feature map, e.g. maximum pooling (Zhou and Chellappa, 1988) which replaces the region with the pooled value (Goodfellow, Bengio, and Courville, 2016, pp. 331–345). Often the pooled maps are reduced in resolution, effectively performing sub-sampling. Figure 3.3 shows these stages together with fully-connected output layers.

3.2.4 Neural Networks as Classifiers

Feed-forward networks are commonly used as discriminative classifiers, learning to map features to classes, or, in other words, the conditional probability distribution of classes given the features. In contrast to classical machine learning approaches, neural networks learn the

features and their mappings while traditional machine learning relies on hand-crafted feature extractors, learning the mapping only (Goodfellow, Bengio, and Courville, 2016, p. 3–5).

Deep learning with CNN-based approaches has become remarkably successful over the last decades. A common theme was the increase in depth of the networks, from LeNet with five layers (LeCun, Haffner, et al., 1999) over AlexNet with eight layers (Krizhevsky, Sutskever, and Hinton, 2012), VGG's 19 layers (Simonyan and Zisserman, 2014), GoogLeNet's 22 layers (Szegedy et al., 2015) to ResNet's 152 layers (He, X. Zhang, et al., 2016). Unfortunately, the infinite addition of layers did not turn out to be the almighty solution since networks run into degradation problems when they become too deep. The performance eventually saturates or even decreases if network become too deep. This problem was addressed by the introduction of special buildings blocks for CNNs, e.g., Inception Units (Szegedy et al., 2015), Residual Units (He, X. Zhang, et al., 2016) or Dense Blocks (G. Huang et al., 2017). Inception units feature parallel micro-pathways, including 1x1 convolutions from the "network-in-network" approach (Lin, Chen, and Yan, 2014). They grow in width instead of depth while keeping the number of parameters within the feasible range. Residual units, like the name suggests, compute a residual which is added to the previous input layer, thus avoiding the degradation problem, and allowing for very deep networks. Dense blocks add additional connections between layers, i.e., a layer is connected to every subsequent layer within a block.

Originally applied to the image domain for object recognition (Krizhevsky, Sutskever, and Hinton, 2012) they have been applied to various data types and domains, including human activity recognition.

3.3 Human Activity Recognition

Human Activity Recognition (HAR) very broadly aims at the analysis and recognition of human actions using information obtained from sensors (Beddiar et al., 2020). It has become a major field of research due to its relevance for video surveillance (Ji et al., 2013), video retrieval (Ramezani and Yaghmaee, 2016), human-computer interaction (Choi et al., 2008), robotics (Koppula and Saxena, 2016) or autonomous driving (Ryoo and J. K. Aggarwal, 2011; Rasouli and Tsotsos, 2018). It has grown very fast, producing a large amount of literature (Beddiar et al., 2020).

Action and activity are terms often used synonymously though some authors distinguish between them. J. Aggarwal and Ryoo, 2011 describes actions as a subset of activities, defined as single person activities, and distinguished from gestures (elementary body movements, e.g., raising an arm), interactions (activities involving two or more people and/or objects, e.g., shaking hands) and group activities (a group of people acting together, e.g., marching). In this work the terms are used synonymously.

Activity recognition attempts to classify what is happening inside a specific time frame given the sensory data. It can be distinguished from action prediction which attempts to predict what will happen after a specific time frame (J. Aggarwal and Ryoo, 2011). Examples for action prediction can be found in Ryoo, 2011; Kong, Kit, and Y. Fu, 2014; Kong, Tao, and Y. Fu, 2017. The activity recognition task can be defined more formally as a set of predefined activities with sensor readings at time points for which a model predicts the activity sequence based on the sensor readings. The model minimizes the discrepancy between predicted activities and ground truth activities typically by utilizing a positive loss function (J. Wang et al., 2019).

There is a variety of sensor types which have been used for HAR and can be broadly divided into sensors carried on the body, e.g., accelerometers, object sensors, e.g., markers or Radio Frequency Identifier (RFID) tags placed on objects, and remote sensors, e.g., cameras (J. Wang et al., 2019). The modalities used for HAR can be aquired directly, e.g., RGB video, skeleton data from motion capturing, or extracted from another modality, e.g., skeleton data

from RGB video or depth video, optical flow from RGB video, etc.

3.3.1 Classical Machine Learning Approaches

HAR methods can also be divided into those with hand-crafted and with learned features, stemming from the distinction between machine learning approaches. The hand-crafted methods employ feature extraction at space-time interest points (STIP) (Laptev, 2005), densely sampled (Uijlings et al., 2015) or along trajectories (H. Wang et al., 2013). The most used descriptors are Histogram of Oriented Gradients (HOG) (Dalal and Triggs, 2005; Laptev et al., 2008), Histogram of Optical Flow (HOF) (Laptev et al., 2008) and Motion Boundary Histograms (MBH) (Dalal, Triggs, and Schmid, 2006). Further proposed were Fourier-based viewpoint invariant motion descriptors (Weinland, Ronfard, and E. Boyer, 2006), anticipatory temporal conditional random fields (Koppula and Saxena, 2016) or 3d cuboid descriptors (Xia and J. K. Aggarwal, 2013). The descriptor are often aggregated by super-vector based encoding methods, e.g., improved Fisher-Vectors (iFV) (Perronnin, Sánchez, and Mensink, 2010), super-normal vector (SNV) (X. Yang and Tian, 2014) or Vector of Locally Aggregated Descriptors (VLAD) (Jégou et al., 2012).

3.3.2 Neural Network Approaches

In the last decade the field has moved more towards neural network approaches, either as end-to-end solutions or connecting classical machine learning with deep learning. Deep neural networks (DNN) served as feature extractors, e.g., from images with 2D CNNs classified with a Support Vector Machine (SVM, Cortes and Vapnik, 1995) (Kong, Tao, and Y. Fu, 2017). They were also used as classifiers for hand-crafted features. Vepakomma et al., 2015 extracted features from wearable sensors, i.e., statistical information from acceleration data, and processed the data with a DNN. Walse, Dharaskar, and Thakare, 2016 used sensor data obtained from a smartphone, performed a PCA and classified them with a DNN.

3.4 Unimodal Human Activitiy Recognition with Deep Neural Networks

RGB video data is one of the most widely used modalities for activity recognition. In terms of processing, it can be treated either as a series of 2D images, each processed independently with a CNN, or as spatio-temporal data processed by 3D CNNs.

Ji et al., 2010 was among the first to use a 3D CNN for HAR. Being developed before AlexNet (Krizhevsky, Sutskever, and Hinton, 2012), the network had a rather shallow architecture with six layers and resembled the HMAX model (Serre, Wolf, and Poggio, 2005; Jhuang et al., 2007). In contrast to HMAX where the features were given by a Gabor filter bank, the CNN was trained to learn the feature representation in Ji et al., 2010. Ji et al., 2013 extended their model (Ji et al., 2010) by an auxiliary feature extraction path. The second path used hand-crafted features (Scale-invariant feature transform (SIFT) (Lowe, 2004)) for regularization of the CNN. G. W. Taylor et al., 2010 also used a 3D convolution, but it was only one layer in a multi-stage approach, consisting of a convolutional Gated Restricted Boltzmann Machine (convGRBM), 3D convolutional layer, abs rectification and spatial and temporal pooling layers. Tran et al., 2015 proposed C3D, a 15-layer 3D CNN, which was also used as a feature extractor with a linear SVM as the final stage in their work. They investigated the dimensionality of spatio-temporal kernel size for the HAR task and found that equally sized kernel performed best (3x3x3 in their work). The 3D convolutional architecture performed well on many data sets and the idea was adopted by others (Varol, Laptev, and Schmid, 2018; Qiu, Yao, and Mei, 2017). Qiu, Yao, and Mei, 2017 introduced Pseudo-3D

residual units which took inspiration from ResNet (He, X. Zhang, et al., 2016) used for image processing. They proposed several designs for the 3D residual unit, separating spatial and temporal convolutions and processing them sequentially or in parallel.

3.4.1 Human Activity Recognition with Deep Learning on Skeleton Data

Skeleton data consist of joint position of the human body over time, stated as 3d-coordinates within a frame of reference. They can be directly measured with motion tracking systems. They usually consist of several cameras with known positions, estimating the position of multiple joints within the cube spanned by the system. These setups are generally expansive and only usable in labs but offer good precision. The Microsoft Kinect is an affordable sensor which can extract joint information from a depth image (Shotton et al., 2011). Joint information can be also extracted from RGB video through pose estimation methods (He, Gkioxari, et al., 2017; Pavllo et al., 2019).

In CNN-based approaches the joint positions are sometimes transformed into 2D structures and processed by 2D CNN. Color coding for temporal dynamics and treating the joint positions as 2D data (Y. Hou et al., 2018), using one image dimension for coding of the spatial structure and the other for temporal dynamics (Du, Y. Fu, and L. Wang, 2015) or projecting the 3D joint positions onto a 2D plane (C. Li et al., 2017) were proposed for using 2D CNN for HAR. Pham et al., 2018 takes the approach from Du, Y. Fu, and L. Wang, 2015 and maps joints to pixel positions, carefully arranging them according to human body physical structure. They use a ResNet for the actual task. Z. Yang et al., 2018 used a tree structural skeleton image and introduced a long-sequence attention mechanism while processing the data with a 2D ResNet. P. Zhang et al., 2019 use the skeleton to image mapping from Du, W. Wang, and L. Wang, 2015, but add a view adaptive to approximate virtual viewpoints on the skeleton. The networks used here were AlexNet (Krizhevsky, Sutskever, and Hinton, 2012), ResNet (He, X. Zhang, et al., 2016), a recurrent neural network (RNN) and a combination of CNN and RNN.

CNN-based approaches can treat the x, y, z-position of joints as separate time series and process them with a time convolutional network (T. S. Kim and Reiter, 2017).

3.4.1.1 Optimal Depth for CNN with Simple Residual Units

> This section summarizes our publication *"Deep Residual Temporal Convolutional Networks for Skeleton-Based Human Action Recognition"*.

As previously stated in 3.2.4 the base units and depth of a network are important factor for the performance of a CNN. In our paper (Khamsehashari, Gadzicki, and Zetzsche, 2019) we have investigated this aspects for residual units, their architectural organization and the overall depth of the network. The task used for the investigation was human activity recognition with skeleton data.

Pham et al., 2018 has investigated residual unit networks (ResNet) for depth between 20 and 110 layers and achieved good results with skeleton data for HAR. Interestingly a temporal convolutional network approach with a modified residual (Res-TCN) unit has achieved comparable results with a much shallower 11-layer architecture (T. S. Kim and Reiter, 2017). The modified units (Figure 3.4c) are simpler than the originally proposed for the image domain (He, X. Zhang, et al., 2016) (Figure 3.4a) or skeleton data (3.4b).

We systematically analyzed the Res-TCN architecture with depth ranging from 11 to 152 layers, using two variants of the architecture called Deep Res-TCN-3 and Deep Res-TCN-4. The evaluation was done on the NTU RGB+D data set (Shahroudy et al., 2016), at that time the largest set featuring multiple modalities (RGB, depth and IR video and skeleton data) with more than 56k training videos across 60 action classes..

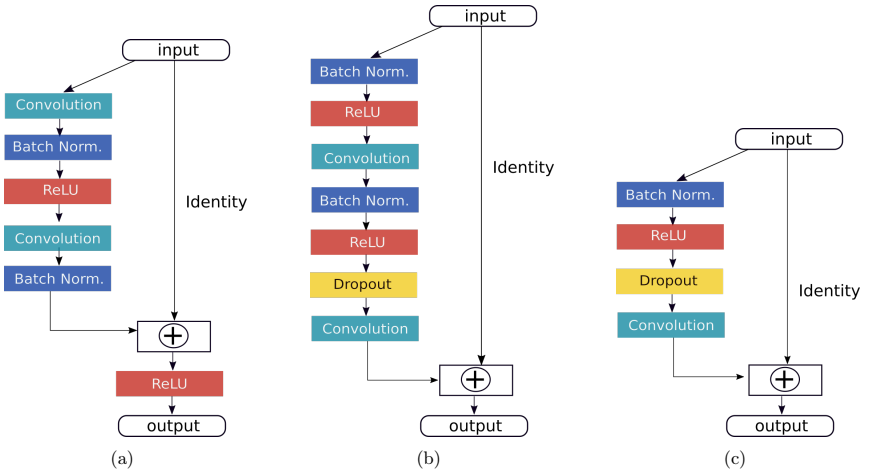


Figure 3.4: The basic residual unit in different approaches. (a) original ResNet (He, X. Zhang, et al., 2016); (b) improved ResNet (Pham et al., 2018); (c) Res-TCN (T. S. Kim and Reiter, 2017) (Source: Khamsehashari, Gadzicki, and Zetzsche, 2019).

Deep Res-TCN-3 sticks with the original architecture of Res-TCN (T. S. Kim and Reiter, 2017), keeping the number of blocks between down-sampling layers at 3, but with additional layers per block. We chose this setting to evaluate the performance of the simpler residual units while remaining as directly comparable as possible. Deep Res-TCN-4 uses a 4-block scheme and stems from the original ResNet architecture (He, X. Zhang, et al., 2016; Pham et al., 2018), using 4 blocks between down-sampling as shown in Figure 3.5 We exchange the original residual units with the simpler one to see whether it limits the performance. In addition to testing the influence of depth on the two ResNet architectures with simple residual units we also tested the influence of hyperparameters found in our paper Gadzicki, Khamsehashari, and Zetzsche, 2018. We used the improved parameters for a further test with the Res-TCN-4 architecture to disentangle the connection between of hyperparameters and architecture. The performance was measured by the accuracy for cross-view and cross-subject tasks.

The main result is that all deeper variants provide a better classification than the 11-layer Res-TCN (T. S. Kim and Reiter, 2017) and the 56-layer ResNet (Pham et al., 2018). The

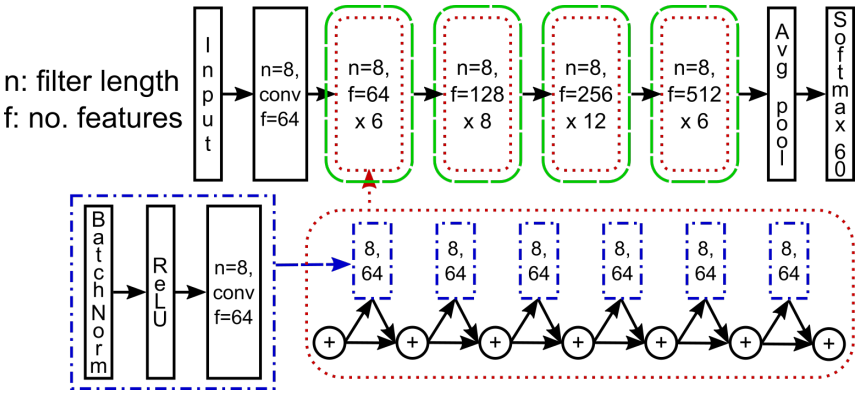


Figure 3.5: Deep Res-TCN-4 architecture with 34 layers (Source: Khamsehashari, Gadzicki, and Zetzsche, 2019).

optimum though is not at the deepest architectural variant, but at rather moderate depth of 18 layers for Deep Res-TCN-4 and 34 layers for Deep Res-TCN-3. For deeper configurations the performance levels off or decreases slightly. This has probably little to do with the architecture in terms of number of layers and blocks (Pham et al., 2018 reaches optimal performance with 56 layers with continuous increase up to that depth), but rather with the simpler design of the residual units. Though He, X. Zhang, et al., 2016 claim that ResNet does not suffer from degradation problems, the simpler units from T. S. Kim and Reiter, 2017 seem to do it, possibly due to having less expression power with one convolution only.

Even though our results were better than the Res-TCN approach when published, they were surpassed for cross-subject (Z. Yang et al., 2018) but not for cross-view. In (Gadzicki, Khamsehashari, and Zetzsche, 2018) we have shown that with an appropriate tuning of the hyperparameters for training of the original Res-TCN, it can outperform the basic ResNet approach for cross-subject. The best performing model for unimodal skeleton data is Shi et al., 2020 right now (April 2021).

3.5 Multimodal Activity Recognition with Neural Networks

Multimodal processing in the stricter sense refers to the processing of data obtained from different sensor types. We use the term also for modalities which are derived from one sensor source, e.g., RGB video and optical flow are different modalities even though the optical flow stems from the RGB video.

3.5.1 Multisensory Processing in Human Perception

Human perception generally involves multisensory processing and can be described as a unified percept of different sensory information sources (Recanzone, 2009). The ventriloquist effect lets us assign the auditory sensory information from the puppeteer to the moving mouth of a puppet as perceived by the visual system (Radeau and Bertelson, 1977). McGurk and MacDonald, 1976 showed in a classic experiment how the disagreeing visual and auditory sensory information form the perception of something entirely different (auditory stimulus 'ba-ba' and visual stimulus 'ga-ga' result in 'da-da' perceived).

Multisensory processing can be described as the interaction between sensory modalities during perception, i.e., one sensory modality can influence another modality during its processing. The goals are to increase the accuracy of perception and the control of perceptually guided actions (Briscoe, 2016). The tracking of an object, e.g., an animal in the ground cover, might be significantly easier with visual and auditory sensory information than with one modality alone. Objects with similar visual appearance might become distinguishable with additional tactile or olfactory information. The integration of multiple sensory modalities, once a common source has been identified, is another form of multisensory processing (Briscoe, 2016).

Areas of higher-order cognition like the neocortex are believed to process mostly multisensory information (Ghazanfar and Schroeder, 2006). Traditional models assume low-level processing of sensory information to be unimodal (Felleman and Essen, 1991), but there is growing evidence for multisensory interactions in the early stages (Schroeder and Foxe, 2005; Ghazanfar and Schroeder, 2006). Auditory-visual interactions have been shown to happen within 200 msec after the stimuli were presented, with earliest observations of interaction patterns in the visual cortex after 40 msec (Giard and Peronnet, 1999). There is evidence for connections from the auditory cortex (A1) to visual cortex (V1) (Falchier et al., 2002).

3.5.2 Exploring Fusion Strategies for Multimodal Activity Recognition with CNN

The fusion of multiple sensory sources is an established concept in science and engineering. The data fusion process can be defined as "associating, combining, integrating and mixing data provided by multiple spatio-temporal data sources" (Bellot, A. Boyer, and Charpillet, 2002). The general goal is to increase the performance of a system which uses the sensory data. More specifically, Bellot, A. Boyer, and Charpillet, 2002 proposed four potential gains of this process:

- **Gain in representation**: a higher level of abstraction or granularity with richer semantic can be reached by fusion, compared to the individual data sources.
- **Gain in certainty**: the fusion process introduces a growth in belief on the data.
- **Gain in accuracy**: the fusion process decreases the standard deviation, noise and errors of the data.
- **Gain in completeness**: the fusion process brings new information to the knowledge about the environment.

Generally fusion of different modalities in machine learning aims at improving the system performance regarding e.g., recognition accuracy or robustness towards noise data. By using multi data source, potential correlations between them might be exploited, helping to disambiguate samples. Baltrušaitis, Ahuja, and Morency, 2019 state five challenges to relating multiple modalities:

- **representation** of multimodal data that exploits the complementary and redundant data sources,
- **translation** of mapping of modalities to another,
- **alignment** of elements from different data sources,
- **fusion** of multiple modalities for prediction, and
- **co-learning** for transferring knowledge from one modality to another.

Fusion is one of the challenges for recognition tasks. The most common fusion approaches are late and early fusion (D'mello and Kory, 2015) and can be distinguished by the position within the processing pipeline where the fusion takes place.

Late fusion (Snoek, Worring, and Smeulders, 2005; Atrey et al., 2010) is the decision-based fusion (Baltrušaitis, Ahuja, and Morency, 2019), basically integrating unimodal approaches into one model for prediction. It is a very flexible method, allowing to combine sophisticated models for modalities into a single model. Here the decisions of individual models, e.g., the predictions in a recognition task, are fused by averaging or major voting. The major drawback of this fusion method is the lack of exploitation of cross-correlations and interactions between individual modalities (Baltrušaitis, Ahuja, and Morency, 2019).

For exploitation of cross-correlations and interactions individual modalities have to be processed together, as in early fusion (Baltrušaitis, Ahuja, and Morency, 2019). It is a feature-based fusion where raw data or early features from each data source must be aligned and synchronized for processing with a single model which has to suit all modalities.

Halfway fusion (Liu et al., 2016) or middle fusion (Damer et al., 2019) place the fusion point somewhere within the model. The model has to consists of several stages in order to allow for it, like e.g., neural networks with multiple layers.

Between early and late fusion at the extreme ends, there exist hybrid fusion approaches trying to combine the properties of both (D'mello and Kory, 2015). As an example, two

modalities are fused in early fashion, but also processes unimodally in parallel and fused via late fusion again (Wu, Cai, and Meng, 2005).

An overview of multi-sensory fusion for activity recognition can be found in Aguileta et al., 2019 and B. Fu et al., 2020.

3.5.3 Multistream Networks

Multi-stream networks were introduced by Karpathy et al., 2014 in the form of multiresolution CNNs. The network used 3D convolutions with an architecture similar to AlexNet (Krizhevsky, Sutskever, and Hinton, 2012). From a video, a foveal stream was extracted by taking the center pixels in full resolution and a context stream was taken by sub-sampling the video. This organization is inspired by the foveal and peripheral processing of the visual system. Both streams served as input to the network and were fused in several ways for comparison. While the model by Karpathy et al., 2014 showed mixed results regarding the recognition performance in comparison to models using hand-crafted features or unimodal CNNs the general idea of multistream networks was picked up by others.

Simonyan and Zisserman, 2014 used a spatial and temporal stream by adding dense optical flow to the original RGB video. Dense optical flow is computed for every pixel in an image and represents the displacement vector in horizontal and vertical position between two images. The input was processed by a 2D CNN with five convolutional, three pooling, two dropout and a final softmax layer. The video and optical frames were processed individually, effectively performing HAR from still images. The individual results were combined in a late fusion manner at the end by averaging or serving as input to a SVM. The model outperformed other approaches which can be generally attributed to the information about temporal dynamics provided by the optical flow path. Individually the optical flow path performed better than the spatial path, but the fusion of both improved the results.

Feichtenhofer, Pinz, and Zisserman, 2016 extend the model by fusion at different levels of the network. The processing was done frame by frame with models from object recognition (Simonyan and Zisserman, 2015). They investigated different strategies for late and mid-level fusion, differentiating between fusion for spatial and temporal streams. Some fusion strategies involved fusing twice by concatenating spatial and temporal path into a single path while maintaining the temporal path and fusing it again late. The double fusion strategy performed well but had the consequence of a significant increase in training parameters. A much simpler late fusion of softmax layers performed nearly equally well with a much smaller overhead.

Carreira and Zisserman, 2017 merged the ideas of 3D CNNs (Ji et al., 2010; Ji et al., 2013; P. Taylor et al., 2015; Tran et al., 2015; Varol, Laptev, and Schmid, 2018) and two-stream networks (Simonyan and Zisserman, 2014; Feichtenhofer, Pinz, and Zisserman, 2016) into the Inflated 3D CNN (I3D). The two streams were RGB video and optical flow which were merged via late fusion. The 3D CNN was an inflated version of a variant (Ioffe and Szegedy, 2015) of the Inception network (Szegedy et al., 2015). Inflated means that a pre-trained 2D CNN is extended by a third dimension resulting in a kernel with 2D weights which remain constant in the time dimension for initialization of the network. The Inception architecture (Figure 3.6) aims at firstly approximating a local sparse structure by dense components and secondly reducing dimensions wherever computational requirements would increase too much (Szegedy et al., 2015). The two-stream I3D model outperformed other approaches with its late fusion of streams. While Carreira and Zisserman, 2017 have investigated other ANN approaches beside I3D in their works, they did not investigate other fusion strategies.

3.5.3.1 Late Fusion of Multiple Modalities

> This section summarizes our work published in *"Multimodal Convolutional Neural Networks for Human Activity Recognition"*.

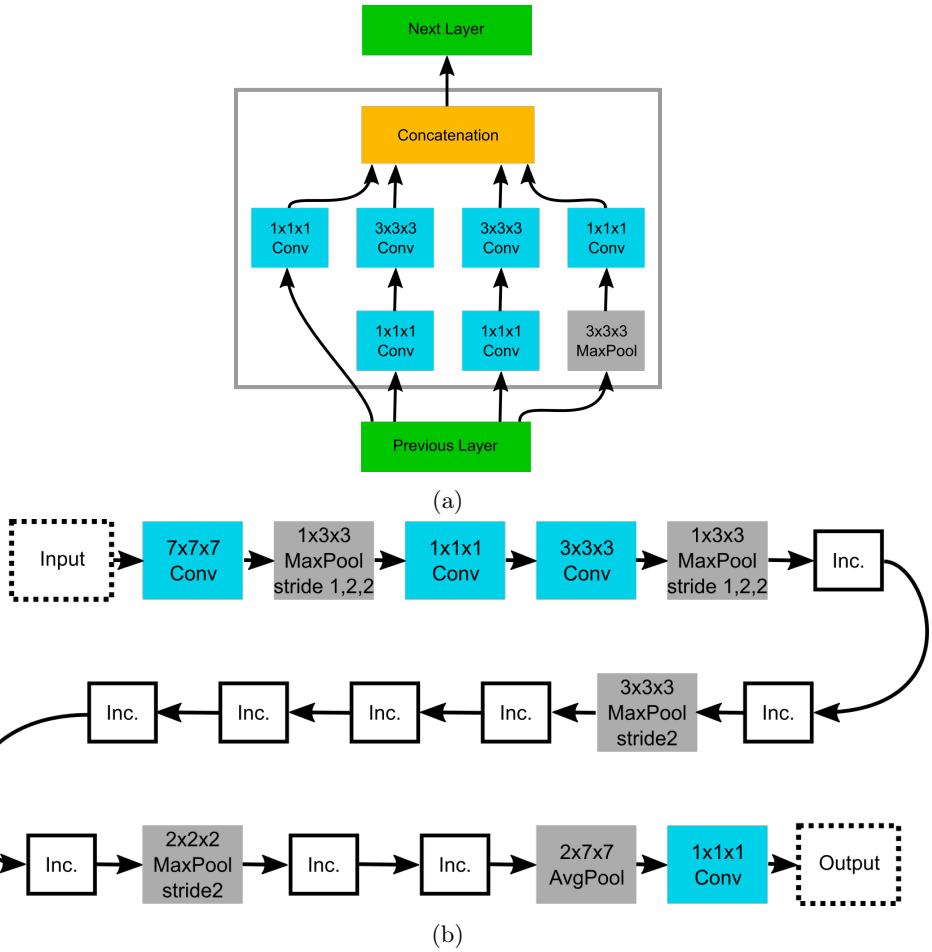


Figure 3.6: (a) Structure of an Inception-v1 block. (b) Layout of the Inception-v1 I3D CNN (Source: Gadzicki, Khamsehashari, and Zetzsche, 2020, adapted from Carreira and Zisserman, 2017)

In our paper (Gadzicki, Khamsehashari, and Zetzsche, 2018) we have investigated late fusion with convolutional neural networks for activity recognition. We have used Inception-v1 CNNs (Szegedy et al., 2015), pre-trained on the "Kinetics" data set (Kay et al., 2017), for RGB and optical flow and Res-TCN CNN (T. S. Kim and Reiter, 2017) for skeleton data which was trained from scratch. Our optical flow has been computed with "Flownet 2.0" (Ilg et al., 2017). We fused the networks in a late fusion fashion and trained the output (dense) layer only. We tested our approach on the "NTU RGB+D" Shahroudy et al., 2016 data set which provides several modalities in the form of RGB video, depth images and skeleton data. It was the largest freely available data set for multiple modalities, featuring over 56k examples, 60 classes and recordings from three different viewpoints. Common evaluations for this data set are cross-subject and cross-view.

During the training of the skeleton network, we found better hyperparameters which delivered state-of-the-art performance in terms of cross-subject recognition. Our late fusion results were dominated by the skeleton network; thus fusion of additional modalities did not improve the performance. The shortcoming of our approach was to use the Inception network merely as a feature extractor for RGB and optical flow. Even though it was trained on the very large "Kinetics" data set and should generalized well, that was not enough to compete

with a skeleton network which achieved comparative accuracy on the "NTU RGB+D" data set.

3.5.3.2 Comparison of Early and Late Fusion

> This section summarizes our work published in *"Early vs Late Fusion in Multimodal Convolutional Neural Networks"*.

In our paper (Gadzicki, Khamsehashari, and Zetzsche, 2020) we compared the performance of the I3D architecture for early and late fusion. For early fusion either raw data or early features, e.g., the output of the first convolutional layer, can be combined. Early fusion for raw data is not necessarily a trivial task since data from different sources are rarely already spatio-temporally aligned regarding resolution or sampling frequencies. Thus, they require a certain amount of pre-processing before being processed by a CNN. The fusion of early features can avoid some of the alignment issues by making sure that the output feature maps are aligned and so can be concatenated trivially. If the input data agree in dimensionality, e.g., all input data are 3D data structures as in the case of videos, the feature map alignment can be enforced by selection of appropriate parameters of the convolutional layer. Late fusion is much easier to realize, since the data sources can be processed individually and only the final outputs are merged. For every modality, a specialized CNN (or other classifier) tailored for the data source can be used if they agree on the output representation of class scores. Figure 3.7 shows the network structure for both, early and late fusion.

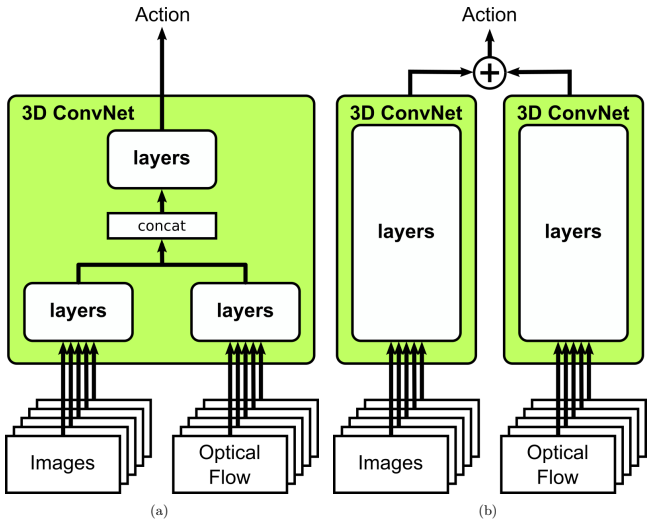


Figure 3.7: CNN with (a) early fusion and (b) late fusion. The modalities shown here are RGB images and Optical Flow. For the early fusion, the action label is directly output by the logits layer of the fused network. For late fusion the outputs of the logits layers of the individual CNNs for each modality are summed (Source: Gadzicki, Khamsehashari, and Zetzsche, 2020).

In our implementation we used early fusion by concatenating the outputs of the first convolutional layer and late fusion by summing the outputs of the softmax layer.

As the basic architecture we used the "Inception v1 I3D" (Carreira and Zisserman, 2017) for video data types and extended it to allow for fusion at arbitrary points in the networks by concatenating the outputs of the previous layers. For the processing of skeleton data we used Res-TCN (T. S. Kim and Reiter, 2017). We tested our approach on the "NTU RGB+D" (Shahroudy et al., 2016) data set.

For our early vs late fusion comparison we have used RGB video and optical flow which has been computed with "Flownet 2.0" (Ilg et al., 2017). In addition, we also used late fusion for RGB video and skeleton data.

As a result, any fusion improved the accuracy in comparison to unimodal processing. Among the fusion strategies tested, early fusion performed best.

3.5.3.3 Fusion of CNN for HAR by others

In the ActionVLAD model Girdhar et al., 2017 use a Vector of Locally Aggregated Descriptors (VLAD) implemented as a network layer and an extension from (Arandjelović et al., 2018). The ActionVLAD layer abstracts from the features by assigning the features to clusters, computing the difference between feature and prototype, and aggregating the differences for a video. The network was tested for early, late and concat fusion of the ActionVLAD layer. The concat fusion is realized in a way that would resemble a middle fusion at the last feature layer when compared to the fusion strategies described so far. The early fusion has little to do with early fusion as described in the other approaches above as the VLADs are taken from the last feature layers. In this approach late fusion performed best.

Temporal linear layer (TLE) (Diba, Sharma, and Van Gool, 2017) compute a low-dimensional embedding of features maps from convolutional layers. It allows to aggregate features from short video sequences or single frames into longer sequences. It was tested with late fusion on different network architectures and performed better than the original networks.

AdaScan (Kar et al., 2017) features adaptive scan pooling of frames from video sequences regarding the importance of a frame for the recognition of the activity. The scan pooling was added on top of a C3D network (Tran et al., 2015) and tested for RGB video and optical flow, merging the streams with late fusion. The addition of adaptive scan pooling improved the performance in comparison to the original network.

Another encoding methods for CNN features was proposed as Spatio-Temporal Vector of Locally Max Pooled Features (ST-VLMPF) (Duta et al., 2017). Features and locations are clustered according to their similarity to learned prototypes and max pooled over space and time resulting in a spatio-temporal feature vector representation, abstracted from the CNN's features maps, but preserving information about features and their positions. The ST-VLMPF representation was extracted from 2D CNNs for RGB and optical flow images and from a 3D CNN for RGB video. This approach performed better than other multistream approaches (Karpathy et al., 2014; Simonyan and Zisserman, 2014) at that time.

Other authors have investigated early fusion in other settings than RGB video with optical flow. Liu et al., 2016 used a "Faster R-CNN" (Ren et al., 2017) for pedestrian detection. Their modalities were RGB and thermal images which were fused in early, late, and middle fashion before the region proposal network split off. They found that middle fusion performed best, but the other fusion strategies all performed better than the unimodal network.

Damer et al., 2019 used a 2D Inception network (Szegedy et al., 2015) for biometric representations (iris and face recognition in their work). They suggested early, late and middle fusion for this network and found that the multimodal representations were more discriminative than the unimodal. Furthermore, the early and late fusions performed better than middle fusion.

Apart from the fusion of modalities into one data stream there are also considerations for interaction between streams on multiple levels. Spatiotemporal multiplier networks (Feichtenhofer, Pinz, and Wildes, 2017) add additive or multiplicative interactions at every layer of a spatiotemporal Resnet (Feichtenhofer, Pinz, and Wildes, 2016). The injection was performed between the residual units, either directly into the result of a residual block (sum of residual and previous block) or into the input of the residual only. Generally unidirectional from the motion to the spatial stream, but bidirectional interactions were tested also. The injection into the residual computation generally performed best, with addition performing

better than multiplication, even though by a relatively small margin.

Beyond optical flow, skeleton data and trajectories are important modalities which show promising results. Action machine (Zhu et al., 2018) uses RGB video as an input, but extracts the skeleton information from this stream as well. The two streams are fused with late fusion. The best performing activity recognition model on the "NTU RGB+D" data set is currently (Davoodikakhki and Yin, 2020). They perform hierarchical action classification with network pruning. For RGB video with trajectories of interest points trajectory-pooled deep convolutional descriptors TDD (L. Wang, Qiao, and Tang, 2015) are used. Context stemming from objects can also be added as an information source. Bobick and Davis, 2001 suggest early fusion via Motion History Images (MHI) for RGB, depth, skeleton from RGB and context from object recognition.

3.6 Human Activity Recognition for Robotics

3.6.1 Human-Robot-Interaction

Human activity recognition plays an important role for the development of robots operating in an environment together with humans. One aspect is the Human-Robot-Interaction (HRI) which covers a broad field of sub-topics. A recent taxonomy for HRI (Onnasch and Roesler, 2020) defines several categories enabling a modular description of scenarios in which humans and robots operate, based on previous HRI framework models. The categories cover the field of application (e.g., industry, service), the exposure to the robot (e.g. in a laboratory or in the field), the robot task specification (e.g. transport, manipulation, information exchange), the robot morphology (e.g. appearance, movement of the robot), the degree of autonomy (e.g. decision making, action implementation), the human role (e.g., supervisor, cooperator), the composition (number of human vs robots), the communication channel (input and output) and proximity between human and robot (physical and temporal).

Human activity recognition has impact on several of these categories. For a robot operating in a domestic environment and being allowed a certain degree of autonomy, it will be crucial to recognize what humans are currently doing to assist them if necessary, avoid them if the robot would disturb them or simply carry out tasks instead of the human. For a high level of autonomy, it is also important to ask how the robot can acquire its skill set. Current robotic systems are equipped with carefully designed routines for manipulation of their environment. Enabling a robot to learn new skills or improve current ones, e.g., by demonstration by humans, could greatly improve a robot's performance.

3.6.2 Human Activity Recognition in Human-to-Robot Pipeline

> This section summarizes our work published in *"From Human to Robot Everyday Activity"*.

In Mason, Gadzicki, et al., 2020 we present a pipeline for the analysis of human activities data. This research has been conducted as part of the collaborative research center "EASE" ("Everyday Activity Science and Engineering", http://www.ease-crc.org) which has the long-term goal to develop cognitive robot capable of performing everyday activities. Within our subproject the goal was to collect data from human activities and transform them into narrative-enabled episodic memories (NEEMs).

The pipeline as shown in Figure 3.8. It starts with recording of human activities in context of kitchen related activities. The determined scenario was setting the table for a given context, given by a set of constraints. Multiple modalities have been recorded from human subjects, including RGB video from seven different perspectives, audio from scene and head-worn microphone, skeleton joint data from motion tracking, eye tracker data and biosignals

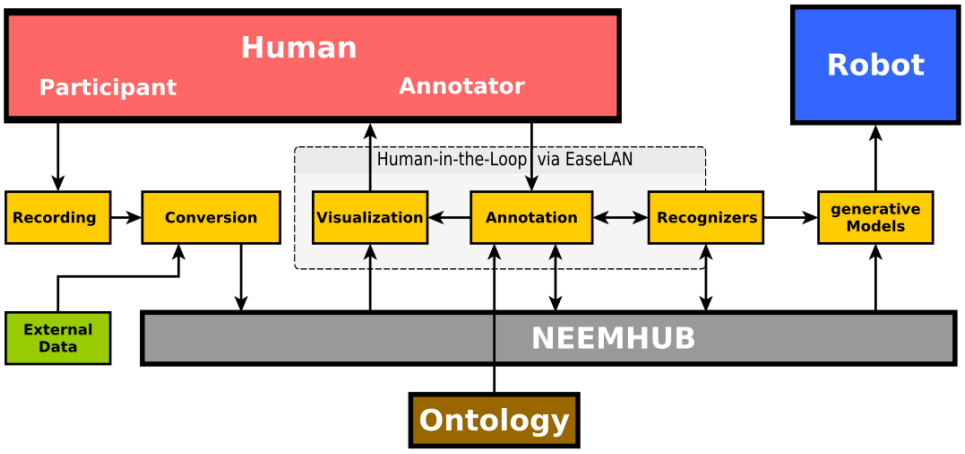


Figure 3.8: The human activities data analysis pipeline (Source: Mason, Gadzicki, et al., 2020).

from muscle and brain activity (Mason, Meier, et al., 2018). The resulting EASE Table Setting Dataset (EASE-TSD) includes 70 sessions of six or more trials.

The data were annotated manually utilizing the EASE Ontology which is designed for knowledge representation and inference for cognitive robots. The annotation scheme reflects the ontology by using hierarchically structured levels of granularity to describe the human activities. It ranges from elementary motions (gestures in the terminology of J. Aggarwal and Ryoo, 2011) of short duration (e.g. reach, grasp, lift) to actions of higher complexity composed of the motions (e.g. picking, carrying an object) to high level task related activities (e.g. planning, object retrieval). Multiple labels could be present at any time step, accounting for actions taking place simultaneously.

Several automatic data annotators have been developed for the task of human activity recognition. They different modalities and approaches, covering the whole range of data source recorded. The multimodal HAR approach with convolutional neural networks (Gadzicki, Khamsehashari, and Zetzsche, 2020) described above (3.5.3) covers the analysis of RGB video and skeleton data. The model has been adjusted to the constraints of the annotation data in the EASE-TSD. There are different variants for the hierarchical annotation levels, covering single levels but also two levels (action and gesture levels). The model output has been changed to a multi-label prediction due to the possibility of multiple labels being simultaneously present. The output is a probability distribution over all labels which is usually not required for other human activity datasets[1].

Apart from the multimodal CNN approach there are more annotators for speech recognition or HAR from biosignals. Together they allow to automatically annotate data based on all available modalities.

The manual and automatic annotations must be transformed into a NEEM-specific format which allows them to be stored, queried and visualized in the OpenEASE database. The NEEM format agrees with the ontological structure used by the cognitive robot, thus enabling the access and processing of human activity data. The OpenEASE database can also be used by other researchers who might want to use the recorded data for development of models or testing of their models.

To our knowledge this is the first attempt of developing a pipeline for the transfer of human activity data to a robotic system with integrated automatic annotators. There are more datasets which address a similar setting of kitchen related activities performed by humans.

[1]Datasets including multiple labels per sample are common for object recognition tasks though.

"EPIC-Kitchens" (Damen et al., 2018) features video from a head-mounted camera recorded during the activities performed by 32 subjects at home. "50 Salads" (Stein and McKenna, 2013) features recordings of meal preparations and includes RGB+D video and accelerometer data from sensors mounted on the objects. "MPII Cooking Activities Dataset" (Rohrbach et al., 2012) provides video recordings for 65 kitchen activities. "TUM Kitchen Dataset" (Tenorth, Bandouch, and Beetz, 2009) provides video, fullbody motion capture data, RFID tag and magnetic sensor readings from objects and the environment.

3.7 Contribution

In Khamsehashari, Gadzicki, and Zetzsche, 2019 we showed that simplified residual units are not particularly suitable for very deep networks, but offer good performance at relatively shallow depths.

Our investigation of late fusion for human activity recognition with convolutional neural networks (Gadzicki, Khamsehashari, and Zetzsche, 2018) that late fusion does not necessarily improve the performance if only the last layer of the network is trained. Our results were dominated by the skeleton CNN which achieved state-of-the-art performance and delivered best cross-subject recognition accuracy at the time of publishing.

In our paper Gadzicki, Khamsehashari, and Zetzsche, 2020 we have investigated early fusion of modalities in comparison to late fusion and unimodal processing. We could show that early fusion performed better than the other variants. Early fusion with convolutional neural networks has been barely investigated in the past. It offers novel insights on fusion of CNNs and shows an interesting direction for future development of multimodal recognition with convolutional neural networks.

The work on a pipeline for transfer of knowledge from human subjects to robots (Mason, Gadzicki, et al., 2020) is unique in its complexity and scope, being "the first to combine multimodal data collection, hierarchical and semantic annotations, and ontological reasoning to enhance cognitive robots" (Mason, Gadzicki, et al., 2020). Our contribution of an automated annotator for multimodal data represents an important puzzle piece for making data recordings of human activities processable by robots.

CHAPTER 4

Sensorimotor Perception and Navigation

The computational models covered in Chapter 2 and 3 are bottom-up, data driven. The models based on sensorimotor processing consist of a bottom-up part, but additionally features a top-down, knowledge driven component.

In this chapter the basics of sensorimotor theory are covered in Section 4.1, the computational models in Section 4.2, and the application to Space exploration in Section 4.3.

4.1 Sensorimotor Theory

4.1.1 Saccadic Eye Movement

As mentioned in 2.1 the human eye has only a small area of high acuity, the fovea. To perceive parts of our environment with high resolution (and in color) the fovea has to be directed at specific points within our field of view. This is achieved with saccadic eye movements which are extremely fast, abrupt movements. They are ballistic in nature, i.e., once initiated the trajectory cannot be altered. If a saccade does not hit the intended target, a corrective saccade follows up very briefly after the first one (Palmer, 1999, pp. 523–524).

Saccades take 150–200ms to plan and execute, the actual ballistic eye movement takes typically 30ms and reaches speeds of up to 900° per second (Goldberg, Eggers, and Gouras, 1991). The eye fixates the region of interest for a variable amount of time, on average 300ms per fixation, giving the visual system time to process the information (Palmer, 1999, p. 523). Saccadic eye movement is not consciously perceived, even though the image motion during a saccade is theoretically perceivable due to the pupil being open. It has been shown that large moving objects at high velocity are perceivable (Burr and Ross, 1982) which would also apply to the image during a saccade. The visual system does not provide information during a saccade. This phenomenon is called saccadic suppression (Palmer, 1999, p. 523).

4.1.1.1 Saccadic Exploration

To fully explore a scene, an observer has to perform a large number of saccades. Due to the limited size of the visual field covered by the fovea, the information gained from a single fixation is low in relation to all the information available (Palmer, 1999, pp. 528–530). For efficient exploration, a guidance system is necessary. Studies by Yarbus, 1967 have shown that human observers fixate regions of interest (e.g. faces, objects). However, fixation points

cannot be estimated from the structure of the image alone, they are task-dependent as well. Noton and Stark, 1971 identified recurring sequences of fixations if one observer viewed an image multiple times (with delay between trials). Though there was variation, and fixation locations between observers did not agree, for a single observer they agreed for 65% of the images.

These findings suggest that there is a) a mechanism to extract potentially relevant fixation locations from the overall scene and b) a top-down cognitive mechanism selecting interest points depending on a high-level task.

4.1.2 Sensorimotor Contingency Theory

In vision science it is generally accepted that the internal representation consist of multiple channels, decomposing the visual sensory information according to specific features (Ginsburg, 1986; Wiesel and Hubel, 1963; Hubel and Wiesel, 1977). It implies that there is a neural image (Robson, 1981) integrating the visual information into a coherent perception. There is also evidence for regions of neurons representing multiple spatial resolutions in the visual areas in a graded fashion (Everson et al., 1998). Computational models reflect this view by featuring multiple scale and orientation channels as in the HVS model or a set of feature maps in CNNs.

The sensorimotor contingency (or dependency) theory (O'Regan and Noë, 2001) states that there is no such neural image which integrates the details into a spatially consistent representation. Instead the external world serves as the representation (O'Regan, 1992) and vision is treated as an exploratory activity. During exploration, the sensory information do not change randomly but in a way governed by the modality. When an eye moves around the projection of light on the retina changes in specific ways, e.g., the projection of a line might change from a straight line (fovea directed at the line) to a curved line in the periphery (fovea directed above the line). These shifts and distortions that appear during eye movement are particular to the modality, defining the sensorimotor dependencies between the exploratory action and the change in sensation. The visual information depends on the properties of an object (size, shape, texture, or color) and its position in space (distance and angle to the observer). By changing position and viewing angle the observer can effectively sample the object's properties. The sensory information will be distorted in a lawful way, basically defining a property by the dependency of distortion and action. The brain will have to abstract from the actual changes which are infinite in number while moving around and describe them in a set of laws which then can be used to code specific object properties. Visual perception is how features or cortical representation changes when movements are undertaken, but the observer must have mastered the laws governing the sensorimotor contingencies of the modality and actively perform the mastery (O'Regan and Noë, 2001).

Different modalities are governed by their respective sensorimotor dependencies. For the auditory system walking will produce different sounds based on the ground (and footwear) material while for the tactile system the amount of friction perceived when rubbing a surface will define the material's properties. Walking or rubbing are already abstracted actions as for the biological system the pattern of muscle fiber activations through the nervous system is most likely the encoding of an action. Proprioception (Tuthill and Azim, 2018) gives information about the execution of an action and the change in sensation for a particular modality codes the properties of objects or the environment by sensorimotor dependencies. New sensorimotor contingencies can be learned, e.g. by augmentation of new sensory information (Kaspar et al., 2014).

Not only object properties but the notion of space can be inferred by sensorimotor contingencies (Philipona, O'Regan, and Nadal, 2003) without any prior knowledge about the environment or the relations between input and output. An organism equipped with sensors and actuators can learn by issuing random commands that there are sensory inputs which

it can fully control by its commands and others which show no relation to the commands. The organism can deduce that there is a body it can fully control and an environment it has only partial control of. The sensory inputs regarding the body are proprioceptive and regarding the environment exteroceptive (Kandel, J. H. Schwartz, and Jessell, 2000). With this distinction the organism can try to understand the environment by monitoring changes in exteroceptive sensor readings due to motor commands. Certain transformations of the sensory input will depend on the order of commands issues. From these rigid transformations the organism can deduce geometry and the properties of the body-environment system. Finally, the organism is able to navigate the environment with a set of laws governing the relations between sensory input and motor commands expressed as sensorimotor contingencies. Discovering the invariants in the sensorimotor laws enables the emergence of the notion of space (Philipona, O'Regan, and Nadal, 2003).

Rachuy, 2020 proposed an algorithm for bootstrapping of mobile agents by processing of sensorimotor interactions with its environment. The agent assigned geometric interpretations to its motor actions utilizing Lie groups as a representation of geometric operations.

Ecological (Gibson, 1979) and Active Perception (Bajcsy, 1988; Ballard, 1991) are closely related theories stressing the importance of actions for perception. For visual perception Gibson, 1979 states that the environment affords viewing sensations, inviting an organism to explore the environment. In active perception Ballard, 1991 sees an advantage for an organism to execute behaviors based on visual information if the visual sensors are actively controlled. The active control of actions are based on expectations about the outcome (Bajcsy, Aloimonos, and Tsotsos, 2018). For ecological and active perception, the actions are means to acquire new sensory information while the action itself does not necessarily contribute to the representation of the environment. In contrast the sensorimotor contingency theory regards action as integral part of the representation, contributing information of same importance as the sensory information (O'Regan and Noë, 2001).

4.2 Computational Models of Sensorimotor Processing

Although sensorimotor models have been extensively tested in psychology with human subjects as well as other organisms (see Gallivan et al., 2018; H. E. Kim, Avraham, and Ivry, 2021 for a review) and in medicine (Hayes et al., 2018), computational models are scarce in comparison.

Kuipers and Levitt, 1988 proposed a four-level spatial semantic hierarchy, consisting of sensorimotor interaction, procedural behaviors, topological and metric mapping, as a cognitive map representation for navigation and mapping. The sensorimotor level describes the input-output relations between the agent and the environment. It is defined as a sequence of view-action pairs. Kuipers and Levitt, 1988 furthermore describe three simulated systems for indoor and outdoor scenarios which implement the spatial hierarchy. The simulated sensory information is highly abstract, e.g., the "vision" in one system is the number of visible landmarks.

4.2.1 Visual Perception

Schill et al., 2001 suggested a model motivated by the saccadic exploration process in human vision for scene recognition. This system actively explores an image by performing saccades represented as sensorimotor features. A sensorimotor feature SMF is a triple

$$SMF := (s_{t-1}, m_{t-1}, s_t) \tag{4.1}$$

with s_{t-1} being the sensory input in the location at the last time step, m_{t-1} the motor action performed at the last time step in order to reach the new location which provides the sensory

input s_t at the current time step. The sensory inputs are locally limited, mimicking the limited visual angle covered by the fovea. The motor action corresponds to the shift of gaze during a saccade.

The potential fixation locations are points with the intrinsically two-dimensional signal (i2D) which yield the least redundant information and correspond well with fixation location by human observers (Zetzsche, Schill, et al., 1998). The extraction is performed by a nonlinear i2D-selective operator (Zetzsche and Krieger, 2001) which is applied to the image. The feature descriptor is extracted from linear orientation filters in the local image area. A scene can now be represented by the potential saccades between fixation locations, or in other words the set of sensorimotor features present in the scene. During supervised learning, after performing a saccade on an image, the sensorimotor feature is considered as evidence for a particular hypothesis (or class) and stored in the knowledge base. By performing saccades on an entire training set with labels a frequency distribution of sensorimotor features for every sample can be obtained. The system uses Dempster-Shafer Theory of Evidence (DS) (Dempster, 1967; Shafer, 1976) for assigning belief mass to a set of hierarchical hypotheses which constitute the knowledge base. The DS theory offers the advantage of representing uncertainty more explicitly than with probabilities.

The system behavior is guided by Inference by Information Gain (IBIG) (Schill, 1997) which is an adaptive strategy suited for situations of incomplete or inconsistent data. Within the sensorimotor framework the idea is to select an action which has the highest expected gain in information with respect to the current belief state. Figure 4.1 illustrates this process. After performing a saccade and receiving new sensory data, the current belief state is updated.

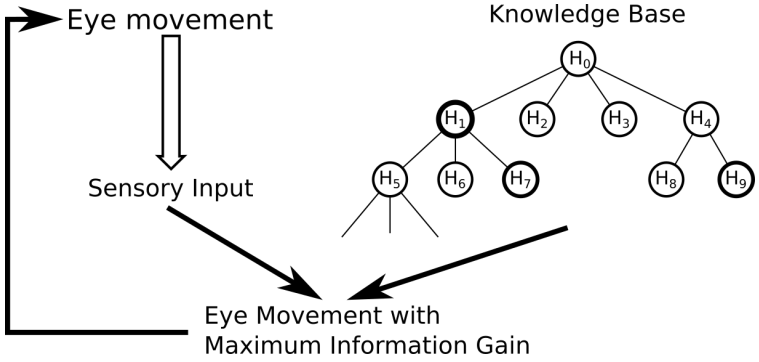


Figure 4.1: The information gain is calculated based on the current sensory input and the belief distribution in the hierarchy. The next eye movement is selected based on the maximum information gain (Adapted from Schill et al., 2001).

The next action is selected by calculating the difference between the current and the potential belief distribution after performing an action.

IBIG relies on a hierarchical representation of the hypothesis space where all leaf nodes are objects (or finest object classes) and non-leaf nodes represent more abstract classes. While leaf nodes are generally associated with the full set of sensorimotor features of an object or scene, the non-leaf nodes carry a subset which is common to the subtree. The hierarchical structure has practical considerations, reducing the computational complexity of inference in Dempster-Shafer Theory (Gordon and Shortliffe, 1985), but is also motivated by the organization of human brain functions. For the visual system it generally believed that the visual cortex is organized hierarchically, abstracting from primitive patterns in the lower layers to complex, abstracted features in the higher layers (Oram and Perrett, 1994). P. Taylor et al., 2015 showed in a data-driven analysis that cognitive functions are hierarchically ordered, forming a continuum of functions (Figure 4.2). “ From tangible sensory inputs, symbolic content may

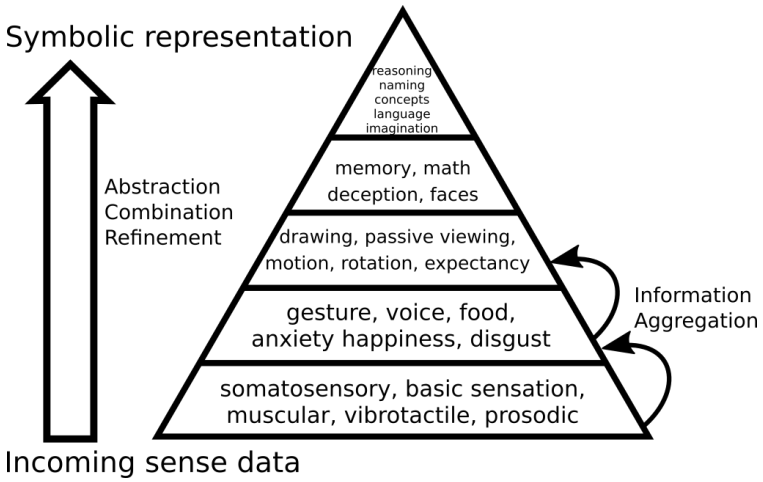


Figure 4.2: Data-based, objective pyramid of cognition showing an oversimplified graphical model of the information representation flow from sensory inputs (bottom) to abstract representations (top) in human cortex. (Adapted from P. Taylor et al., 2015)

progressively emerge as information is processed deeper into the brain's structural network, starting with inputs, and expanding in abstraction or refinement, resulting in intangible or deep symbolic content in the structural pinnacle of the human brain network" (P. Taylor et al., 2015).

4.2.1.1 Unsupervised Learning of Sensorimotor Hierarchies

This section summarizes our work published in "Hierarchical Clustering of Sensorimotor Features".

In the paper (Gadzicki, 2009) the author describes the approach to generate a hierarchy of sensorimotor feature in unsupervised fashion. This was motivated by the fact that the hierarchical hypotheses space described above was hand-crafted. A scene is represented by full set of possible sensorimotor features, representing a fully-connected graph with potential fixation locations as nodes and actions as edges. Since sensorimotor features are directional each pair of fixation locations is represented by two sensorimotor features. The sensorimotor features are abstracted through clustering with a Self-Organizing Map (SOM) (Kohonen, 1990; Kohonen, 2001) which is an artificial neural network with the property of creating a spatially organized representation of features. During training of the SOM, it is fed with all extracted sensorimotor features from the training set. The resulting representation is spatially organized on a 2D-grid where similar features are mapped to the same node or a neighboring one. After training of the SOM, a scene can now be represented by the frequency count of activations of the SOM with the sensorimotor features of a scene being the input. This procedure is illustrated in Figure 4.3.

The hierarchical representation was generated with agglomerative clustering methods with the frequency distribution of activations serving as feature vectors of the samples. Starting with singleton clusters for every object and applying Ward's rule (Ward, 1963) for merging of a pair of clusters, a dendrogram is build. The Tanimoto coefficient (Tanimoto, 1958) was used for calculating the similarity of vectors.

This approach works well the training sets with a certain overlap. For instance, for an hierarchy of objects it was beneficial to have different views of the objects with the overlapping views.

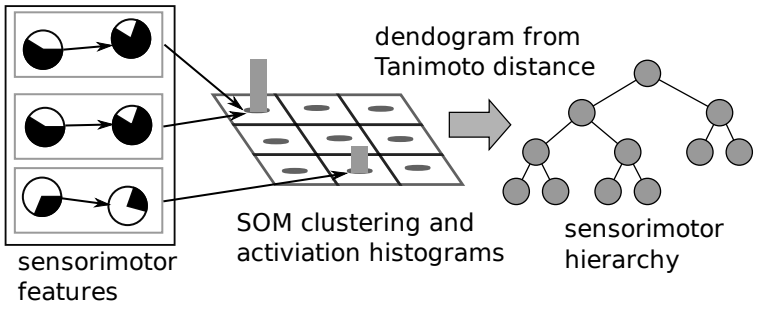


Figure 4.3: A set of sensorimotor features representing a scene is passed through a Self-Organizing Map, resulting in a frequency distribution of the activations (Source: Reineking et al., 2010).

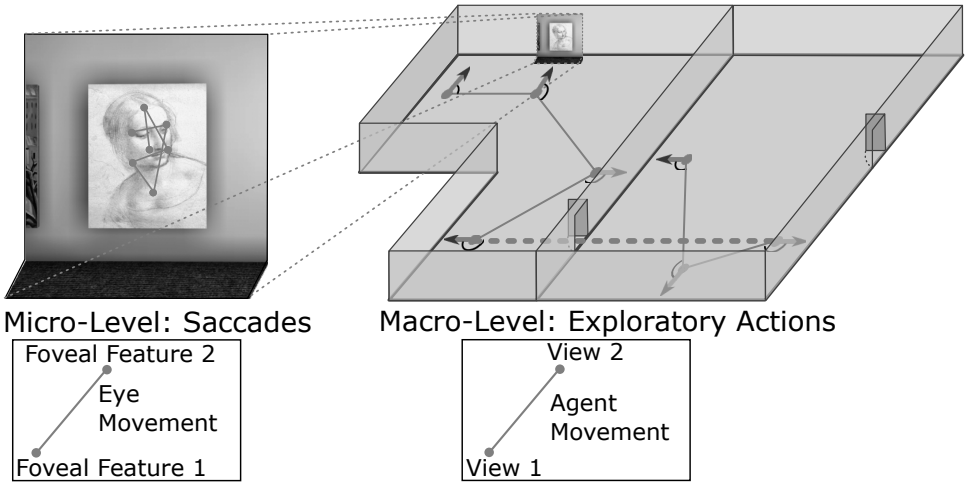


Figure 4.4: Two levels of hierarchical sensorimotor representations (Source: Reineking et al., 2010).

4.2.2 Localization

This section summarizes our work published in *"Bio-inspired Architecture for Active Sensorimotor Localization"*.

The system design from Schill et al., 2001 was extended to localization in a virtual environment in our paper (Reineking et al., 2010), integrating the works from (Gadzicki, 2009; Reineking, 2011; Zetzsche, Wolter, and Schill, 2008; Schill et al., 2001). The simulated agent represents its environment based on sensorimotor features on a micro level for perception and on a macro level for localization and navigation. On the micro level the agent analyzes scenes by performing saccades on the sensory input acquired from the environment. On the macro level the agent explores the environment by performing actions between positions in the environment, obtaining views of the scene. The sensory information is given by the visual input at starting and end locations of the motor action. The motor action is defined by a rotation, a translation and again a rotation. Figure 4.4 illustrates the two levels of sensorimotor features.

Micro and macro level have their own hierarchical hypotheses space and are generated by agglomerative clustering (Gadzicki, 2009) in unsupervised fashion. For the micro level hypotheses space, views are sampled in the environment with individual views being the leaf nodes of the hierarchy and classes of views being the non-leaf nodes. For the macro level

the leafs represent individual rooms and non-leaf nodes room classes. For each node of both, micro and macro level hierarchies, a numerical representation of the sensorimotor features based on the SOM activations is stored.

The selection of the next action is based on the uncertainty minimization strategy from IBIG (Schill, 1997). Given the current belief state and a set of potential actions the local conflict uncertainty measure is used to select the action with the minimum expected uncertainty. After the execution of the action the belief state is updated with the actual feature observed.

4.2.3 Realization with Physical Hardware

A computational model of Sensorimotor Contingency (SMC) theory was implemented on a simple robot (Maye and Engel, 2011; Maye and Engel, 2012). Their approach is based on the assumption that "learning SMCs corresponds to determining the conditional probability of making a sensory observation given an action and a context" (Maye and Engel, 2011) and can be expressed with a Markov model. The system utilize a value system (Maye and Engel, 2011) or utility function (Maye and Engel, 2012) to guide the robot's behavior. The sensorimotor features (or SMCs in these works) are represented by chaining of motor actions and sensory observations which are stored in a tree structure. By exploring the environment, the robot acquires a collection of SMCs describing the environment and enabling the robot to predict the outcome of subsequent actions.

In Högman, Björkman, and Kragic, 2013 a robot arm learns sensorimotor contingencies through pushing of objects. The model uses a probabilistic representation based on Gaussian Process regression. The sensorimotor features are represented by a function of a motor action and previous sensory data (position and orientation of the object) mapping to new sensory data after the action was executed. The system learned to successfully classify objects by applying optimal action selection minimizing the conditional entropy of the class given an action and random outcome.

In Nakath, Kluth, et al., 2014; Kluth et al., 2015 an approach for active sensorimotor object recognition with an robotic arm is proposed. The system can use a robotic arm to move around the object and inspect it from different viewpoints physically but also a simulated arm in VR. It is also possible to feed images from a dataset. The system uses probabilistic sensorimotor features as the representation. The feature descriptor for sensory input is either GIST (Oliva and Torralba, 2006) or SURF (Bay, Tuytelaars, and Van Gool, 2006) features, depending on whether the robotic arm is used or the image feed from a dataset. The features are assigned to clusters stemming from previously learned k-means clustering. The resulting sensorimotor features are processed by a classifier which computes the belief regard the class of the object. The motor actions are selected bases on the uncertainty of the belief state measured as entropy. The optimal action is based on the expected maximum information gain, expressed as expected entropy.

Lanillos, Dean-Leon, and Cheng, 2017 suggested an approach to self-perception in robots, enabling the robot to understand changes in sensory information it perceives. By combining multisensory information from visual, proprioceptive, and tactile sensors with probabilistic reasoning the robot learns to distinguish between cues inside and outside its body. Via sensory contingencies the robot discovers usable objects and how to interact with them.

4.3 Application to Space Exploration

> This section summarizes our publication *"KaNaRiA: Identifying the Challenges for Cognitive Autonomous Navigation and Guidance for Missions to Small Planetary Bodies"*.

Cognitive principles to perception and reasoning can be applied beyond earth as shown in our paper on asteroid mining (Probst et al., 2015). The goal of the project was to develop new approaches for autonomous navigation of spacecraft during deep space missions, e.g., exploration of the main asteroid belt. During deep space missions spacecraft operate at distances to earth which do not allow for real time communication. Time critical parts of missions must be carefully controlled by predefined procedures, anticipating possible problems which might occur during execution of a mission. Autonomous navigation in many mission phases is an approach to overcome these shortcomings. The system was realized as a simulated mission to the asteroid belt consisting of several mission phases covering transfer from parking orbit up to the landing on an asteroid. The overall system architecture was inspired by cognitive agent systems which follow objectives while holding and updating a belief over the current state of the environment. The approach to decision-making was governed by integration of top-down knowledge (*a-priori* knowledge about the environment, e.g. layout of the spacecraft, orbital elements of celestial bodies) and bottom-up knowledge (fused sensory information) in the spirit of works on sensorimotor perception (Schill et al., 2001) and navigation (Reineking et al., 2010).

Various subsystems were designed based on information maximization and active perception as leading principles. Optimal rotation sequences for pointing of navigation instruments towards celestial bodies were suggested based on active perception (Nakath, Rachuy, et al., 2016; Nakath, Clemens, and Schill, 2018). In Nakath, Clemens, and Rachuy, 2020 an active graph based simultaneous localization and mapping (SLAM) in the vicinity of an asteroid was proposed. This approach utilizes the navigation sensors of the spacecraft, an inertial measurement unit (IMU), a star tracker, and LiDAR to estimate the state of the spacecraft relative to the map estimate of the asteroid. Reducing the uncertainty about the state, e.g., the orbit trajectory, and the map, e.g., the asteroid's surface, are conflicting goals. The active perception approach controls the information and localization gain, balancing the conflicting goals, by choosing appropriate orbit trajectories.

4.4 Contribution

In Reineking et al., 2010 we demonstrated that the sensorimotor principle can be applied to different levels of granularity for both, perception and localization. It offers a psychologically and neurobiologically plausible approach for recognition and navigation tasks. By utilizing the information gain principle, the system is able to solve these tasks efficiently. We showed, furthermore, that the hierarchical representation which is necessary for efficient inference can be learned in an unsupervised fashion (Gadzicki, 2009).

Active perception and information gain principles can be applied to navigational tasks in Space exploration (Probst et al., 2015). The approaches suggested here are novel to my best knowledge and show that methods inspired by biological perception can be applied to complex systems like spacecrafts. The concept of using sensory information together with motor actions for representation and information maximization for inference can be virtually applied to any agent operating in any environment.

CHAPTER 5

Conclusion and Outlook

This dissertation presents three different perspectives on biologically inspired pattern recognition. One perspective is represented by computational models of human visual system which implements known properties of neurons from the visual cortex. Another perspective is given by artificial neural networks inspired by the interconnected nature of the cells in the brain, which perform complex operations utilizing simple units. Sensorimotor processing represents the last view combining bottom-up feature extraction with top-down information gain driven processing.

Models of the human visual system are a very direct realization of biologically inspired vision. We have shown that our model was able to predict subjective assessment of streak distortions in printings over the full impairment scale. For streak distortions at the threshold of perception a model predicting a broad range of scales is of limited use. Here a possible direction of development is the modification of the HVS model to predict detection of distortions. The prediction of threshold perception can be modeled well with such a model, but requires respective data from human observers for tuning of the model.

We have proposed a neuro-biologically plausible method for estimation of frequency distribution and auto- and cross-correlation functions. This has been done by utilizing the gain control function known from neurons in the visual pathway. As a next step the neural frequency distributions could be used for vision models utilizing statistical information, e.g., models of peripheral vision. They would benefit with regard to plausibility by utilizing our proposed statistical functions.

Artificial neural network are inspired by parallel distributed processing and connectionist ideas of data processing with a population of simple units, inspired by the brain and its construction from neurons. While convolutional neural networks are inspired by human visual system, they can serve as models for other modalities as well. We have proposed convolutional neural networks for activity recognition. Our focus was on multimodal CNNs with different fusion methods, showing that early fusion performed better than the classical late fusion. Future developments can be the addition of more modalities which might be particularly interesting for early fusion, leading to an increased performance for the recognition task. A challenge here is alignment and synchronization of modalities, especially those with different dimensionality, e.g., video and skeleton data. Another direction could be spatio-temporal interest operators, providing better information about the temporal changes of input data. Furthermore developments from the image domain could be adapted for HAR, e.g., attention networks (Xu et al., 2015) or transformers (Dosovitskiy et al., 2020), and investigated with

regard to multimodal extensions.

Our system for human activity recognition has been integrated in a human to robot pipeline with the long term goal of transferring knowledge from human subjects performing everyday activities to robots. In this first stage the HAR module served as a recognizer for automated labeling. As a next step this approach can become part of generative models of human activities, abstracting from individual human demonstrations and generalizing activities of different levels of granularity.

Finally sensorimotor representations are suitable for computational models of perception and localization. While HVS models and ANN are bottom-up driven approaches, our sensorimotor approaches include a top-down cognitive component. Inference utilizing information gain has been shown to work efficiently for perception tasks using visual information and localization in spatial environments. Based on these principles it is possible to design navigation and localization methods for spacecrafts.

CHAPTER 6

Publications

6.1 Peer-reviewed publications

The following list contains the peer-reviewed papers.

- Gadzicki, K.: **Hierarchical Clustering of Sensorimotor Features**. In *KI 2009: Advances in Artificial Intelligence* Vol. 5803 Lecture Notes in Artificial Intelligence (2009).
  My share: 100%
  This paper is based on my Diploma thesis.

- Reineking, T. and Wolter, J. and Gadzicki, K. and Zetzsche, C.: **Bio-inspired Architecture for Active Sensorimotor Localization**. In *Spatial Cognition VII* (2010), pp.163–178.
  My share: 20%
  I have implemented and integrated the hierarchical clustering, and contributed to the corresponding sections of the manuscript.

- Gadzicki, K. and Zetzsche, C.: **Prediction of the Perceived Quality of Streak Distortions in Offset-Printing with a Psychophysically Motivated Multi-channel Model**. In *Tagungsband 18.Workshop Farbbildverarbeitung 2012*, Darmstadt (2012).
  My share: 90%
  I have implemented and evaluated the model, and written most of the manuscript.

- Gadzicki, K. and Zetzsche, C.: **Prediction of the perceived quality of streak distortions in offset-printing with a psychophysically motivated multi-channel model**. In *Journal of Modern Optics*, 60.14 (2013), pp. 1167–1175.
  My share: 90%
  I have implemented and evaluated the model, and written most the manuscript.

- Zetzsche, C. and Gadzicki, K. and Kluth, T.: **Statistical Invariants of Spatial Form: From Local AND to Numerosity**. In *Proceedings of the Second Interdisciplinary Workshop The Shape of Things* (2013), pp. 163–172.
  My share: 30%
  My part was the implementation and evaluation of the frequency distributions and statistical functions with neurons. I have contributed to these parts of the paper.

- Probst, A., González Peytaví, G., Nakath, D., Schattel, A., Rachuy, C., Lange, P., Clemens, J., Echim, M, Schwarting, V., Srinivas, A., Gadzicki, K., Förstner, R., Eissfeller, B., Schill, K., Büskens, C. and Zachmann, G.: **KaNaRiA: Identifying the Challenges for Cognitive Autonomous Navigation and Guidance for Missions to Small Planetary Bodies**. In *66th International Astronautical Congress (IAC)*, Jerusalem, 2015.
  My share: 5%
  I was the project manager and coordinator of this project. I have guided the general direction of the development, contributing to concepts and discussions on practically all areas of the project. I have contributed to the manuscript in introduction and general description of the project.

- Gadzicki, K. and Khamsehashari, R. and Zetzsche, C.: **Multimodal Convolutional Neural Networks for Human Activity Recognition**. In *IROS 2018: Workshop on Latest Advances in Big Activity Data Sources for Robotics & New Challenges*, Madrid (2018).
  My share: 85%
  I have designed, implemented and evaluated the models. I have written most of the manuscript.

- Khamsehashari, R. and Gadzicki, K. and Zetzsche, C.: **Deep Residual Temporal Convolutional Networks for Skeleton-Based Human Action Recognition**. In *ICVS 2019: Computer Vision Systems* (2019), pp. 376–385.
  My share: 20%
  I have contributed to the concept of the model. I have contributed to the manuscript.

- Gadzicki, K. and Khamsehashari, R. and Zetzsche, C.: **Early vs Late Fusion in Multimodal Convolutional Neural Networks**. In *2020 IEEE 23rd International Conference on Information Fusion (FUSION)*, virtual event (2020).
  My share: 90%
  I have designed, implemented and evaluated the models. I have written the manuscript.

- Mason, C. and Gadzicki, K. and Meier, M. and Ahrens, F. and Kluss, T. and Maldonado, J. and Putze, F. and Fehr, T. and Zetzsche, C. and Herrmann, M. and Schill, K. and Schultz, T.: **From Human to Robot Everyday Activity**. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems IROS 2020*, Las Vegas (2020).
  My share: 20%
  I have contributed to the automated annotators with HAR by multimodal CNNs. I have written the corresponding parts of the manuscript, and contributed to the introduction, general description of the pipeline and conclusion.

6.2 Extended Abstracts

- Zetzsche, C., Rosenholtz, R., Cheema, N., Gadzicki, K., Fridman, L., Schill, K.: **Neural Computation of Statistical Image Properties in Peripheral Vision**. In *Computational and Mathematical Models in Vision (MODVIS)*, 2017.
  My share: 45%
  My part was the implementation and evaluation of the frequency distributions and statistical functions with neurons.

Bibliography

Adelson, E. and J. Bergen (1985). „Spatiotemporal energy models for the perception of motion.“ In: *J. Opt. Soc. Am. A* 2.2, pp. 284–99. ISSN: 0740-3232.

Aggarwal, J. and M. S. Ryoo (Apr. 2011). „Human Activity Analysis: A Review“. In: *ACM Comput. Surv.* 43.3. ISSN: 0360-0300. DOI: 10.1145/1922649.1922653.

Aguileta, A. A., R. F. Brena, O. Mayora, E. Molino-Minero-Re, and L. A. Trejo (2019). „Multi-Sensor Fusion for Activity Recognition—A Survey“. In: *Sensors* 19.17. ISSN: 1424-8220. DOI: 10.3390/s19173808. URL: https://www.mdpi.com/1424-8220/19/17/3808.

Albrecht, D. and D. Hamilton (July 1982a). „Striate cortex of monkey and cat: contrast response function“. In: *J Neurophysiol* 48.1, pp. 217–237.

– (1982b). „Striate Cortex of Monkey and Cat: Function Contrast Response“. In: *J Neurophysiol* 48.1, pp. 217–237.

Arandjelović, R., P. Gronat, A. Torii, T. Pajdla, and J. Sivic (2018). „NetVLAD: CNN Architecture for Weakly Supervised Place Recognition“. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40.6, pp. 1437–1451. DOI: 10.1109/TPAMI.2017.2711011.

Atrey, P. K., M. A. Hossain, A. El Saddik, and M. S. Kankanhalli (Nov. 2010). „Multimodal fusion for multimedia analysis: a survey“. In: *Multimedia Systems* 16.6, pp. 345–379. ISSN: 1432-1882. DOI: 10.1007/s00530-010-0182-0.

Bajcsy, R. (1988). „Active perception“. In: *Proceedings of the IEEE* 76.8, pp. 966–1005. DOI: 10.1109/5.5968.

Bajcsy, R., Y. Aloimonos, and J. K. Tsotsos (Feb. 2018). „Revisiting active perception“. In: *Autonomous Robots* 42.2, pp. 177–196. ISSN: 1573-7527. DOI: 10.1007/s10514-017-9615-3.

Balas, B., L. Nakano, and R. Rosenholtz (2009). „A summary-statistic representation in peripheral vision explains visual crowding.“ In: *J Vis* 9.12, pp. 13.1–18. ISSN: 1534-7362. DOI: 10.1167/9.12.13.

Ballard, D. H. (1991). „Animate vision“. In: *Artificial Intelligence* 48.1, pp. 57–86. ISSN: 0004-3702. DOI: https://doi.org/10.1016/0004-3702(91)90080-4. URL: https://www.sciencedirect.com/science/article/pii/0004370291900804.

Baltrušaitis, T., C. Ahuja, and L. Morency (Feb. 2019). „Multimodal Machine Learning: A Survey and Taxonomy“. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41.2, pp. 423–443. ISSN: 1939-3539. DOI: 10.1109/TPAMI.2018.2798607.

Barlow, H. and D. Berry (July 2011). „Cross- and auto-correlation in early vision.“ In: *Proceedings. Biological sciences / The Royal Society* 278.1714, pp. 2069–75. ISSN: 1471-2954. DOI: 10.1098/rspb.2010.2170.

Bay, H., T. Tuytelaars, and L. Van Gool (2006). „SURF: Speeded Up Robust Features“. In: *Computer Vision – ECCV 2006*. Ed. by A. Leonardis, H. Bischof, and A. Pinz. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 404–417. ISBN: 978-3-540-33833-8.

Beddiar, D. R., B. Nini, M. Sabokrou, and A. Hadid (Nov. 2020). „Vision-based human activity recognition: a survey“. In: *Multimedia Tools and Applications* 79.41, pp. 30509–30555. ISSN: 1573-7721. DOI: 10.1007/s11042-020-09004-3.

Bellot, D., A. Boyer, and F. Charpillet (2002). „A new definition of qualified gain in a data fusion process: application to telemedicine“. In: *Proceedings of the Fifth International Conference on Information Fusion. FUSION 2002.* Vol. 2, pp. 865–872.

Bobick, A. F. and J. W. Davis (2001). „The recognition of human movement using temporal templates“. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 23.3, pp. 257–267. DOI: 10.1109/34.910878.

Briscoe, R. E. (2016). „Multisensory Processing and Perceptual Consciousness: Part I“. In: *Philosophy Compass* 11.2, pp. 121–133. DOI: https://doi.org/10.1111/phc3.12227.

Burr, D. C. and J. Ross (1982). „Contrast sensitivity at high velocities“. In: *Vision Research* 22.4, pp. 479–484. ISSN: 0042-6989. DOI: https://doi.org/10.1016/0042-6989(82)90196-1. URL: https://www.sciencedirect.com/science/article/pii/0042698982901961.

Burt, P. and E. Adelson (1983). „The Laplacian Pyramid as a Compact Image Code“. In: *IEEE Transactions on Communications* 31.4, pp. 532–540. ISSN: 0096-2244. DOI: 10.1109/TCOM.1983.1095851.

Calderón, A., S. Roa, and J. Victorino (2003). „Handwritten Digit Recognition using Convolutional Neural Networks and Gabor filters“. In: *International Congress on Computational Intelligence.*

Campbell, F. W. and J. G. Robson (Aug. 1968). „Application of Fourier analysis to the visibility of gratings“. eng. In: *The Journal of physiology* 197.3. PMC1351748[pmcid], pp. 551–566. ISSN: 0022-3751. DOI: 10.1113/jphysiol.1968.sp008574. URL: https://pubmed.ncbi.nlm.nih.gov/5666169.

Carandini, M. and D. Heeger (July 2012). „Normalization as a canonical neural computation“. In: *Nature Reviews Neurosci* 13, pp. 51–62.

Carreira, J. and A. Zisserman (July 2017). „Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset“. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4724–4733. DOI: 10.1109/CVPR.2017.502.

Chandrakala, S. and S. L. Jayalakshmi (June 2019). „Environmental Audio Scene and Sound Event Recognition for Autonomous Surveillance: A Survey and Comparative Studies“. In: *ACM Comput. Surv.* 52.3. ISSN: 0360-0300. DOI: 10.1145/3322240.

Choi, J., Y.-i. Cho, T. Han, and H. S. Yang (2008). „A View-Based Real-Time Human Action Recognition System as an Interface for Human Computer Interaction“. In: *Virtual Systems and Multimedia.* Ed. by T. G. Wyeld, S. Kenderdine, and M. Docherty. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 112–120. ISBN: 978-3-540-78566-8.

Cortes, C. and V. Vapnik (Sept. 1995). „Support-vector networks“. In: *Machine Learning* 20.3, pp. 273–297. ISSN: 1573-0565. DOI: 10.1007/BF00994018.

Cybenko, G. (Dec. 1989). „Approximation by superpositions of a sigmoidal function“. In: *Mathematics of Control, Signals and Systems* 2.4, pp. 303–314. ISSN: 1435-568X. DOI: 10.1007/BF02551274.

D'mello, S. K. and J. Kory (Feb. 2015). „A Review and Meta-Analysis of Multimodal Affect Detection Systems“. In: *ACM Comput. Surv.* 47.3. ISSN: 0360-0300. DOI: 10.1145/2682899.

Dalal, N. and B. Triggs (2005). „Histograms of oriented gradients for human detection“. In: *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05).* Vol. 1, 886–893 vol. 1. DOI: 10.1109/CVPR.2005.177.

Dalal, N., B. Triggs, and C. Schmid (2006). „Human Detection Using Oriented Histograms of Flow and Appearance“. In: *Computer Vision – ECCV 2006.* Ed. by A. Leonardis, H. Bischof, and A. Pinz. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 428–441. ISBN: 978-3-540-33835-2.

Daly, S. (1993). „The visible differences predictor: an algorithm for the assessment of image fidelity“. In: *Digital images and human vision*. Ed. by A. B. Watson. Cambridge, MA, USA: MIT Press, pp. 179–206. ISBN: 0-262-23171-9.

Damen, D., H. Doughty, G. M. Farinella, S. Fidler, A. Furnari, E. Kazakos, D. Moltisanti, J. Munro, T. Perrett, W. Price, and M. Wray (2018). „Scaling Egocentric Vision: The EPIC-KITCHENS Dataset“. In: *CoRR* abs/1804.02748. arXiv: 1804.02748. URL: http://arxiv.org/abs/1804.02748.

Damer, N., K. Dimitrov, A. Braun, and A. Kuijper (Oct. 2019). „On Learning Joint Multi-biometric Representations by Deep Fusion“. In: *Proceedings of the IEEE 10th International Conference on Biometrics Theory, Applications and Systems (BTAS 2019)*. Tampa, FL, USA.

Daugman, J. G. (1984). „Spatial visual channels in the Fourier plane“. In: *Vis Res* 24.9, pp. 891–910. ISSN: 0042-6989. URL: http://www.ncbi.nlm.nih.gov/pubmed/6506478.

Davoodikakhki, M. and K. Yin (2020). „Hierarchical Action Classification with Network Pruning“. In: *Advances in Visual Computing*. Ed. by G. Bebis, Z. Yin, E. Kim, J. Bender, K. Subr, B. C. Kwon, J. Zhao, D. Kalkofen, and G. Baciu. Cham: Springer International Publishing, pp. 291–305. ISBN: 978-3-030-64556-4.

De Valois, R., D. Albrecht, and L. Thorell (1982). „Spatial frequency selectivity of cells in macaque visual cortex“. In: *Vis Res* 22.5, pp. 545–559.

DeAngelis, G., J. G. Robson, I. Ohzawa, and R. Freeman (1992). „Organization of suppression in receptive fields of neurons in cat visual cortex“. In: *J Neurophysiol* 68.1, pp. 144–163.

Dempster, A. (1967). *A generalization of Bayesian inference*. Tech. rep. Harvard University, Cambridge, MS, Dept of Statistics.

Deng, J., W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei (2009). „ImageNet: A Large-Scale Hierarchical Image Database“. In: *CVPR09*.

Denker, J., W. Gardner, H. Graf, D. Henderson, R. Howard, W. Hubbard, L. D. Jackel, H. Baird, and I. Guyon (1989). „Neural Network Recognizer for Hand-Written Zip Code Digits“. In: *Advances in Neural Information Processing Systems*. Ed. by D. Touretzky. Vol. 1. Morgan-Kaufmann. URL: https://proceedings.neurips.cc/paper/1988/file/a97da629b098b75c294dffdc3e463904-Paper.pdf.

Diba, A., V. Sharma, and L. Van Gool (2017). „Deep Temporal Linear Encoding Networks“. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1541–1550. DOI: 10.1109/CVPR.2017.168.

Dosovitskiy, A., L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby (2020). *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*. arXiv: 2010.11929 [cs.CV].

Du, Y., Y. Fu, and L. Wang (Nov. 2015). „Skeleton based action recognition with convolutional neural network“. In: *2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR)*, pp. 579–583. DOI: 10.1109/ACPR.2015.7486569.

Du, Y., W. Wang, and L. Wang (2015). „Hierarchical recurrent neural network for skeleton based action recognition“. In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1110–1118.

Duta, I. C., B. Ionescu, K. Aizawa, and N. Sebe (2017). „Spatio-Temporal Vector of Locally Max Pooled Features for Action Recognition in Videos“. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3205–3214. DOI: 10.1109/CVPR.2017.341.

Enroth-Cugell, C. and J. G. Robson (Dec. 1966). „The contrast sensitivity of retinal ganglion cells of the cat“. In: *J Physiol* 187.3, pp. 517–552.

Enroth-Cugell, C., J. G. Robson, D. E. Schweitzer-Tong, and A. B. Watson (Aug. 1983). „Spatio-temporal interactions in cat retinal ganglion cells showing linear spatial summation“. In: *J Physiol* 341, pp. 279–307.

Everson, R. M., A. K. Prashanth, M. Gabbay, B. W. Knight, L. Sirovich, and E. Kaplan (1998). „Representation of spatial frequency and orientation in the visual cortex“. In: *Proceedings of the National Academy of Sciences* 95.14, pp. 8334–8338. ISSN: 0027-8424. DOI: 10.1073/pnas.95.14.8334. eprint: https://www.pnas.org/content/95/14/8334.full.pdf. URL: https://www.pnas.org/content/95/14/8334.

Falchier, A., S. Clavagnier, P. Barone, and H. Kennedy (2002). „Anatomical Evidence of Multimodal Integration in Primate Striate Cortex“. In: *Journal of Neuroscience* 22.13, pp. 5749–5759. ISSN: 0270-6474. DOI: 10.1523/JNEUROSCI.22-13-05749.2002. URL: https://www.jneurosci.org/content/22/13/5749.

Feichtenhofer, C., A. Pinz, and A. Zisserman (2016). „Convolutional Two-Stream Network Fusion for Video Action Recognition“. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1933–1941. DOI: 10.1109/CVPR.2016.213.

Feichtenhofer, C., A. Pinz, and R. P. Wildes (2016). „Spatiotemporal residual networks for video action recognition“. In: *Advances in Neural Information Processing Systems (NIPS)*, pp. 3468–3476.

– (2017). „Spatiotemporal Multiplier Networks for Video Action Recognition“. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. IEEE Computer Society, pp. 7445–7454. DOI: 10.1109/CVPR.2017.787.

Felleman, D. and D. V. van Essen (1991). „Distributed hierarchical processing in the primate cerebral cortex.“ In: *Cerebral cortex* 1 1, pp. 1–47.

Field, D. (1987). „Relations between the statistics of natural images and the response properties of cortical cells“. In: *J Opt Soc Am* 4.12, pp. 2379–94. ISSN: 0740-3232.

Foley, J. M. (June 1994). „Human luminance pattern-vision mechanisms: masking experiments require a new model“. In: *J. Opt. Soc. Am. A* 11.6, pp. 1710–1719. DOI: 10.1364/JOSAA.11.001710. URL: http://josaa.osa.org/abstract.cfm?URI=josaa-11-6-1710.

Freeman, J. and E. P. Simoncelli (2011). „Metamers of the ventral stream.“ In: *Nature neuroscience* 14.9, pp. 1195–1201. ISSN: 1546-1726. DOI: 10.1038/nn.2889.

Fu, B., N. Damer, F. Kirchbuchner, and A. Kuijper (2020). „Sensing Technology for Human Activity Recognition: A Comprehensive Survey“. In: *IEEE Access* 8, pp. 83791–83820.

Fukushima, K. (Sept. 1975). „Cognitron: A self-organizing multilayered neural network“. In: *Biological Cybernetics* 20.3, pp. 121–136. ISSN: 1432-0770. DOI: 10.1007/BF00342633.

– (Apr. 1980). „Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position“. In: *Biological Cybernetics* 36.4, pp. 193–202. ISSN: 1432-0770. DOI: 10.1007/BF00344251.

Gabor, D. (Nov. 1946). „Theory of communication. Part 1: The analysis of information“. English. In: *Journal of the Institution of Electrical Engineers - Part III: Radio and Communication Engineering* 93 (26), 429–441(12). ISSN: 0367-7540. URL: https://digital-library.theiet.org/content/journals/10.1049/ji-3-2.1946.0074.

Gadzicki, K. (2009). „Hierarchical Clustering of Sensorimotor Features“. In: *KI 2009: Advances in Artificial Intelligence*. Ed. by B. Mertsching, M. Hund, and Z. Aziz. Vol. 5803. Lecture Notes in Artificial Intelligence. Springer, p. 737. ISBN: 978-3-642-04616-2.

Gadzicki, K. and C. Zetzsche (Sept. 2012). „Prediction of the Perceived Quality of Streak Distortions in Offset-Printing with a Psychophysically Motivated Multi-channel Model“. In: *Tagungsband 18.Workshop Farbbildverarbeitung 2012*. Ed. by P. Urban and M. Goesele. Darmstadt, Germany: TU Darmstadt, Fraunhofer Institut für Graphische Datenverarbeitung, pp. 119–130. ISBN: 978-3-00-039639-7.

– (2013). „Prediction of the perceived quality of streak distortions in offset-printing with a psychophysically motivated multi-channel model“. In: *Journal of Modern Optics* 60.14, pp. 1167–1175. DOI: 10.1080/09500340.2013.809162.

Gadzicki, K., R. Khamsehashari, and C. Zetzsche (2018). „Multimodal Convolutional Neural Networks for Human Activity Recognition“. In: *IROS 2018: Workshop on Latest Advances in Big Activity Data Sources for Robotics & New Challenges* (Madrid).

– (2020). „Early vs Late Fusion in Multimodal Convolutional Neural Networks“. In: *2020 IEEE 23rd International Conference on Information Fusion (FUSION)*, pp. 1–6. DOI: 10.23919/FUSION45008.2020.9190246. URL: https://doi.org/10.23919/FUSION45008.2020.9190246.

Gallivan, J. P., C. S. Chapman, D. M. Wolpert, and J. R. Flanagan (Sept. 2018). „Decision-making in sensorimotor control“. In: *Nature Reviews Neuroscience* 19.9, pp. 519–534. ISSN: 1471-0048. DOI: 10.1038/s41583-018-0045-9.

Geisler, W. and D. Albrecht (1992). „Cortical neurons: isolation of contrast gain control“. In: *Vis Res* 32.8, pp. 1409–1410.

Ghazanfar, A. A. and C. E. Schroeder (2006). „Is neocortex essentially multisensory?“ In: *Trends in Cognitive Sciences* 10.6, pp. 278–285. ISSN: 1364-6613. DOI: https://doi.org/10.1016/j.tics.2006.04.008. URL: https://www.sciencedirect.com/science/article/pii/S1364661306001045.

Giard, M. H. and F. Peronnet (Sept. 1999). „Auditory-Visual Integration during Multimodal Object Recognition in Humans: A Behavioral and Electrophysiological Study“. In: *Journal of Cognitive Neuroscience* 11.5, pp. 473–490. ISSN: 0898-929X. DOI: 10.1162/089892999563544. eprint: https://direct.mit.edu/jocn/article-pdf/11/5/473/1758592/089892999563544.pdf.

Gibson, J. J. (1979). *The Ecological Approach to Visual Perception*. Boston, Massachusetts.

Ginsburg, A. P. (1986). „Spatial filtering and visual form perception.“ In: Handbook of perception and human performance, Vol. 2: Cognitive processes and performance. Oxford, England: John Wiley & Sons, pp. 1–41. ISBN: 0-471-82956-0 (Hardcover); 0-471-82957-9 (Hardcover).

Girdhar, R., D. Ramanan, A. Gupta, J. Sivic, and B. Russell (2017). „ActionVLAD: Learning Spatio-Temporal Aggregation for Action Classification“. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3165–3174. DOI: 10.1109/CVPR.2017.337.

Girod, B. (1993). „What's Wrong with Mean-Squared Error?“ In: *Digital Images and Human Vision*. Cambridge, MA, USA: MIT Press, pp. 207–220. ISBN: 0262231719.

Glorot, X., A. Bordes, and Y. Bengio (Apr. 2011). „Deep Sparse Rectifier Neural Networks“. In: *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*. Ed. by G. Gordon, D. Dunson, and M. Dudík. Vol. 15. Proceedings of Machine Learning Research. Fort Lauderdale, FL, USA: JMLR Workshop and Conference Proceedings, pp. 315–323. URL: http://proceedings.mlr.press/v15/glorot11a.html.

Glünder, H. (1986). „Neural computation of inner geometric pattern relations“. en. In: *Biological Cybernetics* 55.4, pp. 239–251. ISSN: 0340-1200. DOI: 10.1007/BF00355599.

Goldberg, M., H. Eggers, and P. Gouras (1991). „The ocular motor system“. In: *Principles of neural science*. Ed. by E. Kandel, J. Schwartz, and T. Jessell. 3rd edition. Norwalk: Appleton & Lange. Chap. 43, pp. 660–679.

Goodfellow, I., Y. Bengio, and A. Courville (2016). *Deep Learning*. MIT Press.

Goodfellow, I., D. Warde-Farley, M. Mirza, A. Courville, and Y. Bengio (June 2013). „Maxout Networks“. In: *Proceedings of the 30th International Conference on Machine Learning*. Ed. by S. Dasgupta and D. McAllester. Vol. 28. Proceedings of Machine Learning Research 3. Atlanta, Georgia, USA: PMLR, pp. 1319–1327. URL: http://proceedings.mlr.press/v28/goodfellow13.html.

Gordon, J. and E. H. Shortliffe (1985). „A method for managing evidential reasoning in a hierarchical hypothesis space“. In: *Artificial Intelligence* 26.3, pp. 323–357. ISSN: 0004-3702. DOI: https://doi.org/10.1016/0004-3702(85)90064-5. URL: https://www.sciencedirect.com/science/article/pii/0004370285900645.

Graham, N. (1977). „Visual detection of aperiodic spatial stimuli by probability summation among narrowband channels“. In: *Vision Research* 17.5, pp. 637–652. ISSN: 0042-6989. DOI: https://doi.org/10.1016/0042-6989(77)90140-7. URL: https://www.sciencedirect.com/science/article/pii/0042698977901407.

Graham, N. and J. G. Robson (1987). „Summation of very close spatial frequencies: the importance of spatial probability summation“. In: *Vision Research* 27.11, pp. 1997–2007. ISSN: 0042-6989. DOI: https://doi.org/10.1016/0042-6989(87)90063-0. URL: https://www.sciencedirect.com/science/article/pii/0042698987900630.

Grill-Spector, K., K. S. Weiner, J. Gomez, A. Stigliani, and V. S. Natu (2018). „The functional neuroanatomy of face perception: from brain measurements to deep neural networks“. In: *Interface Focus* 8.4, p. 20180013. DOI: 10.1098/rsfs.2018.0013. URL: https://royalsocietypublishing.org/doi/abs/10.1098/rsfs.2018.0013.

Handbuch zur technischen Abnahme von Bogenoffset-Rollenoffsetmaschinen (1996). Tech. rep. Bundesverband Druck und Medien.

Hartline, H. K., H. G. Wagner, and F. Ratliff (May 1956). „Inhibition in the eye of the limulus“. In: *Journal of General Physiology* 39.5, pp. 651–673. ISSN: 0022-1295. DOI: 10.1085/jgp.39.5.651. eprint: https://rupress.org/jgp/article-pdf/39/5/651/1241156/651.pdf.

Hartline, H. K. (1938). „The response of single optic nerve fibers of the vertebrate eye to illumination of the retina“. In: *American Journal of Physiology-Legacy Content* 121.2, pp. 400–415. DOI: 10.1152/ajplegacy.1938.121.2.400.

Hawken, M. J. and A. J. Parker (1987). „Spatial properties of neurons in the monkey striate cortex“. In: *Proceedings of the Royal Society of London. Series B. Biological Sciences* 231.1263, pp. 251–288. DOI: 10.1098/rspb.1987.0044.

Hayes, S. J., M. Andrew, N. C. Foster, D. Elliott, E. Gowen, and S. J. Bennett (2018). „Sensorimotor learning and associated visual perception are intact but unrelated in autism spectrum disorder“. In: *Autism Research* 11.2, pp. 296–304. DOI: https://doi.org/10.1002/aur.1882. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/aur.1882. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/aur.1882.

He, K., G. Gkioxari, P. Dollár, and R. Girshick (2017). „Mask R-CNN“. In: *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 2980–2988. DOI: 10.1109/ICCV.2017.322.

He, K., X. Zhang, S. Ren, and J. Sun (2016). „Deep Residual Learning for Image Recognition“. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778. DOI: 10.1109/CVPR.2016.90.

Heeger, D. (1992). „Normalization of cell responses in cat striate cortex“. In: *Vis Neurosci* 9.2, pp. 181–198.

Heeger, D. and P. Teo (1995). „A Model of Perceptual Image Fidelity“. In: *International Conference on Image Processing, 1995*. Vol. 2. Vol. 2, 1995, pp. 343–346. ISBN: 0818673109.

Högman, V., M. Björkman, and D. Kragic (2013). „Interactive object classification using sensorimotor contingencies“. In: *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 2799–2805. DOI: 10.1109/IROS.2013.6696752.

Hornik, K. (1991). „Approximation capabilities of multilayer feedforward networks“. In: *Neural Networks* 4.2, pp. 251–257. ISSN: 0893-6080. DOI: https://doi.org/10.1016/0893-6080(91)90009-T. URL: https://www.sciencedirect.com/science/article/pii/089360809190009T.

Hou, W., X. Gao, D. Tao, and X. Li (2015). „Blind Image Quality Assessment via Deep Learning“. In: *IEEE Transactions on Neural Networks and Learning Systems* 26.6, pp. 1275–1286. DOI: 10.1109/TNNLS.2014.2336852.

Hou, Y., Z. Li, P. Wang, and W. Li (Mar. 2018). „Skeleton Optical Spectra-Based Action Recognition Using Convolutional Neural Networks“. In: *IEEE Transactions on Circuits and*

Systems for Video Technology 28.3, pp. 807–811. ISSN: 1558-2205. DOI: 10.1109/TCSVT.2016.2628339.

Huang, G., Z. Liu, L. Van Der Maaten, and K. Q. Weinberger (2017). „Densely Connected Convolutional Networks“. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2261–2269. DOI: 10.1109/CVPR.2017.243.

Hubel, D. H. and T. N. Wiesel (1959). „Receptive fields of single neurones in the cat's striate cortex“. In: *J Physiol* 148.3, pp. 574–591.

– (1962). „Receptive fields, binocular interaction and functional architecture in the cat's visual cortex“. In: *J Physiol* 160.1, pp. 106–154.

– (1968). „Receptive fields and functional architecture of monkey striate cortex“. In: *J Physiol* 195.1, pp. 215–243. ISSN: 0084-2230.

– (1977). „Ferrier lecture - Functional architecture of macaque monkey visual cortex“. In: *Proceedings of the Royal Society of London. Series B. Biological Sciences* 198.1130, pp. 1–59. DOI: 10.1098/rspb.1977.0085. eprint: https://royalsocietypublishing.org/doi/pdf/10.1098/rspb.1977.0085. URL: https://royalsocietypublishing.org/doi/abs/10.1098/rspb.1977.0085.

Ilg, E., N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox (2017). „FlowNet 2.0: Evolution of Optical Flow Estimation with Deep Networks“. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. URL: http://lmb.informatik.uni-freiburg.de/Publications/2017/IMSKDB17.

Ioffe, S. and C. Szegedy (2015). „Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift“. In: *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*. ICML15. Lille, France: JMLR.org, pp. 448–456.

Jarrett, K., K. Kavukcuoglu, M. Ranzato, and Y. LeCun (2009). „What is the best multi-stage architecture for object recognition?“ In: *2009 IEEE 12th International Conference on Computer Vision*, pp. 2146–2153. DOI: 10.1109/ICCV.2009.5459469.

Jégou, H., F. Perronnin, M. Douze, J. Sánchez, P. Pérez, and C. Schmid (2012). „Aggregating Local Image Descriptors into Compact Codes“. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 34.9, pp. 1704–1716. DOI: 10.1109/TPAMI.2011.235.

Jhuang, H., T. Serre, L. Wolf, and T. Poggio (2007). „A Biologically Inspired System for Action Recognition“. In: *2007 IEEE 11th International Conference on Computer Vision*, pp. 1–8. DOI: 10.1109/ICCV.2007.4408988.

Ji, S., W. Xu, M. Yang, and K. Yu (2010). „3D Convolutional Neural Networks for Human Action Recognition“. In: *Proceedings of the 27th International Conference on Machine Learning (ICML-10), June 21-24, 2010, Haifa, Israel*. Ed. by J. Fürnkranz and T. Joachims. Omnipress, pp. 495–502. URL: https://icml.cc/Conferences/2010/papers/100.pdf.

– (2013). „3D Convolutional Neural Networks for Human Action Recognition“. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35.1, pp. 221–231. DOI: 10.1109/TPAMI.2012.59.

Julesz, B. (Feb. 1962). „Visual Pattern Discrimination“. In: *IEEE Transactions on Information Theory* 8.2, pp. 84–92. ISSN: 0018-9448. DOI: 10.1109/TIT.1962.1057698.

Kandel, E. R., J. H. Schwartz, and T. M. Jessell (2000). *Principles of neural science*. English. New York: McGraw-Hill, Health Professions Division.

Kar, A., N. Rai, K. Sikka, and G. Sharma (2017). „AdaScan: Adaptive Scan Pooling in Deep Convolutional Neural Networks for Human Action Recognition in Videos“. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5699–5708. DOI: 10.1109/CVPR.2017.604.

Karpathy, A., G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei (2014). „Large-Scale Video Classification with Convolutional Neural Networks“. In: *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition*. CVPR '14. Washington,

DC, USA: IEEE Computer Society, pp. 1725–1732. ISBN: 978-1-4799-5118-5. DOI: 10.1109/CVPR.2014.223.

Kaspar, K., S. König, J. Schwandt, and P. König (2014). „The experience of new sensorimotor contingencies by sensory augmentation“. In: *Consciousness and Cognition* 28, pp. 47–63. ISSN: 1053-8100. DOI: https://doi.org/10.1016/j.concog.2014.06.006. URL: https://www.sciencedirect.com/science/article/pii/S1053810014000920.

Kay, W., J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, M. Suleyman, and A. Zisserman (2017). „The Kinetics Human Action Video Dataset“. In: *CoRR* abs/1705.06950. arXiv: 1705.06950. URL: http://arxiv.org/abs/1705.06950.

Kayargadde, V. and J. Martens (1996). „Perceptual characterization of images degraded by blur and noise: model“. In: *J Opt Soc Am A Opt Image Sci Vis* 13.6, pp. 1178–88. ISSN: 1084-7529.

Khamsehashari, R., K. Gadzicki, and C. Zetzsche (2019). „Deep Residual Temporal Convolutional Networks for Skeleton-Based Human Action Recognition“. In: *Computer Vision Systems*. Ed. by D. Tzovaras, D. Giakoumis, M. Vincze, and A. Argyros. Cham: Springer International Publishing, pp. 376–385. ISBN: 978-3-030-34995-0.

Kim, H. E., G. Avraham, and R. B. Ivry (2021). „The Psychology of Reaching: Action Selection, Movement Implementation, and Sensorimotor Learning“. In: *Annual Review of Psychology* 72.1. PMID: 32976728, pp. 61–95. DOI: 10.1146/annurev-psych-010419-051053.

Kim, J., A. Nguyen, and S. Lee (2019). „Deep CNN-Based Blind Image Quality Predictor“. In: *IEEE Transactions on Neural Networks and Learning Systems* 30.1, pp. 11–24. DOI: 10.1109/TNNLS.2018.2829819.

Kim, T. S. and A. Reiter (July 2017). „Interpretable 3D Human Action Analysis with Temporal Convolutional Networks“. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1623–1631. DOI: 10.1109/CVPRW.2017.207.

King-Smith, P. E. and J. J. Kulikowski (1975). „Pattern and flicker detection analysed by subthreshold summation.“ In: *The Journal of Physiology* 249.3, pp. 519–548. DOI: https://doi.org/10.1113/jphysiol.1975.sp011028. URL: https://physoc.onlinelibrary.wiley.com/doi/abs/10.1113/jphysiol.1975.sp011028.

Kluth, T., D. Nakath, T. Reineking, C. Zetzsche, and K. Schill (2015). „Affordance-Based Object Recognition Using Interactions Obtained from a Utility Maximization Principle“. In: *Computer Vision - ECCV 2014 Workshops*. Ed. by L. Agapito, M. M. Bronstein, and C. Rother. Cham: Springer International Publishing, pp. 406–412. ISBN: 978-3-319-16181-5.

Kohonen, T. (1990). „The self-organizing map“. In: *Proceedings of the IEEE* 78.9, pp. 1464–1480. DOI: 10.1109/5.58325.

– (2001). *Self-organizing maps*. 3rd. Vol. 30. Springer series in information sciences. Berlin, Heidelberg, New York: Springer.

Kong, Y., Z. Tao, and Y. Fu (2017). „Deep Sequential Context Networks for Action Prediction“. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3662–3670. DOI: 10.1109/CVPR.2017.390.

Kong, Y., D. Kit, and Y. Fu (2014). „A Discriminative Model with Multiple Temporal Scales for Action Prediction“. In: *Computer Vision – ECCV 2014*. Ed. by D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars. Cham: Springer International Publishing, pp. 596–611. ISBN: 978-3-319-10602-1.

Koppula, H. S. and A. Saxena (2016). „Anticipating Human Activities Using Object Affordances for Reactive Robotic Response“. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 38.1, pp. 14–29. DOI: 10.1109/TPAMI.2015.2430335.

Krizhevsky, A., I. Sutskever, and G. E. Hinton (Jan. 2012). „ImageNet Classification with Deep Convolutional Neural Networks“. In: *Neural Information Processing Systems* 25. DOI: 10.1145/3065386.

Krupinski, E. A., J. Johnson, H. Roehrig, J. Nafziger, J. Fan, and J. Lubin (Dec. 2004). „Use of a Human Visual System Model to Predict Observer Performance with CRT vs LCD Display of Images“. In: *Journal of Digital Imaging* 17.4, pp. 258–263. ISSN: 1618-727X. DOI: 10.1007/s10278-004-1016-4.

Kuipers, B. J. and T. S. Levitt (June 1988). „Navigation and Mapping in Large Scale Space“. In: *AI Magazine* 9.2, p. 25. DOI: 10.1609/aimag.v9i2.674. URL: https://ojs.aaai.org/index.php/aimagazine/article/view/674.

Kulikowski, J. J., S. Marčelja, and P. O. Bishop (Apr. 1982). „Theory of spatial position and spatial frequency relations in the receptive fields of simple cells in the visual cortex“. In: *Biological Cybernetics* 43.3, pp. 187–198. ISSN: 1432-0770. DOI: 10.1007/BF00319978.

CIE 1976 (1976). *L*a*b* Colour space*. Standard.

Lang, P., X. Fu, M. Martorella, J. Dong, R. Qin, X. Meng, and M. Xie (2020). *A Comprehensive Survey of Machine Learning Applied to Radar Signal Processing*. arXiv: 2009.13702 [eess.SP].

Lanillos, P., E. Dean-Leon, and G. Cheng (2017). „Yielding Self-Perception in Robots Through Sensorimotor Contingencies“. In: *IEEE Transactions on Cognitive and Developmental Systems* 9.2, pp. 100–112. DOI: 10.1109/TCDS.2016.2627820.

Laptev, I., M. Marszalek, C. Schmid, and B. Rozenfeld (2008). „Learning realistic human actions from movies“. In: *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8. DOI: 10.1109/CVPR.2008.4587756.

Laptev, I. (Sept. 2005). „On Space-Time Interest Points“. In: *International Journal of Computer Vision* 64.2, pp. 107–123. ISSN: 1573-1405. DOI: 10.1007/s11263-005-1838-7.

Larson, E. and D. Chandler (2010). „Most apparent distortion: full-reference image quality assessment and the role of strategy“. In: *J Electron Imaging* 19.1, pp. 011006–1–011006–21. DOI: 10.1117/1.3267105.

Laskar, M. N. U., L. G. S. Giraldo, and O. Schwartz (2018). *Correspondence of Deep Neural Networks and the Brain for Visual Textures*. arXiv: 1806.02888 [q-bio.NC].

LeCun, Y., B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel (Dec. 1989). „Backpropagation Applied to Handwritten Zip Code Recognition“. In: *Neural Computation* 1.4, pp. 541–551. ISSN: 0899-7667. DOI: 10.1162/neco.1989.1.4.541.

LeCun, Y., P. Haffner, L. Bottou, and Y. Bengio (1999). „Object Recognition with Gradient-Based Learning“. In: *Shape, Contour and Grouping in Computer Vision*. Berlin, Heidelberg: Springer-Verlag, p. 319. ISBN: 3540667229.

Legge, G. and J. Foley (1980). „Contrast masking in human vision“. In: *J Opt Soc Am* 70.12, pp. 1458–71. ISSN: 0030-3941.

Li, C., Y. Hou, P. Wang, and W. Li (May 2017). „Joint Distance Maps Based Action Recognition With Convolutional Neural Networks“. In: *IEEE Signal Processing Letters* 24.5, pp. 624–628. ISSN: 1558-2361. DOI: 10.1109/LSP.2017.2678539.

Li, Y., L.-M. Po, X. Xu, L. Feng, F. Yuan, C.-H. Cheung, and K.-W. Cheung (2015). „No-reference image quality assessment with shearlet transform and deep neural networks“. In: *Neurocomputing* 154, pp. 94–109. ISSN: 0925-2312. DOI: https://doi.org/10.1016/j.neucom.2014.12.015. URL: https://www.sciencedirect.com/science/article/pii/S0925231214016798.

Lin, M., Q. Chen, and S. Yan (2014). „Network In Network“. In: *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*. Ed. by Y. Bengio and Y. LeCun. URL: http://arxiv.org/abs/1312.4400.

Liu, J., S. Zhang, S. Wang, and D. N. Metaxas (2016). „Multispectral Deep Neural Networks for Pedestrian Detection“. In: *CoRR* abs/1611.02644. arXiv: 1611.02644. URL: http://arxiv.org/abs/1611.02644.

Lowe, D. G. (Nov. 2004). „Distinctive Image Features from Scale-Invariant Keypoints“. In: *International Journal of Computer Vision* 60.2, pp. 91–110. ISSN: 1573-1405. DOI: 10.1023/B:VISI.0000029664.99615.94.

Luan, S., C. Chen, B. Zhang, J. Han, and J. Liu (2018). „Gabor Convolutional Networks“. In: *IEEE Transactions on Image Processing* 27.9, pp. 4357–4366. DOI: 10.1109/TIP.2018.2835143.

Lubin, J. (1993). „The use of psychophysical data and models in the analysis of display system performance“. In: *Digital images and human vision*. Ed. by A. B. Watson. Cambridge, MA, USA: MIT Press, pp. 163–178. ISBN: 0-262-23171-9.

Lubin, J. (1995). „A visual discrimination model for imaging system design and evaluation“. In: *Vision Models for Target Detection and Recognition*, pp. 245–283. DOI: 10.1142/9789812831200_0010.

Maas, A. L., A. Y. Hannun, and A. Y. Ng (2013). „Rectifier nonlinearities improve neural network acoustic models“. In: *in ICML Workshop on Deep Learning for Audio, Speech and Language Processing*.

Mannos, J. and D. Sakrison (1974). „The effects of a visual fidelity criterion of the encoding of images“. In: *IEEE Transactions on Information Theory* 20.4, pp. 525–536. DOI: 10.1109/TIT.1974.1055250.

Mantiuk, R., K. Myszkowski, and H. .-. Seidel (2004). „Visible difference predicator for high dynamic range images“. In: *2004 IEEE International Conference on Systems, Man and Cybernetics (IEEE Cat. No.04CH37583)*. Vol. 3, 2763–2769 vol.3. DOI: 10.1109/ICSMC.2004.1400750.

Marcelja, S. (1980). „Mathematical description of the responses of simple cortical cells“. In: *J Optl Soc Am* 70.11, pp. 1297–300. ISSN: 0030-3941. URL: http://www.ncbi.nlm.nih.gov/pubmed/7463179.

Marmolin, H. (1986). „Subjective MSE Measures“. In: *IEEE Transactions on Systems, Man, and Cybernetics* 16.3, pp. 486–489. DOI: 10.1109/TSMC.1986.4308985.

Mason, C., M. Meier, F. Ahrens, T. Fehr, M. Herrmann, F. Putze, and T. Schultz (2018). „Human activities data collection and labeling using a think-aloud protocol in a table setting scenario“. In: IROS.

Mason, C., K. Gadzicki, M. Meier, F. Ahrens, T. Kluss, J. Maldonado, F. Putze, T. Fehr, C. Zetzsche, M. Herrmann, K. Schill, and T. Schultz (2020). „From Human to Robot Everyday Activity“. In: *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 8997–9004. DOI: 10.1109/IROS45743.2020.9340706. URL: https://doi.org/10.1109/IROS45743.2020.9340706.

Maye, A. and A. K. Engel (2011). „A discrete computational model of sensorimotor contingencies for object perception and control of behavior“. In: *2011 IEEE International Conference on Robotics and Automation*, pp. 3810–3815. DOI: 10.1109/ICRA.2011.5979919.

Maye, A. and A. K. Engel (2012). „Time Scales of Sensorimotor Contingencies“. In: *Advances in Brain Inspired Cognitive Systems*. Ed. by H. Zhang, A. Hussain, D. Liu, and Z. Wang. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 240–249. ISBN: 978-3-642-31561-9.

McClelland, J. L., B. L. McNaughton, and R. C. O'Reilly (July 1995). „Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory“. In: *Psychol Rev* 102.3, pp. 419–457.

McGurk, H. and J. MacDonald (Dec. 1976). „Hearing lips and seeing voices“. In: *Nature* 264.5588, pp. 746–748. ISSN: 1476-4687. DOI: 10.1038/264746a0.

Merigan, W. H. and J. H. R. Maunsell (1993). „How Parallel are the Primate Visual Pathways?“ In: *Annual Review of Neuroscience* 16.1. PMID: 8460898, pp. 369–402. DOI: 10.1146/annurev.ne.16.030193.002101.

Michelson, A. A. (1927). *Studies in optics*. English. Chicago, Ill.: The University of Chicago Press.

Mishkin, M., L. G. Ungerleider, and K. A. Macko (1983). „Object vision and spatial vision: two cortical pathways“. In: *Trends in Neurosciences* 6, pp. 414–417. ISSN: 0166-2236. DOI: https://doi.org/10.1016/0166-2236(83)90190-X. URL: https://www.sciencedirect.com/science/article/pii/016622368390190X.

Miyahara, M. (1988). „Quality assessments for visual service“. In: *IEEE Communications Magazine* 26.10, pp. 51–60. ISSN: 0163-6804. DOI: 10.1109/35.7667.

Miyahara, M., K. Kotani, and V. Algazi (1998). „Objective picture quality scale (PQS) for image coding“. In: *IEEE Transactions on Communications* 46.9, pp. 1215–1226. ISSN: 00906778. DOI: 10.1109/26.718563.

Nadenau, M. J., S. Winkler, D. Alleysson, and M. Kunt (2000). „Human Vision Models for Perceptually Optimized Image Processing – A Review“. In: *Computer Science*.

Nair, V. and G. E. Hinton (2010). „Rectified Linear Units Improve Restricted Boltzmann Machines“. In: *Proceedings of the 27th International Conference on International Conference on Machine Learning*. ICML'10. Haifa, Israel: Omnipress, pp. 807–814. ISBN: 9781605589077.

Nakath, D., J. Clemens, and K. Schill (2018). „Multi-Sensor Fusion and Active Perception for Autonomous Deep Space Navigation“. In: *2018 21st International Conference on Information Fusion (FUSION)*, pp. 2596–2605. DOI: 10.23919/ICIF.2018.8455788.

Nakath, D., J. Clemens, and C. Rachuy (Aug. 2020). „Active Asteroid-SLAM“. In: *Journal of Intelligent & Robotic Systems* 99.2, pp. 303–333. ISSN: 1573-0409. DOI: 10.1007/s10846-019-01103-0.

Nakath, D., T. Kluth, T. Reineking, C. Zetzsche, and K. Schill (2014). „Active Sensorimotor Object Recognition in Three-Dimensional Space“. In: *Spatial Cognition IX*. Ed. by C. Freksa, B. Nebel, M. Hegarty, and T. Barkowsky. Cham: Springer International Publishing, pp. 312–324. ISBN: 978-3-319-11215-2.

Nakath, D., C. Rachuy, J. Clemens, and K. Schill (2016). „Optimal rotation sequences for active perception“. In: *Multisensor, Multisource Information Fusion: Architectures, Algorithms, and Applications 2016*. Ed. by J. J. Braun. Vol. 9872. International Society for Optics and Photonics. SPIE, pp. 20–32. DOI: 10.1117/12.2223027.

Noton, D. and L. Stark (1971). „Scanpaths in Eye Movements during Pattern Perception“. In: *Science* 171.3968, pp. 308–311. ISSN: 0036-8075. DOI: 10.1126/science.171.3968.308.

O'Regan, J. K. (1992). „Solving the "real" mysteries of visual perception: The world as an outside memory.“ In: *Canadian Journal of Psychology/Revue canadienne de psychologie* 46.3, pp. 461–488. ISSN: 0008-4255(Print). DOI: 10.1037/h0084327.

O'Regan, J. K. and A. Noë (2001). „A sensorimotor account of vision and visual consciousness“. In: *Behavioral and Brain Sciences* 24, pp. 939–973.

Oleskiw, T. D., J. D. Lieber, J. A. Movshon, and E. P. Simoncelli (May 2020). „Testing a two-stage model of stimulus selectivity in macaque V2“. In: *Annual Meeting, Vision Sciences Society*. Vol. 20. St. Petersburg, Florida.

Oleskiw, T. D. and E. P. Simoncelli (Nov. 2018). „Learning a model for visual texture selectivity from natural images“. In: *Annual Meeting, Neuroscience*.

– (May 2019). „A canonical computational model of cortical area V2“. In: *Annual Meeting, Vision Sciences Society*. Vol. 19. St. Petersburg, Florida.

Oliva, A. and A. Torralba (2006). „Chapter 2 Building the gist of a scene: the role of global image features in recognition“. In: *Visual Perception*. Ed. by S. Martinez-Conde, S. Macknik, L. Martinez, J.-M. Alonso, and P. Tse. Vol. 155. Progress in Brain Research. Elsevier, pp. 23–36. DOI: https://doi.org/10.1016/S0079-6123(06)55002-2. URL: https://www.sciencedirect.com/science/article/pii/S0079612306550022.

Onnasch, L. and E. Roesler (June 2020). „A Taxonomy to Structure and Analyze Human–Robot Interaction“. In: *International Journal of Social Robotics*. ISSN: 1875-4805. DOI: 10.1007/s12369-020-00666-5.

Oram, M. W. and D. I. Perrett (1994). „Modeling visual recognition from neurobiological constraints“. In: *Neural Networks* 7.6. Models of Neurodynamics and Behavior, pp. 945–

972. ISSN: 0893-6080. DOI: https://doi.org/10.1016/S0893-6080(05)80153-4. URL: https://www.sciencedirect.com/science/article/pii/S0893608005801534.

Palmer, S. E. (1999). *Vision Science - Photons to Phenomenology*. MIT Press, Cambridge. ISBN: 9780262161831.

Parthasarathy, N. and E. P. Simoncelli (Feb. 2020). „Learning a texture model for representing cortical area V2“. In: *Computational and Systems Neuroscience (CoSyNe)*. Denver, CO.

Pavllo, D., C. Feichtenhofer, D. Grangier, and M. Auli (2019). „3D human pose estimation in video with temporal convolutions and semi-supervised training“. In: *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Peli, E. (Oct. 1990). „Contrast in complex images“. In: *J. Opt. Soc. Am. A* 7.10, pp. 2032–2040. DOI: 10.1364/JOSAA.7.002032. URL: http://josaa.osa.org/abstract.cfm?URI=josaa-7-10-2032.

– (1997). „In search of a contrast metric: Matching the perceived contrast of gabor patches at different phases and bandwidths“. In: *Vision Research* 37.23, pp. 3217–3224. ISSN: 0042-6989. DOI: https://doi.org/10.1016/S0042-6989(96)00262-3. URL: https://www.sciencedirect.com/science/article/pii/S0042698996002623.

Pelli, D. and B. Farell (1995). „Psychophysical methods“. In: *Handbook of Optics: Fundamentals, techniques and design*. Ed. by M. Bass, E. W. V. Stryland, D. R. Williams, and W. L. Wolfe. 2nd edition. Vol. 1. Chap. 29.

Perronnin, F., J. Sánchez, and T. Mensink (2010). „Improving the Fisher Kernel for Large-Scale Image Classification“. In: *Computer Vision – ECCV 2010*. Ed. by K. Daniilidis, P. Maragos, and N. Paragios. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 143–156. ISBN: 978-3-642-15561-1.

Pham, H.-H., L. Khoudour, A. Crouzil, P. Zegers, and S. A. Velastin (May 2018). „Exploiting deep residual networks for human action recognition from skeletal data“. In: *Computer Vision and Image Understanding* 170, pp. 51–66. ISSN: 1077-3142. DOI: 10.1016/j.cviu.2018.03.003.

Philipona, D., J. K. O'Regan, and J.-P. Nadal (Sept. 2003). „Is There Something Out There? Inferring Space from Sensorimotor Dependencies“. In: *Neural Computation* 15.9, pp. 2029–2049. ISSN: 0899-7667. DOI: 10.1162/089976603322297278. eprint: https://direct.mit.edu/neco/article-pdf/15/9/2029/815702/089976603322297278.pdf.

Pouget, A., P. Dayan, and R. S. Zemel (2003). „INFERENCE AND COMPUTATION WITH POPULATION CODES“. In: *Annual Review of Neuroscience* 26.1. PMID: 12704222, pp. 381–410. DOI: 10.1146/annurev.neuro.26.041002.131112.

Probst, A., G. G. Peytaví, D. Nakath, A. Schattel, C. Rachuy, P. Lange, J. Clemens, M. Echim, V. Schwarting, A. Srinivas, K. Gadzicki, R. Förstner, B. Eissfeller, K. Schill, C. Büskens, and G. Zachmann (2015). „KaNaRiA: Identifying the Challenges for Cognitive Autonomous Navigation and Guidance for Missions to Small Planetary Bodies“. In: *66th International Astronautical Congress (IAC)*. Jerusalem, Israel.

Qi, C. R., H. Su, M. Nießner, A. Dai, M. Yan, and L. J. Guibas (2016). „Volumetric and Multi-view CNNs for Object Classification on 3D Data“. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5648–5656. DOI: 10.1109/CVPR.2016.609.

Qiu, Z., T. Yao, and T. Mei (2017). „Learning Spatio-Temporal Representation with Pseudo-3D Residual Networks“. In: *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 5534–5542. DOI: 10.1109/ICCV.2017.590.

Quick, R. F. (June 1974). „A vector-magnitude model of contrast detection“. In: *Kybernetik* 16.2, pp. 65–67. ISSN: 1432-0770. DOI: 10.1007/BF00271628.

Rachuy, C. (2020). *Manifold-Based Sensorimotor Representations for Bootstrapping of Mobile Agents*. DOI: 10.26092/elib/24.

Radeau, M. and P. Bertelson (Mar. 1977). „Adaptation to auditory-visual discordance and ventriloquism in semirealistic situations“. In: *Perception & Psychophysics* 22.2, pp. 137–146. ISSN: 1532-5962. DOI: 10.3758/BF03198746.

Ramezani, M. and F. Yaghmaee (Dec. 2016). „A review on human action analysis in videos for retrieval applications“. In: *Artificial Intelligence Review* 46.4, pp. 485–514. ISSN: 1573-7462. DOI: 10.1007/s10462-016-9473-y.

Rasouli, A. and J. K. Tsotsos (2018). „Joint Attention in Driver-Pedestrian Interaction: from Theory to Practice“. In: *CoRR* abs/1802.02522. arXiv: 1802.02522. URL: http://arxiv.org/abs/1802.02522.

Recanzone, G. H. (Dec. 2009). „Interactions of auditory and visual stimuli in space and time“. eng. In: *Hearing research* 258.1-2. S0378-5955(09)00096-3[PII], pp. 89–99. ISSN: 1878-5891. DOI: 10.1016/j.heares.2009.04.009. URL: https://pubmed.ncbi.nlm.nih.gov/19393306.

ITU-R BT.500-11 (2002). *Recommendation: Methodology for the subjective assessment of the quality of television pictures*. Tech. rep. International Telecommunication Union, Geneva, pp. 1–48.

Reineking, T., J. Wolter, K. Gadzicki, and C. Zetzsche (Aug. 2010). „Bio-inspired Architecture for Active Sensorimotor Localization“. In: *Spatial Cognition VII*. Lecture Notes in Artificial Intelligence. Portland, Oregon: Springer, pp. 163–178.

Reineking, T. (2011). „Particle filtering in the Dempster–Shafer theory“. In: *International Journal of Approximate Reasoning* 52.8, pp. 1124–1135. ISSN: 0888-613X. DOI: https://doi.org/10.1016/j.ijar.2011.06.003. URL: https://www.sciencedirect.com/science/article/pii/S0888613X11000922.

Ren, S., K. He, R. Girshick, and J. Sun (2017). „Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks“. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39.6, pp. 1137–1149. DOI: 10.1109/TPAMI.2016.2577031.

Resnikoff, H. and R. Wells (1984). *Mathematics in Civilization*. Popular Science Series. Dover. ISBN: 9780486246741.

Reyneri, L., V. Colla, and M. Vannucci (2011). „Estimate of a Probability Density Function through Neural Networks“. In: *Advances in Computational Intelligence*. Ed. by J. Cabestany, I. Rojas, and G. Joya. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 57–64. ISBN: 978-3-642-21501-8.

Riesenhuber, M. and T. Poggio (Nov. 1999). „Hierarchical models of object recognition in cortex“. In: *Nature Neuroscience* 2.11, pp. 1019–1025. ISSN: 1546-1726. DOI: 10.1038/14819.

Robson, J. G. (1981). „Neural images: The Physiological Basis of Spatial Vision“. In: *Visual Coding and Adaptatbility*. Ed. by C. S. Harris. 1st edition. New York: Psychology Press.

– (1993). „Contrast Sensitivity: One Hundred Years of Clinical Measurement“. English. In: *Contrast sensitivity*. Ed. by R. M. Shapley and D. M.-K. Lam. Cambridge, Mass.: MIT Press.

Rodieck, R. (1965). „Quantitative analysis of cat retinal ganglion cell response to visual stimuli“. In: *Vision Research* 5.12, pp. 583–601. ISSN: 0042-6989. DOI: https://doi.org/10.1016/0042-6989(65)90033-7. URL: https://www.sciencedirect.com/science/article/pii/0042698965900337.

Rohrbach, M., S. Amin, M. Andriluka, and B. Schiele (June 2012). „A Database for Fine Grained Activity Detection of Cooking Activities“. In: pp. 1194–1201. ISBN: 978-1-4673-1226-4. DOI: 10.1109/CVPR.2012.6247801.

Rosenblatt, F. (1958). „The perceptron: A probabilistic model for information storage and organization in the brain.“ In: *Psychological Review* 65.6, pp. 386–408. ISSN: 0033-295X. DOI: 10.1037/h0042519.

– (1962). *Principles of neurodynamics; perceptrons and the theory of brain mechanisms*. Washington: Spartan Books.

Rosenholtz, R., J. Huang, and K. Ehinger (2012). „Rethinking the role of top-down attention in vision: effects attributable to a lossy representation in peripheral vision.“ In: *Front Psychol* 3, p. 13. ISSN: 1664-1078. DOI: 10.3389/fpsyg.2012.00013.

Rosenholtz, R., J. Huang, A. Raj, B. J. Balas, and L. Ilie (2012). „A summary statistic representation in peripheral vision explains visual search.“ In: *J Vis* 12.4, pp. 1–17. ISSN: 1534-7362. DOI: 10.1167/12.4.14.

Rumelhart, D. E., G. E. Hinton, and J. L. McClelland (1986). „A General Framework for Parallel Distributed Processing“. In: *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 1: Foundations*. Cambridge, MA, USA: MIT Press, pp. 45–76. ISBN: 026268053X.

Rumelhart, D. E., G. E. Hinton, and R. J. Williams (1986a). „Learning internal representations by error propagation“. In: *Parallel Distributed Processing*. Ed. by D. E. Rumelhart and J. L. McClelland. Vol. 1. Campbridge: MIT Press. Chap. 8, pp. 318–362.

Rumelhart, D. E., G. E. Hinton, and R. J. Williams (Oct. 1986b). „Learning representations by back-propagating errors“. In: *Nature* 323.6088, pp. 533–536. ISSN: 1476-4687. DOI: 10.1038/323533a0.

Rumelhart, D. E., J. L. McClelland, and the PDP Research Group (1986). *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*. Cambridge: MIT Press.

Ryoo, M. S. (2011). „Human activity prediction: Early recognition of ongoing activities from streaming videos“. In: *2011 International Conference on Computer Vision*, pp. 1036–1043. DOI: 10.1109/ICCV.2011.6126349.

Ryoo, M. S. and J. K. Aggarwal (June 2011). „Stochastic Representation and Recognition of High-Level Group Activities“. In: *International Journal of Computer Vision* 93.2, pp. 183–200. ISSN: 1573-1405. DOI: 10.1007/s11263-010-0355-5.

Schiessl, I. and N. McLoughlin (2003). „Optical imaging of the retinotopic organization of V1 in the common marmoset“. In: *NeuroImage* 20.3, pp. 1857–1864. ISSN: 1053-8119. DOI: https://doi.org/10.1016/j.neuroimage.2003.07.023. URL: https://www.sciencedirect.com/science/article/pii/S1053811903004440.

Schill, K. (1997). „Decision Support Systems with Adaptive Reasoning Strategies“. In: *Foundations of Computer Science: Potential - Theory - Cognition, to Wilfried Brauer on the occasion of his sixtieth birthday*. London, UK: Springer-Verlag, pp. 417–427. ISBN: 3-540-63746-X.

Schill, K., E. Umkehrer, S. Beinlich, G. Krieger, and C. Zetzsche (Jan. 2001). „Scene analysis with saccadic eye movements: Top-down and bottom-up modeling“. In: *Journal of Electronic Imaging* 10.1, pp. 152–160.

Schroeder, C. E. and J. Foxe (2005). „Multisensory contributions to low-level, ‘unisensory’ processing“. In: *Current Opinion in Neurobiology* 15.4. Sensory systems, pp. 454–458. ISSN: 0959-4388. DOI: https://doi.org/10.1016/j.conb.2005.06.008. URL: https://www.sciencedirect.com/science/article/pii/S0959438805000991.

Serre, T., L. Wolf, and T. Poggio (2005). „Object recognition with features inspired by visual cortex“. In: *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*. Vol. 2, 994–1000 vol. 2. DOI: 10.1109/CVPR.2005.254.

Shafer, G. (1976). *A mathematical theory of evidence*. Princeton, N.J.: Princeton University Press.

Shahroudy, A., J. Liu, T.-T. Ng, and G. Wang (2016). „NTU RGB+D: A Large Scale Dataset for 3D Human Activity Analysis“. In: *CoRR* abs/1604.02808. arXiv: 1604.02808. URL: http://arxiv.org/abs/1604.02808.

Sheikh, H. and A. C. Bovik (2006). „Image information and visual quality“. In: *IEEE Transactions on Image Processing* 15.2, pp. 430–44. ISSN: 1057-7149.

Shi, L., Y. Zhang, J. Cheng, and H. Lu (2020). „Skeleton-Based Action Recognition With Multi-Stream Adaptive Graph Convolutional Networks“. In: *IEEE Transactions on Image Processing* 29, pp. 9532–9545. DOI: 10.1109/TIP.2020.3028207.

Shotton, J., A. Fitzgibbon, A. Blake, A. Kipman, M. Finocchio, B. Moore, and T. Sharp (June 2011). „Real-Time Human Pose Recognition in Parts from a Single Depth Image“. In:

CVPR. IEEE. URL: https://www.microsoft.com/en-us/research/publication/real-time-human-pose-recognition-in-parts-from-a-single-depth-image/.

Simoncelli, E. P., W. T. Freeman, E. H. Adelson, and D. J. Heeger (1992). „Shiftable multiscale transforms“. In: *IEEE Transactions on Information Theory* 38.2, pp. 587–607. DOI: 10.1109/18.119725.

Simoncelli, E. P. and W. T. Freeman (Oct. 1995). „The Steerable Pyramid: A flexible architecture for multi-scale derivative computation“. In: *Proc 2nd IEEE Int'l Conf on Image Proc (ICIP)*. Vol. III. Washington, DC: IEEE Sig Proc Society, pp. 444–447. DOI: 10.1109/ICIP.1995.537667.

Simonyan, K. and A. Zisserman (2014). „Two-stream Convolutional Networks for Action Recognition in Videos“. In: *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 1*. NIPS'14. Montreal, Canada: MIT Press, pp. 568–576.

– (2015). „Very Deep Convolutional Networks for Large-Scale Image Recognition“. In: *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*. Ed. by Y. Bengio and Y. LeCun. URL: http://arxiv.org/abs/1409.1556.

Snoek, C. G. M., M. Worring, and A. W. M. Smeulders (2005). „Early versus Late Fusion in Semantic Video Analysis“. In: *Proceedings of the 13th Annual ACM International Conference on Multimedia*. MULTIMEDIA '05. Hilton, Singapore: Association for Computing Machinery, pp. 399–402. ISBN: 1595930442. DOI: 10.1145/1101149.1101236.

Stein, S. and S. J. McKenna (2013). „Combining Embedded Accelerometers with Computer Vision for Recognizing Food Preparation Activities“. In: *Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. UbiComp '13. Zurich, Switzerland: ACM, pp. 729–738. ISBN: 978-1-4503-1770-2. DOI: 10.1145/2493432.2493482.

Szegedy, C., Wei Liu, Yangqing Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich (2015). „Going deeper with convolutions“. In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–9. DOI: 10.1109/CVPR.2015.7298594.

Tanimoto, T. (1958). *An Elementary Mathematical Theory of Classification and Prediction*. Armonk, New York: International Business Machines Corporation.

Taylor, C. C., Z. Pizlo, J. P. Allebach, and C. A. Bouman (1997). „Image quality assessment with a Gabor pyramid model of the human visual system“. In: *Human Vision and Electronic Imaging II*. Vol. 3016. Proc. SPIE, pp. 58–69. DOI: 10.1117/12.274541.

Taylor, G. W., R. Fergus, Y. LeCun, and C. Bregler (2010). „Convolutional Learning of Spatio-temporal Features“. In: *Computer Vision – ECCV 2010*. Ed. by K. Daniilidis, P. Maragos, and N. Paragios. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 140–153. ISBN: 978-3-642-15567-3.

Taylor, P., J. N. Hobbs, J. Burroni, and H. T. Siegelmann (Dec. 2015). „The global landscape of cognition: hierarchical aggregation as an organizational principle of human cortical networks and functions“. In: *Scientific Reports* 5.1, p. 18112. ISSN: 2045-2322. DOI: 10.1038/srep18112.

Technische Richtlinien Abnahme von Bogenoffsetdruckmaschinen (2005). Tech. rep. Bundesverband Druck und Medien.

Tenorth, M., J. Bandouch, and M. Beetz (2009). „The TUM Kitchen Data Set of Everyday Manipulation Activities for Motion Tracking and Action Recognition“. In: *IEEE International Workshop on Tracking Humans for the Evaluation of their Motion in Image Sequences (THEMIS), in conjunction with ICCV2009*. IEEE, pp. 1089–1096.

Teo, P. and D. Heeger (1994). „Perceptual image distortion“. In: *IEEE Int. Conf. Image Processing ICIP-94*. Vol. 2. Vol. 2, 1994, pp. 982–986. ISBN: 0818669500. DOI: 10.1109/ICIP.1994.413502.

Tran, D., L. Bourdev, R. Fergus, L. Torresani, and M. Paluri (2015). „Learning Spatiotemporal Features with 3D Convolutional Networks“. In: *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*. ICCV '15. Washington, DC, USA: IEEE Computer Society, pp. 4489–4497. ISBN: 978-1-4673-8391-2. DOI: 10.1109/ICCV.2015.510.

Tschulakow, A. V., T. Oltrup, T. Bende, S. Schmelzle, and U. Schraermeyer (Mar. 2018). „The anatomy of the foveola reinvestigated“. eng. In: *PeerJ* 6. 4482[PII], e4482–e4482. ISSN: 2167-8359. DOI: 10.7717/peerj.4482. URL: https://doi.org/10.7717/peerj.4482.

Tuthill, J. C. and E. Azim (Mar. 2018). „Proprioception“. In: *Current Biology* 28.5, R194–R203. ISSN: 0960-9822. DOI: 10.1016/j.cub.2018.01.064.

Uijlings, J., I. C. Duta, E. Sangineto, and N. Sebe (Mar. 2015). „Video classification with Densely extracted HOG/HOF/MBH features: an evaluation of the accuracy/computational efficiency trade-off“. In: *International Journal of Multimedia Information Retrieval* 4.1, pp. 33–44. ISSN: 2192-662X. DOI: 10.1007/s13735-014-0069-5.

Uttal, W. (1975). *An Autocorrelation Theory of Form Detection*. John Wiley & Sons Inc. ISBN: 047089654X.

Varol, G., I. Laptev, and C. Schmid (2018). „Long-Term Temporal Convolutions for Action Recognition“. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40.6, pp. 1510–1517. DOI: 10.1109/TPAMI.2017.2712608.

Vepakomma, P., D. De, S. K. Das, and S. Bhansali (2015). „A-Wristocracy: Deep learning on wrist-worn sensing for recognition of user complex activities“. In: *2015 IEEE 12th International Conference on Wearable and Implantable Body Sensor Networks (BSN)*, pp. 1–6. DOI: 10.1109/BSN.2015.7299406.

Walse, K. H., R. V. Dharaskar, and V. M. Thakare (2016). „PCA Based Optimal ANN Classifiers for Human Activity Recognition Using Mobile Sensors Data“. In: *Proceedings of First International Conference on Information and Communication Technology for Intelligent Systems: Volume 1*. Ed. by S. C. Satapathy and S. Das. Cham: Springer International Publishing, pp. 429–436. ISBN: 978-3-319-30933-0.

Wandell, B. A. (1995). *Foundations of Vision:* Sunderland, Massachusetts: Sinauer Associates.

Wang, H., A. Kläser, C. Schmid, and C.-L. Liu (May 2013). „Dense Trajectories and Motion Boundary Descriptors for Action Recognition“. In: *International Journal of Computer Vision* 103.1, pp. 60–79. ISSN: 1573-1405. DOI: 10.1007/s11263-012-0594-8.

Wang, J., Y. Chen, S. Hao, X. Peng, and L. Hu (2019). „Deep learning for sensor-based activity recognition: A survey“. In: *Pattern Recognition Letters* 119. Deep Learning for Pattern Recognition, pp. 3–11. ISSN: 0167-8655. DOI: https://doi.org/10.1016/j.patrec.2018.02.010. URL: https://www.sciencedirect.com/science/article/pii/S016786551830045X.

Wang, L., Y. Qiao, and X. Tang (2015). „Action recognition with trajectory-pooled deep-convolutional descriptors“. In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4305–4314. DOI: 10.1109/CVPR.2015.7299059.

Wang, Z., A. C. Bovik, H. Sheikh, and E. P. Simoncelli (2004). „Image quality assessment: from error visibility to structural similarity“. In: *IEEE Transactions on Image Processing* 13.4, pp. 600–12. ISSN: 1057-7149.

Wang, Z., A. C. Bovik, and E. P. Simoncelli (2005). „Structural Approaches to Image Quality Assessment“. In: *Handb. Image Video Process.*, pp. 961–974. DOI: 10.1016/B978-012119792-6/50119-4.

Wang, Z., E. P. Simoncelli, and A. C. Bovik (2003). „Multiscale structural similarity for image quality assessment“. In: *The Thirty-Seventh Asilomar Conference on Signals, Systems & Computers, 2004*. Vol. 2. Vol. 2, 2003, pp. 1398–1402. ISBN: 0-7803-8104-1. DOI: 10.1109/ACSSC.2003.1292216.

Ward, J. (Mar. 1963). „Hierarchical Grouping to Optimize an Objective Function“. In: *Journal of the American Statistical Association* 58.301, pp. 236–244. URL: http://www.jstor.org/stable/2282967?seq=1.

Watson, A. B. (1979). „Probability summation over time“. In: *Vision Research* 19.5, pp. 515–522. ISSN: 0042-6989. DOI: https://doi.org/10.1016/0042-6989(79)90136-6. URL: https://www.sciencedirect.com/science/article/pii/0042698979901366.

Watson, A. B. and A. J. Ahumada Jr. (Oct. 2005). „A standard model for foveal detection of spatial contrast“. In: *Journal of Vision* 5.9, pp. 6–6. ISSN: 1534-7362. DOI: 10.1167/5.9.6.

Watson, A. B. and J. Solomon (1997). „Model of visual contrast gain control and pattern masking“. In: *J Opt Soc Am A Opt Image Sci Vis* 14.9, pp. 2379–91. ISSN: 1084-7529.

Weinland, D., R. Ronfard, and E. Boyer (Nov. 2006). „Free Viewpoint Action Recognition Using Motion History Volumes“. In: *Comput. Vis. Image Underst.* 104.2, pp. 249–257. ISSN: 1077-3142. DOI: 10.1016/j.cviu.2006.07.013.

Westen, S., R. Lagendijk, and J. Biemond (1995). „Perceptual image quality based on a multiple channel HVS model“. In: *1995 Int. Conf. on Acoustics, Speech, and Signal Processing ICASSP-95*. Vol. 4. Vol. 4, 1995, pp. 2351–2354. ISBN: 0-7803-2431-5. DOI: 10.1109/ICASSP.1995.479964.

Widrow, B. and M. E. Hoff (Aug. 1960). „Adaptive Switching Circuits“. In: *1960 IRE WESCON Convention Record, Part 4*. Institute of Radio Engineers. New York: Institute of Radio Engineers, pp. 96–104. URL: http://www-isl.stanford.edu/~widrow/papers/c1960adaptiveswitching.pdf.

Wiesel, T. N. and D. H. Hubel (1963). „Single-cell responses in striate cortex of kittens deprived of vision in one eye“. In: *Journal of Neurophysiology* 26.6. PMID: 14084161, pp. 1003–1017. DOI: 10.1152/jn.1963.26.6.1003.

Wiskott, L. (2009). „How Does Our Visual System Achieve Shift and Size Invariance?“ In: *23 Problems in Systems Neuroscience*. Ed. by J. van Hemmen and T. Sejnowski. New York: Oxford University Press. Chap. 16. ISBN: 9780195148220.

Wu, Z., L. Cai, and H. Meng (2005). „Multi-level Fusion of Audio and Visual Features for Speaker Identification“. In: *Advances in Biometrics*. Ed. by D. Zhang and A. K. Jain. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 493–499. ISBN: 978-3-540-31621-3.

Xia, L. and J. K. Aggarwal (2013). „Spatio-temporal Depth Cuboid Similarity Feature for Activity Recognition Using Depth Camera“. In: *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2834–2841. DOI: 10.1109/CVPR.2013.365.

Xu, K., J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio (July 2015). „Show, Attend and Tell: Neural Image Caption Generation with Visual Attention“. In: *Proceedings of the 32nd International Conference on Machine Learning*. Ed. by F. Bach and D. Blei. Vol. 37. Proceedings of Machine Learning Research. Lille, France: PMLR, pp. 2048–2057. URL: http://proceedings.mlr.press/v37/xuc15.html.

Yang, X. and Y. Tian (2014). „Super Normal Vector for Activity Recognition Using Depth Sequences“. In: *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 804–811. DOI: 10.1109/CVPR.2014.108.

Yang, Z., Y. Li, J. Yang, and J. Luo (2018). „Action Recognition with Visual Attention on Skeleton Images“. In: *CoRR* abs/1801.10304. arXiv: 1801.10304. URL: http://arxiv.org/abs/1801.10304.

Yarbus, A. L. (1967). *Eye Movements and Vision*. Springer US. ISBN: 978-1-4899-5379-7. DOI: 10.1007/978-1-4899-5379-7.

Zetzsche, C. and E. Barth (1990). „Fundamental limits of linear filters in the visual processing of two-dimensional signals“. In: *Vision Res* 30.7, pp. 1111–1117.

Zetzsche, C., K. Gadzicki, and T. Kluth (Apr. 2013). „Statistical Invariants of Spatial Form: From Local AND to Numerosity“. In: *Proceedings of the Second Interdisciplinary Workshop The Shape of Things*. CEUR-WS.org, pp. 163–172.

Zetzsche, C. and G. Hauske (1989a). „Multiple channel model for the prediction of subjective image quality“. In: *Human Vision, Visual Processing and Digital Display*. Vol. 1077. Proc. SPIE, Vol. 1077, 1989, pp. 209–216.

– (1989b). „Principal features of human vision in the context of image quality models“. In: *IEEE 3rd Int. Conf. Image Proc. and its Appl.* 1989, pp. 102–106.

Zetzsche, C. and G. Krieger (2001). „Nonlinear mechanisms and higher-order statistics in biological vision and electronic image processing: review and perspectives“. In: *J Electronic Imaging* 10.1, pp. 56–99.

Zetzsche, C. and U. Nuding (2005). „Nonlinear and higher-order approaches to the encoding of natural scenes“. In: *Network* 16.2–3, pp. 191–221.

Zetzsche, C., K. Schill, H. Deubel, G. Krieger, E. Umkehrer, and S. Beinlich (1998). „Investigation of a sensorimotor system for saccadic scene analysis: an integrated approach“. In: *Proceedings of the fifth international conference on simulation of adaptive behavior on From animals to animats 5*. Univ. of Zurich, Zurich, Switzerland: MIT Press, pp. 120–126. ISBN: 0-262-66144-6.

Zetzsche, C., J. Wolter, and K. Schill (May 2008). *Sensorimotor representation and knowledge-based reasoning for spatial exploration and localisation.*

Zetzsche, C., R. Rosenholtz, N. Cheema, K. Gadzicki, L. Fridman, and K. Schill (2017). „Neural Computation of Statistical Image Properties in Peripheral Vision“. In: *Computational and Mathematical Models in Vision (MODVIS)*.

Zhang, P., C. Lan, J. Xing, W. Zeng, J. Xue, and N. Zheng (2019). „View Adaptive Neural Networks for High Performance Skeleton-Based Human Action Recognition“. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41.8, pp. 1963–1978. DOI: 10.1109/TPAMI.2019.2896631.

Zhang, X. and B. A. Wandell (1997). „A spatial extension of CIELAB for digital color-image reproduction“. In: *Journal of the Society for Information Display* 5.1, pp. 61–63. ISSN: 1938-3657. DOI: 10.1889/1.1985127.

Zhou, Y. T. and R. Chellappa (1988). „Computation of optical flow using a neural network“. In: *IEEE 1988 International Conference on Neural Networks*, pp. 71–78. DOI: 10.1109/ICNN.1988.23914.

Zhou Wang and A. C. Bovik (2002). „A universal image quality index“. In: *IEEE Signal Processing Letters* 9.3, pp. 81–84. DOI: 10.1109/97.995823.

Zhu, J., W. Zou, L. Xu, Y. Hu, Z. Zhu, M. Chang, J. Huang, G. Huang, and D. Du (2018). „Action Machine: Rethinking Action Recognition in Trimmed Videos“. In: *CoRR* abs/1812.05770. arXiv: 1812.05770. URL: http://arxiv.org/abs/1812.05770.

Ziemba, C. M., J. Freeman, J. A. Movshon, and E. P. Simoncelli (May 2016). „Selectivity and tolerance for visual texture in macaque V2“. In: *Proc. Nat'l Academy of Sciences* 113.22, E3140–E3149. DOI: 10.1073/pnas.1510847113.

Accumulated Publications

Collected Papers

Prediction of the Perceived Quality of Streak Distortions in Offset-Printing with a Psychophysically Motivated Multi-channel Model

Konrad Gadzicki, Christoph Zetzsche

Universität Bremen
Kognitive Neuroinformatik
Enrique-Schmidt-Str. 5, 28359 Bremen
eMail: konny@informatik.uni-bremen.de,
zetzsche@informatik.uni-bremen.de
URL:http://www.informatik.uni-bremen.de/cog_neuroinf/

Abstract The evaluation of printing machines poses the problem of how distortions like streaks caused by the machine can be detected and assessed automatically. Although luminance variations in prints can be measured quite precisely, the measured functions bear little relevance for the lightness of streaks and other distortions of prints as perceived by human observers. First, the measurements sometimes indicate changes of luminance in regions which are perceived as homogeneous by humans. Second, the measured strength of a distortion correlates often weakly with its perceived strength, which is influenced by a variety of factors, like the shape of a streak's luminance profile and the distribution of luminance variations in its spatial surround. We have used a model of human perception, based on fundamental neurophysiological and psychophysical properties of the visual system, in order to predict the strength of streak distortions as perceived by human observers from a measured luminance signal. For the evaluation of the model, tests with naive and expert observers have been conducted. They showed that the model has a good correlation (> 0.8) to the assessments of human observers and is therefore suited for use in an automatic evaluation system.

1 Introduction

Modern offset printing machines are able to produce prints at high speed and quality. Nevertheless certain distortions, like streaks, are generated almost inevitably due to vibrations of the machine or an inadequate configuration. Streak distortions run orthogonally to the printing direction and result from slight shifts of the ink. All variations in the printing pattern can be captured metrologically very precisely, resulting in a signal in which the slightest changes of the density of a print are recorded. Unfortunately, the perception of humans deviates in several respects from the recorded signal. First, areas containing only small random fluctuations are perceived as homogeneous by humans. It is clear that such fluctuations, though measurable, have little practical relevance for the judgment of the quality of the printing process. Second, the perceived lightness of a streak differs substantially from its recorded intensity profile, both in dependence of the shape of the profile (as opposed to its mere height or contrast), and in dependence of the spatial surround of the streak (for example neighboring streaks or paper borders).
The evaluation of a printing machine thus requires a human expert in order to assess the prints with regard to the severity of streak distortions. Though procedures exist for metrological evaluation, the current methods either locate only severe distortions which are not arguable, while ignoring disputable distortions below their threshold which can still be visible to humans. Or, with the threshold lowered, they report relevant deviations in areas perceived as homogeneous by humans. In addition, they cannot take the surround of a streak into account for the evaluation of its strength. An automatic system for evaluation according to human perception essentially requires the incorporation of a model of the human visual system.
Several such models based on the properties of the human visual system have been used to assess image quality in the past [1][2][3][4]. These models commonly incorporate major properties like luminance invariance, sensitivity to frequency and orientations, and masking effects. Early models used a point-wise non-linearity or explicit gain modifications by a masking signal to account for masking effects while later vision models used divisive inhibition pooling acting over both spatial positions and frequency-selective channels [5]. The model suggested here takes these developments into account.

2 System and Model

The model is intended as part of a future system for the automatic evaluation of prints produced by SID[1]. It is designed to work with scanned images of printings, but can also be adapted to other input sources (e.g. densitometers). Hardware and software for scanning and preprocessing of the scans in the current study was provided by SID. The scans were generated with 5000 dpi which provides sufficient resolution for discrimination of closely neighboring streaks.

2.1 Model

On the top level, the model processes two inputs in parallel. As shown in Figure 1, signal+background and background alone are processed in the same manner by a system

[1]Sächsisches Institut für die Druckindustrie

explained in detail below. Streaks and other distortions are seen as the signal, while the background contributes to masking effects on this signal, e.g. by high-contrast luminance eges at the print borders. The local vector norm between the multi-channel outputs of the two pathways determines the final model output and thus the perceived strength of local distortion.

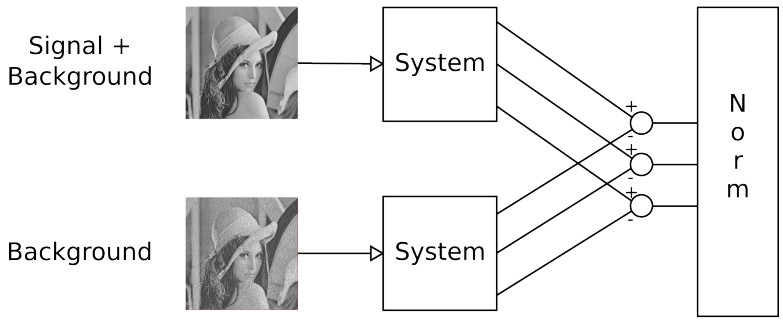


Figure 1: Model overview. The system (shown in detail in Figure 2) generates a multi-channel representation of both signal and signal+background. The distance between the two representations is assumed to be the perceived strength of the signal, and is computed by the difference norm shown at the right.

The system used for both pathways in Figure 1 is shown in detail in Figure 2. Its first stage is a luminance adaptivity stage, where local contrast is computed by a nonlinear multi-scale decomposition. In the next stage frequency- and orientation-selective linear band-pass filters are applied, leading to a further decomposition of each scale into its orientations. Finally, the filter outputs are normalized by local gain control mechanisms which pool over space, scales and orientations. These stages are now described in detail.

2.1.1 Luminance Invariance

The first step in the model is to transform the absolute luminance intensities. According to Weber's Law, the crucial variable for discrimination of luminance variations is contrast and not the absolute difference of luminance. The model uses the Ratio of Gaussian (ROG) operator [1] in order to calculate the contrast values. As the name suggests, the ROG is divisive operation of two low-pass inputs with different cut-off frequencies resulting in a non-linear band-pass output. Figure 3 illustrates the response of the operator.
For a computationally efficient implementation a pyramid scheme, similar to that in the Laplacian pyramid [6], is used to build the nonlinear representation of contrast, decomposed into several spatial scales.

2.1.2 Frequency- and Orientation Selective Filters

The majority of cells in early visual cortex are so-called Simple and Complex Cells which are responsive to patterns with a specific orientation and size. Orientation selectivity was the major finding of Hubel and Wiesel [7][8][9], and spatial-frequency selectivity was measured later in cats [10] and primates [11].
Mathematically the receptive field of such cells can be best described by Gabor functions [12], which offer the optimal resolution and selectivity in order to represent the response of

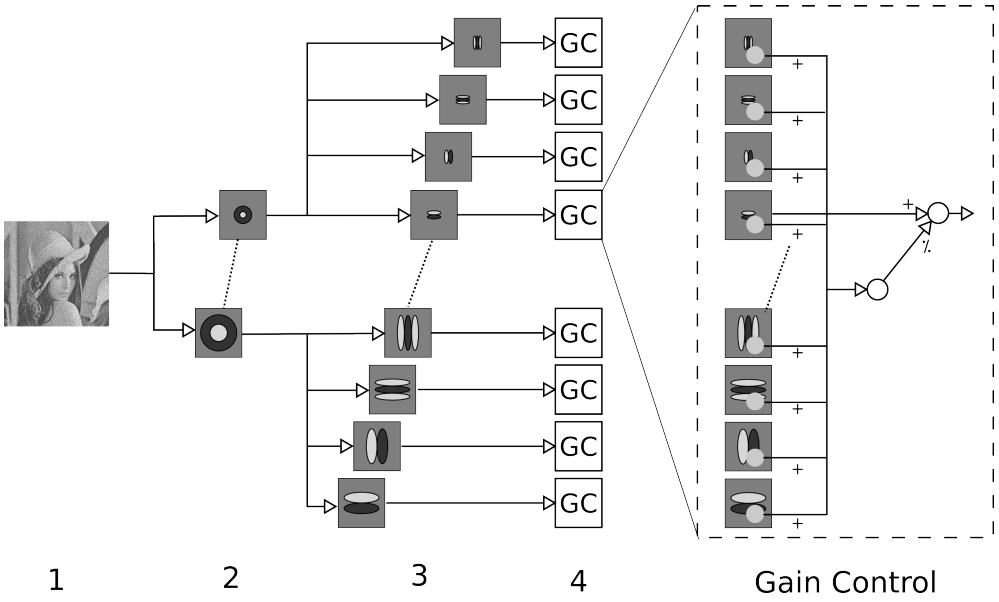
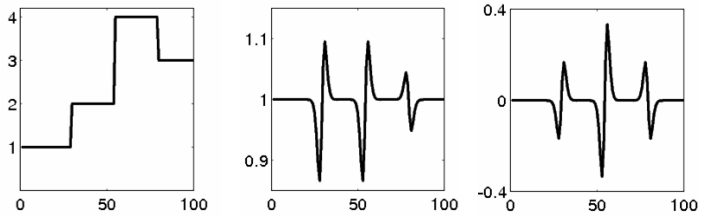


Figure 2: Overview over the system (as used for both pathways in Figure 1). From the input (1), the contrast is computed by non-linear ROG filters (2) and passed to a set of frequency- and orientation-selective linear filters (3). The outputs are then passed through gain control mechanisms (4). One of these is shown in detail on the right hand side. Hence each channel is normalized by spatial pooling over the other channels.



(a) (b) (c)

Figure 3: Response of a Ratio of Gaussian operator to luminance step edges. (a) input. (b) the ROG output represents luminance contrast. (c) linear DOG response shown for comparison.

V1 cells [13][14]. Gabor functions can be separated into even- and odd-symmetric parts, which respond best to saddles or edges respectively.
Though Gabor filters fit the psychophysical data well, they can be problematic in practice due to the DC component (zero frequency) of the even-symmetric part which does not drop to zero for filters with typical bandwidth. A Log-Gabor function [15], essentially a Gabor on a logarithmic scale, can be used to avoid this problem.

2.1.3 Masking

In visual perception the term masking usually refers to the reduced detectability of a stimulus in the presence of another stimulus. Such contrast masking has been modeled by a non-linear transducer function in early models [16]. Psychophysical experiments investigating the contrast response function of neurons from cats and monkeys [17] introduced the Naka-Rushton function, a hyperbolic ratio, as the best fitting description for neural response to contrast. Further investigations of the response properties of neurons revealed a pooling effect [18][19][20]. Masking effects can thus be explained by a suppressive signal derived by pooling over neurons. This mechanism is commonly referred to as Cortical Gain Control. It is modeled by a divisive pooling over space and channels and was included into recent visual system models, e.g. by Watson [5].

3 Methods

For the purpose of data acquisition, several experiments have been run. The first experiment employed naive observers with an on-screen display of distorted patterns in order to establish a baseline for the model in a noise-free environment. The only source of noise in this setup was effectively perceptual noise of the observers.
The second set of experiments were evaluations of real printings conducted by experts from the printing industry. In this setup several sources contributed to the noise of the system, namely the printing process itself, the scanner system, and perceptual noise of the observers.

3.1 Scale

Participants rated distortions on a qualitative grade scale similar to the EBU scale used in television picture quality assessment [21]. The grades used here ranged from 0 to 6 in 0.5 steps with 6 meaning a severe distortion and 1 a barely visible one. 0 was reserved for undetected distortions, thus it was not assigned by the participants directly but rather during the evaluation in case that participants did not see particular distortions in trials. The scale used here is reversed in comparison to the EBU scale and is actually the one used historically by the printing industry.

3.2 Experiments with Naive Observers

The experiments with naive observers used patterns presented on-screen only, in order to establish full control of the presentation and eliminate any additional sources of noise which are inevitably introduced in the full process of printing and scanning of patterns. The patterns were presented on analog screens with analog input which allowed for a luminance resolution equivalent to roughly 10 bits.
The set of patterns consisted of 30 images with a total of 75 streak distortions. For each participant three session were conducted. The whole set was presented during each session. The participants marked the position of a distortion by clicking with the mouse. The level of distortion (grade) and the classification (edge or saddle) were entered with the keyboard.

The distortions for the whole set were generated in a random fashion. This randomized set was used for all participants and sessions. The generation parameters included the number, position and type (Gauss or sharp edge) of distortions, width of the distortion, width of the edge, and contrast.
The participants were students with no prior involvement with the printing industry except for being exposed to printed products like books, magazines etc. in their daily life. They were not trained to perform this specific evaluation of printings task that was required in this experiment.
In total 21 participants participated in the experiment. Six of them were excluded from the evaluation because they either did not finish all sessions or had a significantly higher deviation of responses from the average.

3.3 Experiments with Experts from the Printing Industry

This experiment involved evaluation of real printings and was done by experts from various companies (Heidelberger Druckmaschinen, Koenig & Bauer, Manroland) from the printing industry.
In total, 52 printings were specially produced for the purpose of this experiment. Among them, 36 contained artificially generated distortions while the remaining 16 printings contained realistic distortions produced by an imprecisely configured printing machine. In the case of the artificial patterns, contrast, position, width of the distortion, and width of edges were varied.
The evaluation process was conducted under standardized conditions according to [22]. Each participant completed three sessions. The positions for the realistic-distortion patterns were fixed by the experts before the actual assessment. The positions of artificial distortions were known beforehand, but in the case of additional distortions as a result of the printing process, the positions were mapped manually by experts as well. The level of distortion at a specific position was recorded by an assistant. Details on the experiment can be found in [23].
In contrast to former experiments with on-screen presentation, several sources of noise influenced the results of this experimental setup. The printing process itself introduced considerable noise so that in the case of artificial patterns, the signal was noisy, even though the parameters used for generation were known. With regard to the realistic-distortion patterns, the actual signal of the pattern was not known at all.

3.4 Estimation of Model Parameters

For evaluation, model parameters were fitted to the data. The first approach was to use a simplex search [24] which turned out to be error-prone. This approach was not able to cope with the non-linear nature of this optimization problem, and produced unstable results. Finally a global search using particle swarm optimization [25][26][27] was used to fit the model parameters.

3.5 Evaluation

A local maximum search was used for comparison between model output and assessments. For the on-screen experiments this was due to the fact that the non-expert participants

were often sloppy with regard to clicking at the exact location of the distortion, so that the recorded positions were often several pixel off. In the case of the expert experiments with printings the positions were given in millimeters and were very precise, but the scans of the printings mapped roughly five pixels to a millimeter. Due to this the position in millimeters only indicated an area in the model output.
In order to assess participants' performance, the standard deviation of the participants' responses was calculated. The deviations of the responses indicate how stable the observers' assessments were.
The performance of the model was measured by its correlation with the participants' average assessment for each distortion.

4 Results

4.1 Stability of Assessments

Table 1 shows the standard deviations of the responses of the naive and expert observers. The deviations were calculated intra-individually, e.g. average for each participant, and inter-individually, thus average for each distortion. The intra-individually averaged responses provide information on the consistency of an observer with regard to his or her own responses, i.e. the extent to what the participant can reproduce his or her own responses over subsequent sessions. On the other hand the inter-individually average responses show how the participants agree with each others assessments.

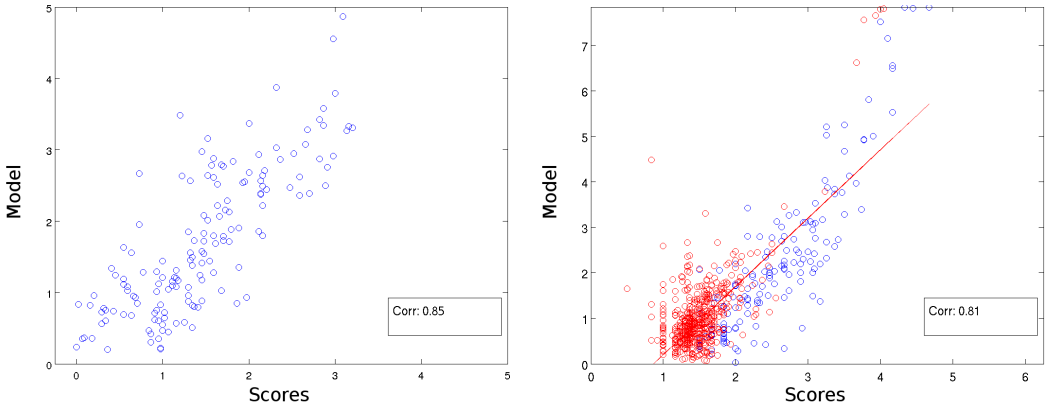
Table 1: Deviations of observers' assessments

	Naive	Expert
Intra-individually averaged	0.382	0.309
Inter-individually	0.707	0.454

One can see that both naive and expert observers were able to reproduce their own responses in a quite reliable way. Though the experts were slightly more consistent, the gap to the naive observers was not large and only differed by 0.073.
With regard to the deviation of responses over all participants for each distortion, the experts were much more consistent as a group. The experts were actually trained to assess distortions on an absolute scale due to their daily work which requires this particular skill. The naive observers on the other hand – even though consistent in their own opinion – have only received a brief explanation of the scale with examples for particular absolute grades. But with no further feedback on their own assessment, it is not surprising that the results of this group varied to a much larger degree.

4.2 Correlation between Model and Assessments

The correlation is shown for inter-individually averaged data where the average assessment for each distortion was mapped to the model output. Figure 4a shows the results for naive observers while figure 4b represents the expert observers.
Interestingly the naive observers achieved a higher correlation than the expert observers. With regard to the deviations of assessments (Table 1) this result is somewhat surprising.



(a) (b)

Figure 4: Correlation between model and inter-individually averaged data for (a) naive observers and (b) experts.

Though the processing pipeline for the experiment with printings introduced additional noise in comparison to the on-screen experiments, it is not clear whether this circumstance already accounts for the better results with naive observers.
A further difference between the two experiments is the respective set of test patterns. While in the on-screen experiments the strength of the streaks was approximately uniformly distributed over the grade scale, the expert experiment used a high percentage of realistic prints, which contain many barely visible streaks. Ratings performed close to the threshold of perception can be a problem, as shown below.

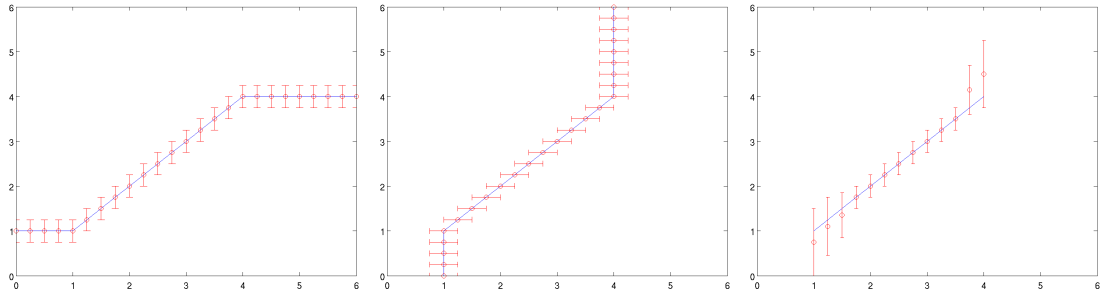
4.3 Patterns at Threshold of Perception

Printings are generally noisy with regard to the luminance signal. Therefore many weak streaks are difficult to detect even for trained observers. This is illustrated by the fact that naive observers who have performed the assessment task with realistic printings found approximately three streaks per print whereas the experts found up to 25. However, almost all additional streaks found by the expert group were rated with 1 (barely visible).

4.3.1 Illustration of the Problem

Figure 5a illustrates the distortions of the grades which can occur at the boundaries of the grade scale. First, in the lower part of the scale, near the threshold of perception, a range of different signal levels, from weakly visible to almost invisible signals, is mapped to grade 1. (Since the locations of the streaks are marked, they will typically receive a grade 1, even if critical testing without markings would presumably reveal that some participants are not able to see them.) Ideally there should be a proportionality between signal level and grade also in this regime, but since the grade 1 is used for even the least visible signal, the mapping runs into a plateau.
A similar problem can occur at the upper boundary of the grade scale as well, exaggerated by the fact that participants tend to avoid to use the worst grade, irrespective of how strong the streaks are. Hence the mapping of signal to grade scale can have a less steep

slope and may reach a plateau also in this range



(a) (b) (c)

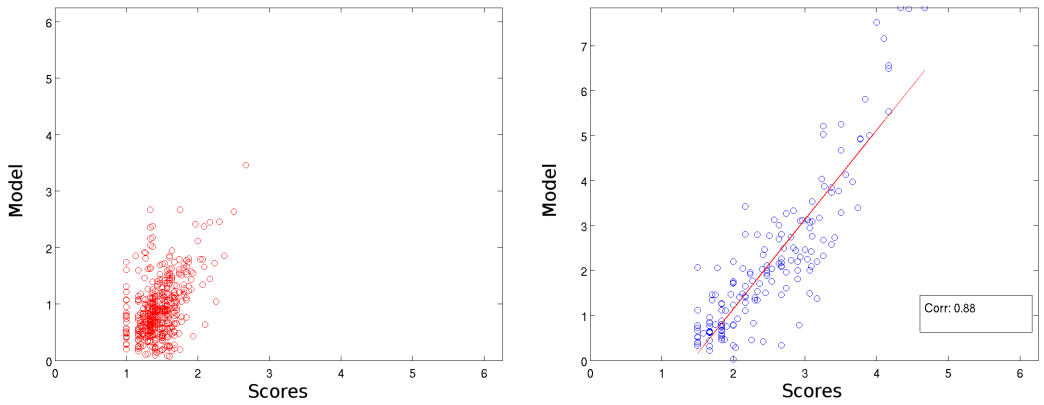
Figure 5: Visualization of distortions of the grade scale at the lower and upper boundary of the scale.
(a) hypothetical effective signal mapped to the grade scale (b) grades mapped to the ideal model. (c) apparent deviations of the model as caused by the grade scale distortion effects.

Figure 5b illustrates the problem from the point of view of an ideal model. For this the axes from figure 5a have to be switched. From the view of the model the minimum grade is assigned to range of values of the model output. At the upper end of the model outputs the same effect is visible, i.e. grades at the upper end will be mapped to a wider range of model outputs.
Figure 5c shows the expected effects with regard to the resulting apparent deviations of the ideal model from the participants' responses. In the interior range of the grade scale the model output corresponds well to the reponses of the participants. The deviations here reflect the "true" deviations of the model, while at the boundaries the apparent deviation of the model is artificially increased.

4.3.2 Correlation between Assessments and Model with Varying Percentage of Near-threshold Distortions

The correlations between model and participants' assessments for the whole set, including the afore-mentioned distortions near the threshold of perception, have already been presented in figure 4b. We now analyze the influence of those near-threshold responses on the overall performance.
Subsets of the data with a varying percentage of near-threshold responses have been investigated with regard to their effect on the correlation of the model predictions. The criterion for assigning a particular streak to the near-threshold group was the percentage of minimum grades given to that streak. As the minimum grade was effectively given to streaks which were barely visible or not seen at all by some observers, this criterion offers a flexible way to determine the subset of near-threshold streaks. The hardest version of the criterion is to mark a streak as near-threshold if at least one minimum grade has been assigned to it by one of the participants. In subsequent plots assessments of distortions are plotted in red if they contain a minimum grade response.
Figure 6a shows the subset consisting of inter-individually averaged assessments, which received at least one minimum grade response. As one can see, the apparent variance

of the model output is quite high, which is in correspondence to the aforementioned hypothesis taht in this regime different signal levels are mapped to the minimum grade. The omission of this subset, i.e. of streaks which have received at least one minimum grade from the evaluation, leads to an increase in the correlation for the averaged data from 0.81 to 0.88, as shown in Figure 6b. Note that although the overall dynamic range of the data is reduced, the correlation increases. This corroborates our hypothesis that the excluded subset is not well behaved.



(a) (b)

Figure 6: Correlation between model and inter-individually averaged assessments for ratings not containing the minimum grade (b) and for ratings containing at least one minimum grade response (a).

5 Conclusion

We have presented a model for the automatic prediction of the perceived strength of streak distortions in offset printing. The model is based on recent neurophysiological and psychophysical results. It can describe shape effects due to the shape of the luminance function of a streak and masking effects as caused by the configuration in its spatial surround.

In experiments with naive and expert observers we have collected assessments on a large number of streak patterns. We have then shown that the model shows a good correlation in predicting the perceived strength of these distortions.

The prediction of distortions close to the perceptual threshold proved to be problematic, resulting in a big deviation of model responses for barely visible distortions. The removal of distortions which were rated with the minimum grade increased the correlation of the model predictions, although the dynamic range of the data has been reduced by this removal.

This model of the human visual system seems thus to be suitable for the automatic evaluation of printings.

Acknowledgement

This work was supported by NCT Bremen, Sächsiches Institut für die Druckindustrie Leipzig and Forschungsgesellschaft Druckmaschinen e.V. Frankfurt am Main.

References

[1] Christoph Zetzsche and G Hauske. Multiple channel model for the prediction of subjective image quality. In *Human Vision, Visual Processing and Digital Display*, volume 1077 of *Proc. SPIE*, pages 209–216, 1989.

[2] Christoph Zetzsche and G Hauske. Principal features of human vision in the context of image quality models. In *IEEE 3rd Int. Conf. Image Proc. and its Appl*, Proc. of IEEE, pages 102–106, 1989.

[3] Scott Daly. The visible differences predictor: an algorithm for the assessment of image fidelity. In Andrew B. Watson, editor, *Digital images and human vision*, pages 179–206. MIT Press, Cambridge, MA, USA, 1993.

[4] Jeffrey Lubin. The use of psychophysical data and models in the analysis of display system performance. In Andrew B Watson, editor, *Digital images and human vision*, pages 163–178. MIT Press, Cambridge, MA, USA, 1993.

[5] A.B. Watson and J.A. Solomon. Model of visual contrast gain control and pattern masking. *Journal of the Optical Society of America.*, 14(9):2379–91, September 1997.

[6] P. Burt and E. Adelson. The Laplacian Pyramid as a Compact Image Code. *IEEE Transactions on Communications*, 31(4):532–540, April 1983.

[7] D.H. Hubel and TN Wiesel. Receptive fields of single neurones in the cat's striate cortex. *The Journal of Physiology*, 148(3):574–591, 1959.

[8] D.H. Hubel and TN Wiesel. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of Physiology*, 160(1):106–154, 1962.

[9] D.H. Hubel and TN Wiesel. Receptive fields and functional architecture of monkey striate cortex. *The Journal of Physiology*, 195(1):215–243, 1968.

[10] F. W. Campbell, G. F. Cooper, and Christina Enroth-Cugell. The spatial selectivity of the visual cells of the cat. *The Journal of Physiology*, 203(1):223–235, 1969.

[11] R.L. De Valois, K.K. De Valois, J. Ready, and H. Blanckensee. Spatial frequency tuning of macaque striate cortex cells. *Assoc. Res., Vision Ophthalmol*, 15, 1975.

[12] D Gabor. Theory of communication. Part 1: The analysis of information. *Engineers-Part III: Radio and Communication*, 1946.

[13] S Marcelja. Mathematical description of the responses of simple cortical cells. *Journal of the Optical Society of America*, 70(11):1297–300, November 1980.

[14] J G Daugman. Spatial visual channels in the Fourier plane. *Vision Research*, 24(9):891–910, January 1984.

[15] David J Field. Relations between the statistics of natural images and the response properties of cortical cells. *Journal of the Optical Society of America.*, 4(12):2379–94, December 1987.

[16] G E Legge and J M Foley. Contrast masking in human vision. *Journal of the Optical Society of America*, 70(12):1458–71, December 1980.

[17] Duane G. Albrecht and David B. Hamilton. Striate Cortex of Monkev and Cat: Function Contrast Response. *Journal of Neurophysiology*, 48(l), 1982.

[18] GC DeAngelis, JG Robson, I Ohzawa, and RD Freeman. Organization of suppression in receptive fields of neurons in cat visual cortex. *Journal of Neurophysiology*, 68:144–163, 1992.

[19] WS Geisler and DG Albrecht. Cortical neurons: isolation of contrast gain control. *Vision Research*, 32(8):1409–1410, 1992.

[20] DJ Heeger. Normalization of cell responses in cat striate cortex. *Visual Neuroscience*, 9:181–198, 1992.

[21] ITU. Recommendation ITU-R BT.500-11, Methodology for the subjective assessment of the quality of television pictures. *International Telecommunication Union, Geneva*, pages 1–48, 2002.

[22] Norm ISO 3664:2009-04. Technical report, International Organization for Standardization, April 2009.

[23] Versuchsdurchführung bei der Bewertung von Druckbogen. Technical report, FOGRA, 2010.

[24] J C Lagarias, J A Reeds, M H Wright, and P E Wright. Convergence properties of the Nelder-Mead simplex method in low dimensions. *SIAM Journal of Optimization*, 9:112–147, 1998.

[25] James Kennedy and R. Eberhart. Particle swarm optimization. In *Proceedings of ICNN'95 - International Conference on Neural Networks*, volume 4, pages 1942–1948. IEEE, 1995.

[26] Kennedy J, RC Eberhart, and YH Shi. *Swarm Intelligence*. Academic Press, 2001.

[27] Said M. Mikki and Ahmed a. Kishk. *Particle Swarm Optimizaton: A Physics-Based Approach*, volume 3. January 2008.

[28] R L De Valois, E W Yund, and N Hepler. The orientation and direction selectivity of cells in macaque visual cortex. *Vision Research*, 22(5):531–544, 1982.

[29] AB Watson. Detection and recognition of simple spatial forms. In O. J. Braddick and A. C. Sleigh, editors, *Physical and Biological Processing of Images*, 1983.

Journal of Modern Optics
Vol. 00, No. 00, 10 January 2010, 1–12

Prediction of the Perceived Quality of Streak Distortions in Offset-Printing with a Psychophysically Motivated Multi-channel Model

Konrad Gadzicki† and Christoph Zetzsche

Cognitive Neuroinformatics, University of Bremen, Germany;
(*Received 00 Month 200x; final version received 00 Month 200x*)

The evaluation of printing machines poses the problem of how distortions like streaks caused by the machine can be detected and assessed automatically. Although luminance variations in prints can be measured quite precisely, the measured functions bear little relevance for the lightness of streaks and other distortions of prints as perceived by human observers. First, the measurements sometimes indicate changes of luminance in regions which are perceived as homogeneous by humans. Second, the measured strength of a distortion correlates often weakly with its perceived strength, which is influenced by a variety of factors, like the shape of a streak's luminance profile and the distribution of luminance variations in its spatial surround. We have used a model of human perception, based on fundamental neurophysiological and psychophysical properties of the visual system, in order to predict the perceptual strength of streaks (i.e. the distortion as perceived by a human observer) from the measured physical luminance signal. For the evaluation of the model, tests with naive and expert observers have been conducted. The results show that the model yields a good correlation (> 0.8) to the assessments of human observers and is thus well suited for use in an automatic evaluation system.

Keywords: image quality; offset printing; human visual system; quality control

1. Introduction

Modern offset printing machines can rapidly produce prints at high quality. Nevertheless certain distortions, like streaks, are generated almost inevitably due to vibrations of the machine or an inadequate configuration. Streak distortions run orthogonally to the printing direction and result from slight shifts of the ink. All variations in the printing pattern can be captured metrologically very precisely, resulting in a signal in which the slightest changes of the density of a print are recorded. Unfortunately, the perception of humans deviates in several respects from the recorded signal. First, areas containing only small random fluctuations are perceived as homogeneous by humans. It is clear that such fluctuations, though measurable, have little practical relevance for the judgement of the quality of the printing process. Second, the perceived lightness of a streak differs substantially from its recorded intensity profile, both in dependence of the shape of the profile (as opposed to its mere height or contrast), and in dependence of the spatial surround of the streak (for example neighbouring streaks or paper borders).

The evaluation of a printing machine thus requires a human expert in order to assess the prints with regard to the severity of streak distortions. Since this is a tedious process and bears the risk of a subjective bias of the expert there have been attempts to create a standardized and automated evaluation method. The method currently used in the German printing industry uses the difference in CIELAB colour ΔE^* every 2.5mm over the length of a print for the calculation of the distortion strength and applies a threshold to classify them as annoying or not [1, 2]. In an

†Corresponding author. Email: konny@informatik.uni-bremen.de

ongoing discussion and evaluation several print-related research institutions in Germany (FGD, SID, FOGRA) have analysed the merits and drawbacks of this standardized procedure. Various modifications of the spatial difference in CIELAB colours have been tested, for example by SID. These brought some improvements, but showed to be not accurate enough in their prediction of the perception of human experts.

The current methods either locate only severe distortions which are not arguable, while ignoring disputable distortions below their threshold which can still be visible to humans. Or, with the threshold lowered, they report relevant deviations in areas perceived as homogeneous by humans. In addition, they cannot take the surround of a streak into account for the evaluation of its strength. It has been concluded by the research institutions that an automatic system for evaluation of perceived streak distortions essentially requires the incorporation of a model of the human visual system (HVS). This prompted the present investigation and the associated development of the model suggested here.

Several image quality models based on the properties of the human visual system have been developed in the past [3–10]. These models commonly incorporate major properties like luminance invariance, sensitivity to frequency and orientations, and masking effects. Early models used a point-wise non-linearity or explicit gain modifications by a masking signal to account for masking effects. More recent models of the visual system used divisive inhibition pooling acting over both spatial positions and frequency-selective channels [11]. The model suggested here takes these developments into account.

In addition to the aforementioned systems, there exists a large number of approaches which are not solely based on the HVS. These approaches include feature based scales, e.g. [12, 13], multi-dimensional impairment scales, e.g. [14], a spatial extension to CIELAB [15], structural similarity index [16, 17], image information fidelity [18] or most apparent distortion [19]. Such approaches can be of special advantage if one model should be applied to heterogeneous stimulus domains. See [20] for a recent comparison of a number of different approaches. Beside the problem of modelling human perception, there are also cognitive issues which can influence the human judgements [21]. Such issues can be hoped to be not as critical in the specific context of the present investigation, where trained specialists operate in a quite restricted domain.

2. System and Model

The model is intended as part of a future system for the automatic evaluation of prints produced by SID (Sächsisches Institut für die Druckindustrie). It is designed to work with scanned images of printings, but can also be adapted to other input sources (e.g. densitometers).

2.1. *Input and Devices*

The test patterns used for the evaluation of a print machine are homogeneous prints in a specific base colour, usually cyan, which are printed with a accurately defined machine configuration. Figure 1 shows an example of a print with streak distortions.

Hardware and software for scanning and preprocessing of the scans in the current study was provided by SID. The scanner is a RGB scanner by Colortrac which has been calibrated in a similar way to ICC. The reference image consisted of cyan patches of varying density since this is the only relevant colour for the purpose of evaluation. The CIELAB values of the patches were measured with a colorimeter and in a second step a quadratic function was fitted to adjust the transformed scanner's elements' RGB values to match the measured CIELAB values.

The prints were scanned twice in opposite direction in order to counter systematic bias from sampling in one direction. The prints contain markers which allow to match the images after the scan process. The matching was not done with sub-pixel accuracy since this level of accu-

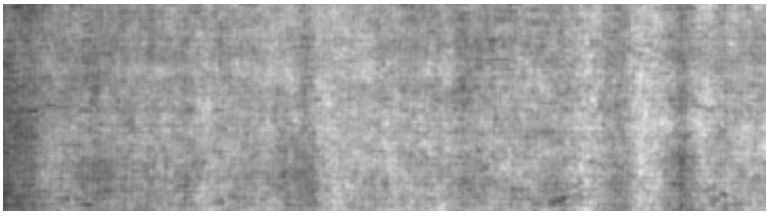


Figure 1. Example of streak distortions. The image shows an extract from a scan, the contrast has been increased dramatically. The printing direction runs horizontally, the streak distortions are thus vertical.

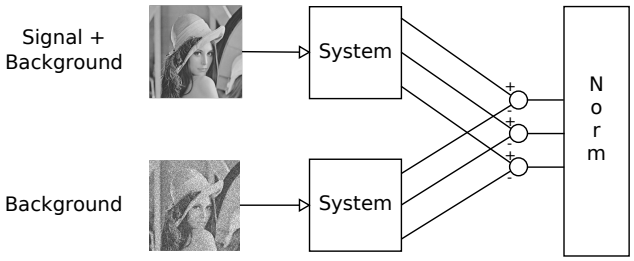


Figure 2. Model overview. The system (shown in detail in Figure 3) generates a multi-channel representation of both signal and signal+background. The distance between the two representations is assumed to be the perceived strength of the signal, and is computed by the difference norm shown at the right.

racy is not required for detection of streaks which are much more coarse. The images are then preprocessed by de-screening and correction of row and column shading, averaging over several pixels. A side effect of the preprocessing is that two-dimensional signal variations are reduced in adaption to the one-dimensional nature of the streak patterns.

2.2. *Model*

On the top level, the model processes two inputs in parallel. As shown in Figure 2, signal + background and background alone are processed in the same manner by a system explained in detail below. Streaks and other distortions are seen as the signal, while the background contributes to masking effects on this signal, e.g. by high-contrast luminance edges at the print borders. A local vector norm between the multi-channel outputs of the two pathways determines the final model output and thus the perceived strength of a local distortion. The system used for both pathways in Figure 2 is shown in detail in Figure 3. Its first stage is a luminance adaptivity stage, where a multi-scale presentation of local contrast is computed by a nonlinear filter decomposition. In the next stage frequency- and orientation-selective linear band-pass filters are applied, leading to a further decomposition of each scale into its orientations. Finally, the filter outputs are normalized by local gain control mechanisms which pool over space, scales and orientations.

For the evaluation of the model, those parameters of the model which are expected to be specifically related to the task were fitted with Particle Swarm optimization [22]. Other parameters, like the filter bandwidths for the ROG and Log-Gabor filters, and the orientation selectivity were set to the typical values known from psychophysical and neurobiological research (see below). The contrast sensitivity function was fitted indirectly by the Log-Gabor filter amplitudes for different scales. A further parameter that has been fitted is the exponent of the vector norm.

The stages of the model are now described in detail.

2.2.1. Luminance Invariance

The first step in the model is to transform the absolute luminance intensities. According to Weber's Law, the crucial variable for discrimination of luminance variations is contrast and

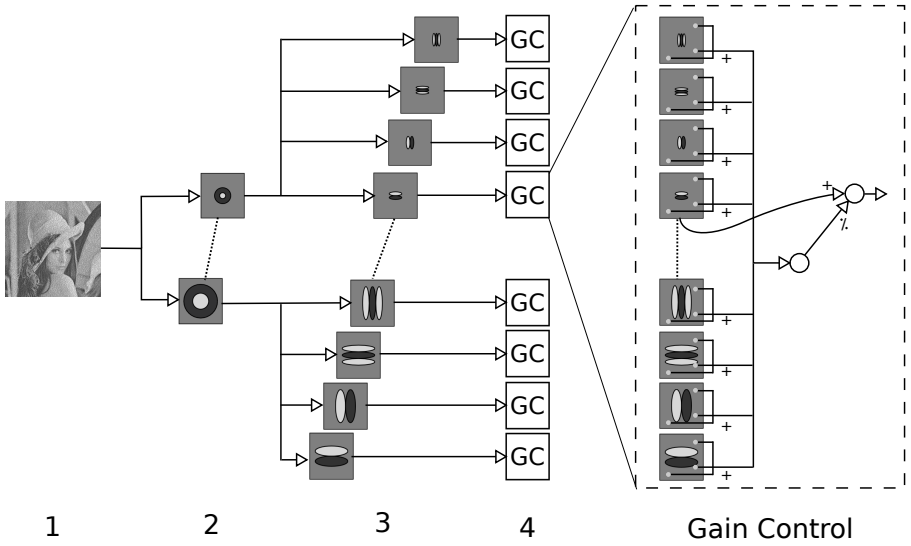


Figure 3. Overview over the system (as used for both pathways in Figure 2). From the input (1), the contrast is computed by non-linear ROG filters (2) and passed to a set of frequency- and orientation-selective linear filters (3). The outputs are then passed through gain control mechanisms (4). One of these is shown in detail on the right hand side. Hence each channel is normalized by spatial pooling over the other channels.

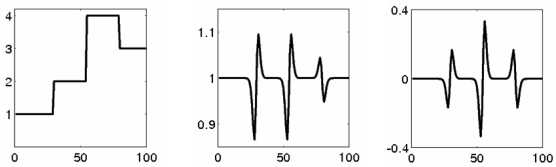


Figure 4. Response of a Ratio of Gaussian operator to luminance step edges. (a) input. (b) the ROG output represents luminance contrast. (c) linear DOG response shown for comparison.

not the absolute difference of luminance. Contrast is defined as the ratio between luminance and (local) background (average luminance of the surrounding). The model is intended for luminance as input but it can also be used with the L^* channel from a signal in CIELAB colour space so that the lightness values correlate already better with human perception. The current data set provides such L^* values. Although Weber's Law is to a certain degree incorporated in the L^* scale, our generalized model incorporates a local contrast adaptation stage which is able to model local effects that cannot be captured by the global normalization of the CIELAB scale.

The model uses the Ratio of Gaussian (ROG) operator [3] in order to calculate the contrast values. As the name suggests, the ROG is divisive operation of two low-pass inputs with different cut-off frequencies resulting in a non-linear band-pass output (1).

$$g_{i+1}(x,y) = \frac{l(x,y) * \frac{1}{2\pi\sigma_i^2} \exp\left(-\left(\frac{x^2+y^2}{2\sigma_i^2}\right)\right)}{\left(l(x,y) * \frac{1}{2\pi\sigma_{i+1}^2} \exp\left(-\left(\frac{x^2+y^2}{2\sigma_{i+1}^2}\right)\right)\right) + c}. \tag{1}$$

In the model the sigma ratio is $\sigma_{i+1} = 2\sigma_i$, starting with $\sigma_0 = 8$ pixel. The constant c is set to 2.0. Figure 4 illustrates the response of the operator. For a computationally efficient implementation a pyramid scheme, similar to that in the Laplacian pyramid [23], is used to build the nonlinear representation of contrast, decomposed into five spatial scales.

2.2.2. Frequency- and Orientation Selective Filters

The majority of cells in early visual cortex are so-called Simple and Complex Cells which are responsive to patterns with a specific orientation and size. Orientation selectivity was the major finding of Hubel and Wiesel [24–26], and spatial-frequency selectivity was measured later in cats [27] and primates [28]. Mathematically the receptive field of such cells can be best described by Gabor functions [29], which offer the optimal resolution and selectivity in order to represent the response of V1 cells [30, 31]. The 2D kernel (2) of a Gabor filter consists of a complex sinusoid (3) weighted by a Gaussian function (4):

$$h(x, y) = g(x, y)s(x, y) \tag{2}$$

with

$$s(x, y) = \exp(-j2\pi u_0 x) \tag{3}$$

$$g(x, y) = \frac{1}{2\pi\sigma_x\sigma_y} \exp\left(-\pi\left(\frac{x^2}{2\sigma_x^2} + \frac{y^2}{2\sigma_y^2}\right)\right) \tag{4}$$

Here x, y are the spatial coordinates, σ_x, σ_y the deviations of the Gaussian function and u_0 the wave length of the sinusoid. The orientation of the filter kernel can be achieved by a rotation of the coordinate system.

For a representation in the frequency domain, (2) can be transformed, resulting in the filter function

$$H(u, v) = \exp\left(-\pi\frac{(u - u_0)^2}{(2\sigma_u)^2} + \pi\frac{v^2}{(2\sigma_v)^2}\right) \tag{5}$$

where u, v are the frequencies corresponding to x, y, u_0 is the centre frequency of the filter and $\sigma_u = \frac{1}{2\pi\sigma_x}$, $\sigma_v = \frac{1}{2\pi\sigma_y}$ are the bandwidths in x- and y-direction. The centre frequency u_0 places the centre of the Gaussian on a specific frequency. σ_u, σ_v define the cut-off frequency where the function value declines to 0.465, effectively determining the bandwidth in frequency and orientation.

Though Gabor filters fit the psychophysical data well, they can be problematic in practice due to the DC component (zero frequency) of the even-symmetric part which does not drop to zero for filters with typical bandwidth. A Log-Gabor function [32], essentially a Gabor on a logarithmic scale, can be used to avoid this problem. The model uses a Log-Gabor in polar coordinates (6)

$$H_{log}(\rho, \phi) = j^k \exp\left(-\frac{(\log\frac{\rho}{\rho_0})^2}{2(\log\sigma_\rho)^2}\right)\left[-\exp\left(\frac{(\phi - \phi_0)^2}{2\sigma_\phi^2}\right) - (-1^k)\exp\left(\frac{(\phi - \phi_0 - \pi)^2}{2\sigma_\phi^2}\right)\right] \tag{6}$$

with ρ being the radial frequency and ϕ the angle. The preferred frequency and orientation of this filter function is given by ρ_0 and ϕ_0. Bandwidth and orientation selectivity are determined by σ_ρ and σ_ϕ. The even symmetric function is real valued ($k = 0$) and the odd symmetric filter function imaginary ($k = 1$). The amplitudes of the filters used for the five scales returned by the optimization are 4314.25, 5420.88, 3915.05, 3305.01 and 3115.31. For each level of the pyramid structure, the preferred frequency ρ_0 is matched to corresponding cut-off frequency σ_{i+1} of the denominator of the ROG operation. The bandwidths used in the model are not subject to optimization but are fixed and follow the values known from neurobiology: 1 octave for frequencies [33] and 30° for orientations [34].

The model uses an orientation-selective filter decomposition since it should not be restricted to the processing of 1-D signals but should allow the evaluation of all types of distortions in two-dimensional prints. It should be noted that even the streak patterns considered here are not strictly one-dimensional signals. In some cases the width and the profile of a streak shows small but visible variations along the length of the streak. Such streaks are easier to detect and are perceived as stronger than the corresponding perfectly straight streaks. The deviations from straightness can in principle be captured by a model with multiple orientation channels since they cause additional activity in neighboring orientation channels. However, our investigations revealed that this effect can not be thoroughly evaluated with the current data set. In the preprocessing applied with the scanning, the averaging in streak direction smoothed out the two-dimensional variations such that only insufficient energy remained in neighbouring orientation channels. With the current data, the model thus acts as an essentially one-dimensional model, but with other distortions and/or recording methods it acts as a 2-D model.

2.2.3. Masking

In visual perception the term masking usually refers to the reduced detectability of a stimulus in the presence of another stimulus. Such contrast masking has been modelled by a non-linear transducer function in early models [35]. Psychophysical experiments investigating the contrast response function of neurons from cats and monkeys [36] introduced the Naka–Rushton function, a hyperbolic ratio, as the best fitting description for neural response to contrast. Further investigations of the response properties of neurons revealed a pooling effect [37–39]. Masking effects can thus be explained by a suppressive signal derived by pooling over neurons, a mechanism commonly referred to as Cortical Gain Control. It is modelled by a divisive pooling over space and channels (7) and was included into recent visual system models, e.g. [11]. In the model pooling is done over neighbouring scales and over all orientations of these scales. Pooling over space is performed by a convolution with a low-pass filter kernel.

$$\bar{r}_{s,o}(x,y) = \frac{r_{s,o}(x,y)^p}{c^q + \sum_{\bar{s}=s-1}^{s+1} \sum_{\bar{o}=1}^{6} (r_{\bar{s},\bar{o}} * k_{\bar{s},\bar{o}})(x,y)^q} \tag{7}$$

The responses r are the outputs of linear filters, indexed by s for scale and o for orientation. As for the exponents used by the model, $p = 2.4622$ was returned by the optimization and $q = 2$ was fixed. The constant c defines the point where saturation begins which was set to 0.5 in the model. The convolution operation is denoted by $*$ with $k(x, y)$ being the kernel. The model uses a Gaussian low pass kernel with 0.0625 as cut-off frequency (Fourier domain based convolution).

2.2.4. Vector Norm

The final result is the local vector norm (Minkowski distance) between the two multi-channel pathways:

$$d = \left(\sum |\bar{r}_{\text{signal+background}} - \bar{r}_{\text{background}}|^p \right)^{1/p} \tag{8}$$

The optimization returned $p = 2.5055$ as the best fitting exponent.

3. Methods

For the purpose of data acquisition, several experiments have been run. The first experiment employed naive observers with an on-screen display of distorted patterns in order to establish a

baseline for the model in a noise-free environment. The only source of noise in this setup was effectively perceptual noise of the observers. The second set of experiments were evaluations of real printings conducted by experts from the printing industry. In this setup several sources contributed to the noise of the system, namely the printing process itself, the scanner system, and perceptual noise of the observers.

3.1. *Scale*

Participants rated distortions on a qualitative grade scale similar to the EBU scale used in television picture quality assessment [40]. The grades used here ranged from 0 to 6 in 0.5 steps with 6 meaning a severe distortion and 1 a barely visible one. 0 was reserved for undetected distortions, thus it could not be chosen by the participants directly but was assigned during the evaluation in case that participants did not see particular distortions in certain trials. The scale used here is reversed in comparison to the EBU scale and is actually the one used historically by the printing industry.

3.2. *Experiments with Naive Observers*

The experiments with naive observers used patterns presented on-screen only, in order to establish full control of the presentation and eliminate any additional sources of noise which are inevitably introduced in the full process of printing and scanning of patterns. The patterns were presented on analogue screens with analogue input which allowed for a luminance resolution equivalent to roughly 10 bits. The set of patterns consisted of 30 images with a total of 75 streak distortions. For each participant three session were conducted. The whole set was presented during each session. The participants marked the position of a distortion by clicking with the mouse and entered the level of distortion (grade) with the keyboard. The distortions for the whole set were generated in a random fashion. This randomized set was used for all participants and sessions. The generation parameters included the number, position and type (Gauss or sharp edge) of distortions, width of the distortion, width of the edge, and contrast.

The participants were students with no prior involvement with the printing industry (except for being exposed to printed products like books, magazines etc. in their daily life). They had no experience with an evaluation of printings task as required in this experiment. In total 21 participants participated in the experiment. Six of them were excluded from the evaluation because they either did not finish all sessions or had a significantly higher deviation of responses from the average.

3.3. *Experiments with Experts from the Printing Industry*

This experiment involved evaluation of real printings and was done by experts from various companies from the printing industry (Heidelberger Druckmaschinen, Koenig & Bauer, Manroland). In total, 52 printings were specially produced for the purpose of this experiment. Among them, 36 contained artificially generated distortions while the remaining 16 printings contained realistic distortions produced by an imprecisely configured printing machine. In the case of the artificial patterns, contrast, position, width of the distortion, and width of edges were varied.

The evaluation process was conducted under standardized conditions according to [41]. It requires the printed matter to be illuminated by a standard illuminant D50, the surround and backing of the matter to be neutral and matt, and the visual environment to be arranged in a way that interference with the viewing task is minimized. In practice a special arrangement (tiltable table with grey, matt surface and standard illuminant) is used in an environment shielded against external light. Each participant completed three sessions. The positions for the

Table 1. Deviations of observers' assessments

	Naive	Expert
Intra-individually averaged	0.382	0.309
Inter-individually	0.707	0.454

realistic-distortion patterns were fixed by the experts before the actual assessment. The positions of artificial distortions were known beforehand, but in the case of additional distortions as a result of the printing process, the positions were mapped manually by experts as well. The level of distortion at a specific position was recorded by an assistant. Details on the experiment can be found in [42].

In contrast to former experiments with on-screen presentation, several sources of noise influenced the results of this experimental setup. The printing process itself introduced considerable noise so that in the case of artificial patterns, the signal was noisy, even though the parameters used for generation were known. With regard to the realistic-distortion patterns, the actual signal of the pattern was not known at all.

3.4. *Evaluation*

A local maximum search was used for mapping assessments to the corresponding positions in the model output. For the on-screen experiments this was due to the fact that the non-expert participants were often sloppy with regard to clicking at the exact location of the distortion, so that the recorded positions were often several pixel off. In the case of the expert experiments with printings the positions were given in millimetres and were very precise, but the scans of the printings mapped roughly five pixels to one millimetre. Due to this the position in millimetres only indicated an area in the model output.

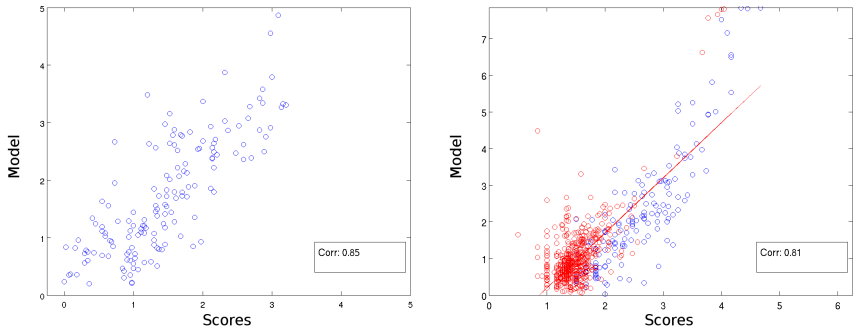
In order to assess participants' performance, the standard deviation of the participants' responses was calculated. The deviations of the responses indicate how stable the observers' assessments were. The performance of the model was measured by its correlation with the participants' average assessment for each distortion.

4. Results

4.1. *Stability of Assessments*

Table 1 shows the standard deviations (s.d.) of the responses of the naive and expert observers. The deviations were calculated intra-individually (average for each participant) and inter-individually (average for each distortion). The intra-individually averaged responses provide information on the consistency of an observer with regard to his or her own responses, i.e. the extent to which the participant can reproduce his or her own responses over subsequent sessions. The inter-individually averaged responses, on the other hand, show how the participants agree with each others assessments.

One can see that both naive and expert observers were able to reproduce their own responses in a quite reliably. The experts were slightly more consistent, but the difference to the naive observers was only 0.073. With regard to the inter-individual deviations (deviations of responses over all participants for each distortion), the experts were much more consistent as a group. The experts were actually trained to assess distortions on an absolute scale, since their daily work requires this particular skill. The naive observers, on the other hand — even though consistent in their own opinion — have only received a brief explanation of the scale with some examples of the distortion levels to be expected. Different cognitive strategies [21] can thus be expected to be present in this group. With no further feedback on the ratings of the other subjects it is



(a) (b)

Figure 5. Correlation between model and inter-individually averaged data for (a) naive observers and (b) experts.

thus not surprising that the results of this group varied to a much larger degree.

4.2. *Correlation Between Model and Assessments*

The model parameters were fitted with 3-fold cross-validation in order to check for effects of over-fitting. The data was split according to random subsets of subjects. The maximum correlations achieved in individual runs were 0.811 and 0.805 without indications of over-fitting. For the results presented here the model has been fitted with the whole data set.

The correlation is shown for inter-individually averaged data where the average assessment for each distortion was mapped to the model output. Figure 5(a) shows the results for naive observers while Figure 5(b) represents the expert observers. Interestingly the naive observers achieved a higher correlation than the expert observers. With regard to the deviations of assessments (Table 1) this result is somewhat surprising. Though the processing pipeline for the experiment with printings introduced additional noise in comparison to the on-screen experiments, it is not clear whether this circumstance already accounts for the better results with naive observers.

A further difference between the two experiments is the respective set of test patterns. While in the on-screen experiments the strength of the streaks was approximately uniformly distributed over the grade scale, the expert experiment used a high percentage of realistic prints, which contain many barely visible streaks. Ratings performed close to the threshold of perception can be a problem, as shown below.

4.3. *Patterns at Threshold of Perception*

Printings are generally noisy with regard to the luminance signal. Therefore many weak streaks are difficult to detect even for trained observers. This is illustrated by the fact that naive observers who have performed the assessment task with realistic printings found approximately three streaks per print whereas the experts found up to 25. However, almost all additional streaks found by the expert group were rated with 1 (barely visible)[1]. These distortions pose a problem. While the model can generate output lower than 1 (if the signal is weak), the observers are stuck with the 1 as the lowest grade available (0 is reserved for non-detected streaks). Even though the signal would require a lower grade, the scale does not offer it and a sort of floor effect becomes visible.

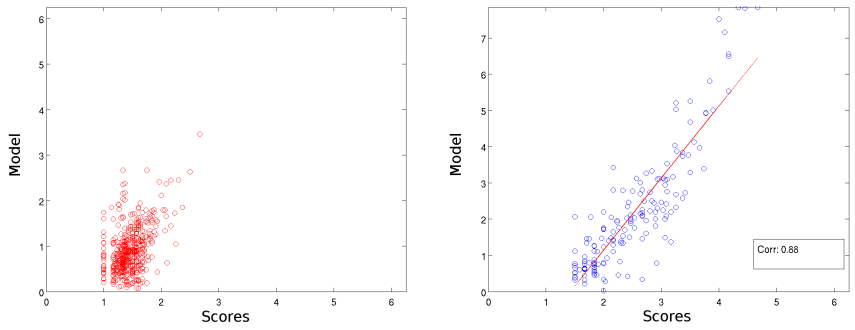
[1]Since the locations of the streaks are marked, they will typically receive a grade 1, even if critical testing without markings would presumably reveal that some participants are not able to see them.

4.3.1. *Correlation Between Assessments and Model with Varying Percentage of Near-threshold Distortions*

The correlations between model and participants' assessments for the whole set, including the afore-mentioned distortions near the threshold of perception, have already been presented in figure 5(b). We now analyse the influence of those near-threshold responses on the overall performance.

Subsets of the data with a varying percentage of near-threshold responses have been investigated with regard to their effect on the correlation of the model predictions. The criterion for assigning a particular streak to the near-threshold group was the percentage of minimum grades given to that streak. As the minimum grade was effectively given to streaks which were barely visible or not seen at all by some observers, this criterion offers a flexible way to determine the subset of near-threshold streaks. The strictest version of the criterion is to mark a streak as near-threshold if at least one minimum grade has been assigned to it by one of the participants. In subsequent plots assessments of distortions are plotted in red if they contain a minimum grade response.

Figure 6(a) shows the subset consisting of inter-individually averaged assessments, which received at least one minimum grade response. As one can see, the apparent variance of the model output is quite high, which is in correspondence to the aforementioned hypothesis that in this regime different signal levels are mapped to the minimum grade. The omission of this subset, i.e. of streaks which have received at least one minimum grade from the evaluation, leads to an increase in the correlation for the averaged data from 0.81 to 0.88, as shown in Figure 6(b). Note that although the overall dynamic range of the data is reduced, the correlation increases. This corroborates our hypothesis that the excluded subset is not well behaved.



(a) (b)

Figure 6. Correlation between model and inter-individually averaged assessments for ratings containing at least one minimum grade response (a) and for ratings not containing the minimum grade (b).

5. Conclusion

We have presented a model for the automatic prediction of the perceived strength of streak distortions in offset printing. The model is based on recent neurophysiological and psychophysical results. It can describe shape effects due to the shape of the luminance function of a streak and masking effects as caused by the configuration in its spatial surround.

Assessments on a large number of streak patterns have been collected in experiments with naive and expert observers. We have then shown that the model shows a good correlation in predicting the perceived strength of these distortions. The differences between experiments on-screen and with prints can be attributed to different participants as well as the different media used for

presentation. It is further possible that mechanisms for adaptation of the visual system might perform differently for self-luminous or illuminated media. Since the L^* values differed from luminance values only by a constant for the current data set, specific effects on the predictions from use of the ROG mechanism are not to be expected. This is supported by cross-checks between model fits from on-screen and prints experiments, which showed consistent results with the on-screen data being luminance values and the print data L^*.

The prediction of distortions close to the perceptual threshold proved to be problematic, resulting in relatively large variations of the model responses for barely visible distortions. The removal of distortions which were rated with the minimum grade increased the correlation of the model predictions substantially, although the dynamic range of the data has been reduced by this removal. If specifically a modelling of barely visible distortions is intended, a better approach would be to perform 2AFC threshold measurements instead of quality judgements, and use these for appropriate model fits in the threshold range. However, for applications which require a graded evaluation the suggested model of the human visual system seems quite useful. It thus appears well suited as part of a system for the automatic evaluation of printings.

Acknowledgements

This work was supported by NCT Bremen, Sächsisches Institut für die Druckindustrie Leipzig and Forschungsgesellschaft Druckmaschinen e.V. Frankfurt am Main. The authors would like to thank the accompanying working group of specialists (FGD Arbeitskreis Streifenmessung) for the fruitful discussions. Parts of this work has been presented at the 18. Workshop of the German Color Group in Darmstadt, Germany in 2012.

References

[1] *Handbuch zur technischen Abnahme von Bogenoffset-Rollenoffsetmaschinen*; Technical report for Bundesverband Druck und Medien, 1996.
[2] *Technische Richtlinien Abnahme von Bogenoffsetdruckmaschinen*; Technical report for Bundesverband Druck und Medien, 2005.
[3] Zetzsche, C.; Hauske, G. Multiple channel model for the prediction of subjective image quality. In: *Human Vision, Visual Processing and Digital Display*, Proc. SPIE, Vol. 1077, 1989; pp 209–216.
[4] Zetzsche, C.; Hauske, G. Principal features of human vision in the context of image quality models. In: *IEEE 3rd Int. Conf. Image Proc. and its Appl*, 1989; pp 102–106.
[5] Daly, S. The visible differences predictor: an algorithm for the assessment of image fidelity. In *Digital images and human vision*; Watson, A.B. Ed.; MIT Press: Cambridge, MA, USA, 1993; pp 179–206.
[6] Lubin, J. The use of psychophysical data and models in the analysis of display system performance. In *Digital images and human vision*; Watson, A.B. Ed.; MIT Press: Cambridge, MA, USA, 1993; pp 163–178.
[7] Teo, P.C.; Heeger, D.J. Perceptual image distortion. In: *IEEE Int. Conf. Image Processing ICIP-94*, Vol. 2, 1994; pp 982–986.
[8] Heeger, D.J.; Teo, P.C. A Model of Perceptual Image Fidelity. In: *International Conference on Image Processing, 1995*, Vol. 2, 1995; pp 343–346.
[9] Westen, S.J.P.; Lagendijk, R.L.; Biemond, J. Perceptual image quality based on a multiple channel HVS model. In: *1995 Int. Conf. on Acoustics, Speech, and Signal Processing ICASSP-95*, Vol. 4, 1995; pp 2351–2354.
[10] Taylor, C.C.; Pizlo, Z.; Allebach, J.P.; et al. Image quality assessment with a Gabor pyramid model of the human visual system. In: *Human Vision and Electronic Imaging II*, Proc. SPIE, Vol. 3016, 1997; pp 58–69.
[11] Watson, A.; Solomon, J. Model of visual contrast gain control and pattern masking. *J Opt Soc Am A Opt Image Sci Vis* **1997**, *14* (9), 2379–91.
[12] Miyahara, M. Quality assessments for visual service. *IEEE Communications Magazine* **1988**, *26* (10), 51–60.
[13] Miyahara, M.; Kotani, K.; Algazi, V. Objective picture quality scale (PQS) for image coding. *IEEE Transactions on Communications* **1998**, *46* (9), 1215–1226.
[14] Kayargadde, V.; Martens, J.B. Perceptual characterization of images degraded by blur and noise: model. *J Opt Soc Am A Opt Image Sci Vis* **1996**, *13* (6), 1178–88.
[15] Zhang, X.; Wandell, B.a. A spatial extension of CIELAB for digital color-image reproduction. *SID Journal* **1997**, *5* (1), 61.
[16] Wang, Z.; Simoncelli, E.P.; Bovik, A.C. Multiscale structural similarity for image quality assessment. In: *The Thirty-Seventh Asilomar Conference on Signals, Systems & Computers, 2004*, Vol. 2, 2003; pp 1398–1402.
[17] Wang, Z.; Bovik, A.C.; Sheikh, H.R.; et al. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing* **2004**, *13* (4), 600–12.

[18] Sheikh, H.R.; Bovik, A.C. Image information and visual quality. *IEEE Transactions on Image Processing* **2006**, *15* (2), 430–44.
[19] Larson, E.C.; Chandler, D.M. Most apparent distortion: full-reference image quality assessment and the role of strategy. *J Electron Imaging* **2010**, *19* (1), 011006–1–011006–21.
[20] Sheikh, H.R.; Sabir, M.F.; Bovik, A.C. A statistical evaluation of recent full reference image quality assessment algorithms. *IEEE Transactions on Image Processing* **2006**, *15* (11), 3440–51.
[21] de Ridder, H. Cognitive issues in image quality measurement. *J Electron Imaging* **2001**, *10* (1), 47–55.
[22] Kennedy, J.; Eberhart, R. Particle swarm optimization. In: *Proceedings of ICNN'95 - IEEE International Conference on Neural Networks*, Vol. 4, 1995; pp 1942–1948.
[23] Burt, P.; Adelson, E. The Laplacian Pyramid as a Compact Image Code. *IEEE Transactions on Communications* **1983**, *31* (4), 532–540.
[24] Hubel, D.; Wiesel, T. Receptive fields of single neurones in the cat's striate cortex. *J Physiol* **1959**, *148* (3), 574–591.
[25] Hubel, D.; Wiesel, T. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *J Physiol* **1962**, *160* (1), 106–154.
[26] Hubel, D.; Wiesel, T. Receptive fields and functional architecture of monkey striate cortex. *J Physiol* **1968**, *195* (1), 215–243.
[27] Campbell, F.W.; Cooper, G.F.; Enroth-Cugell, C. The spatial selectivity of the visual cells of the cat. *J Physiol* **1969**, *203* (1), 223–235.
[28] De Valois, R.L.; De Valois, K.K.; Ready, J.; et al. Spatial frequency tuning of macaque striate cortex cells. *Assoc. Res., Vision Ophthalmol* **1975**, *15*, 16.
[29] Gabor, D. Theory of communication. Part 1: The analysis of information. *J Electrical Engineers-Part III: Radio and Communication* **1946**, *93* (26), 429–441.
[30] Marcelja, S. Mathematical description of the responses of simple cortical cells. *J Optl Soc Am* **1980**, *70* (11), 1297–300.
[31] Daugman, J.G. Spatial visual channels in the Fourier plane. *Vis Res* **1984**, *24* (9), 891–910.
[32] Field, D.J. Relations between the statistics of natural images and the response properties of cortical cells. *J Opt Soc Am* **1987**, *4* (12), 2379–94.
[33] De Valois, R.L.; Albrecht, D.G.; Thorell, L.G. Spatial frequency selectivity of cells in macaque visual cortex. *Vis Res* **1982**, *22* (5), 545–559.
[34] De Valois, R.L.; Yund, E.W.; Hepler, N. The orientation and direction selectivity of cells in macaque visual cortex. *Vision Research* **1982**, *22* (5), 531–544.
[35] Legge, G.E.; Foley, J.M. Contrast masking in human vision. *J Opt Soc Am* **1980**, *70* (12), 1458–71.
[36] Albrecht, D.G.; Hamilton, D.B. Striate Cortex of Monkey and Cat: Function Contrast Response. *J Neurophysiol* **1982**, *48* (1), 217–237.
[37] DeAngelis, G.; Robson, J.; Ohzawa, I.; et al. Organization of suppression in receptive fields of neurons in cat visual cortex. *J Neurophysiol* **1992**, *68* (1), 144–163.
[38] Geisler, W.S.; Albrecht, D.G. Cortical neurons: isolation of contrast gain control. *Vis Res* **1992**, *32* (8), 1409–1410.
[39] Heeger, D.J. Normalization of cell responses in cat striate cortex. *Vis Neurosci* **1992**, *9* (2), 181–198.
[40] *Recommendation ITU-R BT.500-11, Methodology for the subjective assessment of the quality of television pictures*; Technical report for International Telecommunication Union, Geneva, 2002.
[41] *ISO 3664:2009-04 Graphic technology and photography. Viewing conditions*; Technical report for ISO, 2009.
[42] *Versuchsdurchführung bei der Bewertung von Druckbogen*; Technical report for FOGRA, 2010.

Statistical Invariants of Spatial Form: From Local AND to Numerosity

Christoph ZETZSCHE [1], Konrad GADZICKI and Tobias KLUTH
Cognitive Neuroinformatics, University of Bremen, Germany

Abstract Theories of the processing and representation of spatial form have to take into account recent results on the importance of holistic properties. Numerous experiments showed the importance of "set properties", "ensemble representations" and "summary statistics", ranging from the "gist of a scene" to something like "numerosity". These results are sometimes difficult to interpret, since we do not exactly know how and on which level they can be computed by the neural machinery of the cortex. According to the standard model of a local-to-global neural hierarchy with a gradual increase of scale and complexity, the ensemble properties have to be regarded as high-level features. But empirical results indicate that many of them are primary perceptual properties and may thus be attributed to earlier processing stages. Here we investigate the prerequisites and the neurobiological plausibility for the computation of ensemble properties. We show that the cortex can easily compute common statistical functions, like a probability distribution function or an autocorrelation function, and that it can also compute abstract invariants, like the number of items in a set. These computations can be performed on fairly early levels and require only two well-accepted properties of cortical neurons, linear summation of afferent inputs and variants of nonlinear cortical gain control.

Keywords. shape invariants, peripheral vision, ensemble statistics, numerosity

Introduction

Recent evidence shows that our representation of the world is essentially determined by holistic properties [1,2,3,4,5,6]. These properties are described as "set properties", "ensemble properties", or they are characterized as "summary statistics". They reach from the average orientation of elements in a display [1] over the "gist of a scene"[7,8], to the "numerosity" of objects in a scene [9]. For many of these properties we do not exactly know by which kind of neural mechanisms and on which level of the cortex they are computed. According to the standard view of the cortical representation of shape, these properties have to be considered as high-level features because the cortex is organized in form of a local-to-global processing hierarchy in which features with increasing order of abstraction are computed in a progression of levels [10]. At the bottom, simple and locally restricted geometrical features are computed, whereas global and complex properties are represented at the top levels of the hierarchy. Across levels, invariance is system-

[1]Corresponding Author: Christoph Zetzsche, Cognitive Neuroinformatics, FB3, University of Bremen, P.O. Box 330 440, 28334 Bremen, Germany; E-mail: zetzsche@informatik.uni-bremen.de
Research supported by **DFG** (SFB/TR 8 Spatial Cognition, A5-[ActionSpace])

atically increased such that the final stages are independent of translations, rotations, size changes, and other transformations of the input. However convincing this view seems on first sight, it creates some conceptual difficulties.

The major difficulty concerns the question of what exactly is a low-level and a high-level property. Gestalt theorists already claimed that features considered high-level according to a structuralistic view are primary and basic in terms of perception. Further doubts have been raised by global precedence effects [11]. Similar problems arise with the recently discovered ensemble properties. The gist of a scene, a high-level feature according to the classical view, can be recognized in 150 msec [7,12,13,14] and can be modeled using low-level visual features [8]. In addition, categories can be shown to be faster processed than basic objects, contrary to the established view of the latter as entry-level representations [15]. A summary statistics approach, also based on low-level visual features, can explain the holistic processing properties in the periphery of the visual field [4,16,17]. What is additionally required in these models are statistical measures, like probability distributions and autocorrelation functions, from which it is not known how and on which level of the cortical hierarchy they can be realized.

One of the most abstract ensemble properties seems to be the number of elements in a spatial configuration. However, the ability to recognize this number is not restricted to humans with mature cognitive abilities but has also been found in infants and animals [9,18], recently even in invertebrates [19]. Neural reactions to numerosity are fast (100 msecs in macaques [20]). And finally there is evidence for a "direct visual sense for number" since number seems to be a primary visual property like color, orientation or motion, to which the visual system can be adapted by prolonged viewing [21].

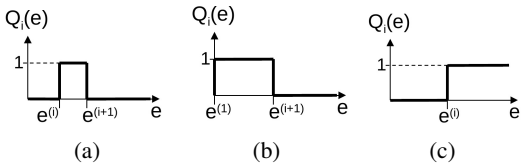
The above observations on ensemble properties raise a number of questions, from which the following are addressed in this paper: Sect. 1: Can the cortex compute a probability distribution? Sect. 2: And also an autocorrelation function? By which kind of neural hardware can this be achieved? Sect.3: Can the shape of individual objects also be characterized by such mechanisms? Sect. 4: What is necessary to compute such an abstract property like the number of elements in a spatial configuration? Can this be achieved in early sensory stages?

1. Neural Computation of a Probability Distribution

Formally, the probability density function $p_e(e)$ of a random variable $\mathbf{e}$ is defined via the cumulative distribution function: $p_e(e) \triangleq \frac{dP_e(e)}{de}$ with $P_e(e) = Pr[\mathbf{e} \leq e]$. Their empirical counterparts, the histogram and the cumulative histogram, are defined by use of *indicator functions*. For this we divide the real line into m bins $(e^{(i)}, e^{(i+1)}]$ with bin size $\Delta e = e^{(i+1)} - e^{(i)}$. For each bin i, an indicator function is defined as

$$Q_i(e) = 1_i(e) = \begin{cases} 1, \text{ if } e^{(i)} < e \leq e^{(i+1)} \\ 0, \text{ else} \end{cases} \tag{1}$$

An illustration of such a function is shown in Fig. 1a. From N samples e_k of the random variable $\mathbf{e}$ we then obtain the histogram as $h(i) = \frac{1}{N} \sum_{k=1}^{N} Q_i(e_k)$. The cumulative histogram $H_e(e)$ can be computed by changing the bins to $(e^{(1)}, e^{(i+1)}]$ (cf. Fig. 1b), and by performing the same summation as for the normal histogram. The reverse cumulative



(a) (b) (c)

Figure 1. Indicator functions. Basic types are: (a) indicator function for computation of a classical histogram. (b) indicator function for a cumulative histogram. (c) indicator function for a reverse cumulative histogram.

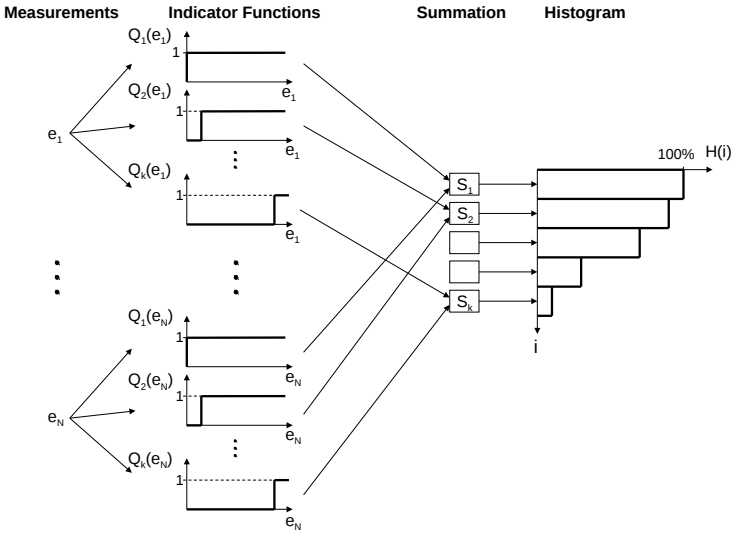
histogram $\bar{H}(i)$ is simply the reversed version of the cumulative histogram. The corresponding bins are $\Delta e_i = (e^{(i)}, e^{(m+1)}]$ and the indicator functions are defined as (Fig. 1c)

$$Q_i(e) = 1_i(e) = \begin{cases} 1, \text{ if } e \geq e^{(i)} \\ 0, \text{ else} \end{cases} \tag{2}$$

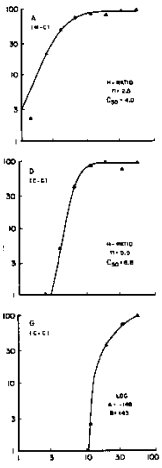
The corresponding system is shown in Fig. 2.

The three types of histograms have identical information content since they are related to each other as

$$h(i) = H((i+1)) - H(i) = \bar{H}(i) - \bar{H}(i+1) \quad \text{and} \quad H(i) = 1 - \bar{H}(i) = \sum_{j=1}^{i} h(j). \tag{3}$$



(a)



Albrecht and Hamilton (1982)

(b)

Figure 2. Computation of the reverse cumulative histogram. (a) shows the set of input variables e_1 to e_n over which the histogram should be computed. Each of these variables is input to a set of *indicator functions* $Q_i(e_k)$. For each bin of the histogram there is a summation unit S_i which sums over all indicator function outputs with index i, i.e. over all $Q_i(e_k)$.
(b) The response functions of three neurons in the visual cortex [22]. They show a remarkable similarity to the indicator functions for the reverse cumulative histogram. First, they come with different sensitivities. Second, they exhibit an independence on the input strength: once the threshold and the following transition range is exceeded the output remains constant and does no longer increase when the input level is increased.

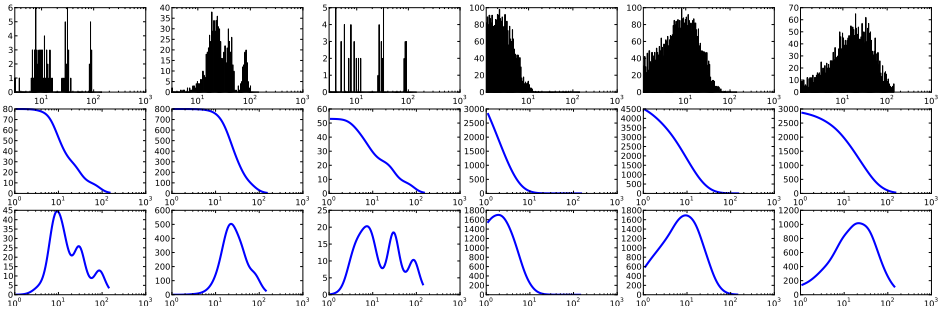


Figure 3. Neurobiological computation of a reverse cumulative histogram. The upper row shows several examples of input probability distributions. The second row shows the corresponding reverse cumulative histograms computed by a dense set of simulated neurons. The third row shows the estimated probability distributions as derived from the neural representation by use of Eq. (3).

How does all this relate to visual cortex? Has the architecture shown in Fig. 2a any neurobiological plausibility? The final summation stage is no problem since the most basic capability of neurons is computation of a linear sum of their inputs. But how about the indicator functions? They have two special properties: First, the indicator functions come with different *sensitivities*. An individual function does only generate a non-zero output if the input e exceeds a certain level, a kind of threshold, which determines the sensitivity of the element $e^{(i)}$ in Eq. (2) and Fig. 1c. To cover the complete range of values, different functions with different sensitivities are needed (Fig. 2a). Second, the indicator functions exhibit a certain *independence of the input level*. Once the input is clearly larger than the threshold, the output remains constant (Fig. 1c).

Do we know of neurons which have such properties, a range of different sensitivities, and a certain independence of the input strength? Indeed, cortical gain control (or normalization), as first described in early visual cortex (e.g. [22]) but now believed to exist throughout the brain [23], yields exactly these properties. Gain-controlled neurons (Fig. 2b) exhibit a remarkable similarity to the indicator functions used to compute the reverse cumulative histogram, since they (i) come with different sensitivities, and (ii) provide an independence of the input strength in certain response ranges.

The computation of a reverse cumulative histogram thus is well in reach of the cortex. We only have to modify the architecture of Fig. 2a by the smoother response functions of cortical neurons. The information about a probability distribution available to the visual cortex is illustrated in Fig. 3. The reconstructed distributions, as estimated from the neural reverse cumulative histograms, are a kind of Parzen-windowed (lowpass-filtered) versions of the original distributions.

2. Neural Implementation of Auto- and Cross-Correlation Functions

A key feature of the recent statistical summary approach to peripheral vision [4,6,24,16] is the usage of auto- and cross-correlation functions. These functions are defined as

$$h(i) = \frac{1}{N} \sum_{k=-N/2+1}^{N/2} e(k) \circ g(i+k), \tag{4}$$

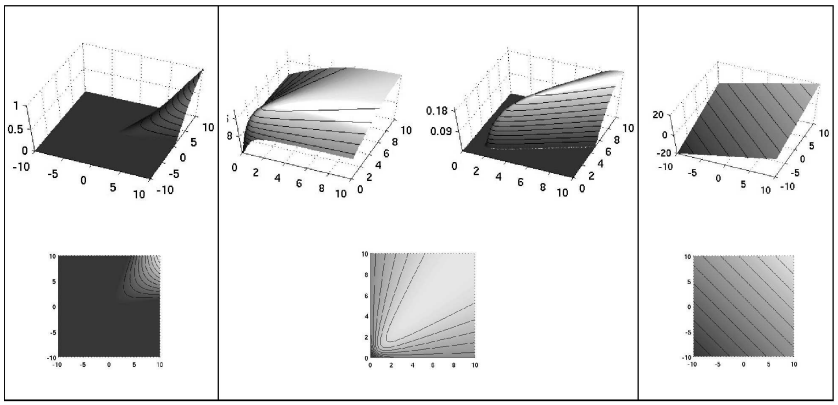


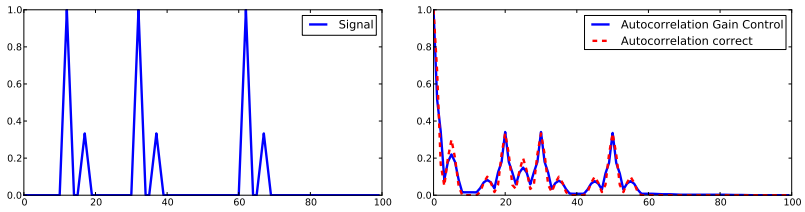
Figure 4. Different types of AND-like functions. Each function is of the type $g_k = g(s_i, s_j)$, i.e. assigns an output value to each combination of the two input values. The upper row shows the functions as surface plots, the lower row as iso-response curves. Left: Mathematical multiplication of two inputs. Center: AND-like combinations that can be obtained by use of cortical gain control (normalization). The upper left figure shows the classical gain control without additional threshold. The upper right figure shows the same mechanism with an additional threshold. This results in a full-fledged AND with a definite zero response in case that only one of the two inputs is active. Right: The linear sum of the two input values for comparison purposes.

where autocorrelation results if $e(k) = g(k)$ and where $\circ$ indicates multiplication. With respect to their neural computation, the outer summation is no problem, but the crucial function is the *nonlinear multiplicative interaction between two variables*. A neural implementation could make use of the Babylonian trick $ab = \frac{1}{4}[(a+b)^2 - (a-b)^2]$ [25,26,27], but this requires two or more neurons for the computation and thus far there is neither evidence for such a systematic pairing of neurons nor for actual multiplicative interactions in the visual cortex. However, exact multiplication is not the key factor: a reasonable statistical measure merely requires provision of a matching function such that $e(k)$ and $g(i+k)$ generate a large contribution to the autocorrelation function if they are similar, and a small contribution if they are dissimilar. For this, it is sufficient to provide a neural operation which is AND-like [27,28]. Surprisingly, such an AND-like operation can be achieved by the very same neural hardware as used before, the cortical gain control mechanism, as shown in [28]. Cortical gain control [22,29] applied to two different features $s_i(x,y)$ and $s_j(x,y)$ can be written as

$$g_k(x,y) = g(s_i(x,y), s_j(x,y)) := \max\left(0, \frac{s_i + s_j}{(\sqrt{s_i^2 + s_j^2} + \varepsilon)\sqrt{2}} - \Theta\right) \tag{5}$$

where $k = k(i,j)$, ε is a constant which controls the steepness of the response and Θ is a threshold. The resulting nonlinear combination is comparable with an AND-like operation of two features and causes a substantial nonlinear increase of the neural selectivity, as illustrated in Fig. 4.

Of course there will be differences between a formal autocorrelation function and the neurobiological version, but the essential feature, the signaling of good matches in dependence of the relative shifts will be preserved (Fig. 5).



(a) (b)

Figure 5. Mathematical and neurobiological autocorrelation functions. (a) shows a test input and (b) the corresponding mathematical (red dotted) and neurobiological (blue) autocorrelation function.

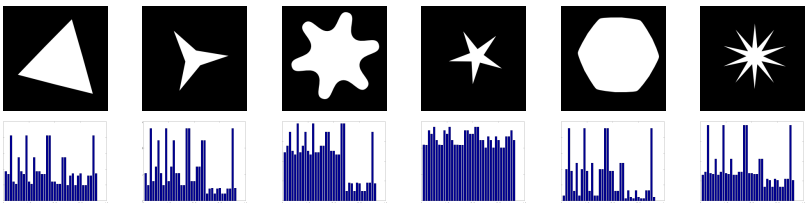


Figure 6. Different shapes and the corresponding integral features. We used parameter combinations of six different orientations $\theta_i = (i-1)\pi/6,\ i = 1,\ldots,6$, and four different scales $r_i = 2^{-i},\ i = 1,\ldots,4$. The radial half-bandwidth was set to $f_{r,h} = \frac{1}{3}r$ and the angular half-bandwidth was constant with $f_{\theta,h} = \pi/12$. Each parameter combination creates pairs of variables for each x,y-position which are AND-combined by the gain control mechanism described in Eq. (5) as $g_k(x,y) = g(s_i(x,y), s_j(x,y))$.

3. Figural Properties from Integrals

We extracted different features $s_{r,\theta}$ from the image luminance function $l = l(x,y)$ by applying a Gabor-like filter operation $s_{r,\theta}(x,y) = (l * \mathscr{F}^{-1}(H_{r,\theta}))(x,y)$ where $\mathscr{F}^{-1}$ denotes the inverse Fourier transformation and the filter kernel $H_{r,\theta}$ is defined in the spectral space. We distinguish two cases (even and odd) which can be seen in the following definition in polar coordinates:

$$H^{even}_{r,\theta}(f_r, f_\theta) := \begin{cases} \cos^2\left(\frac{\pi}{2}\frac{f_r - r}{2f_{r,h}}\right)\cos^2\left(\frac{\pi}{2}\frac{f_\theta - \theta}{2f_{\theta,h}}\right) & ,(f_r,f_\theta) \in \Omega_{r,\theta} \\ 0 & ,\text{else,} \end{cases}$$

with $\Omega_{r,\theta} := \{(f_r,f_\theta) | f_r \in [r-2f_{r,h}, r+2f_{r,h}] \wedge f_\theta \in [\theta - 2f_{\theta,h}, \theta + 2f_{\theta,h}] \cap [\theta + \pi - 2f_{\theta,h}, \theta + \pi + 2f_{\theta,h}]\}$, where $f_{r,h}$ denotes the half-bandwidth in radial direction and $f_{\theta,h}$ denotes the half-bandwidth in angular direction. $H^{odd}_{r,\theta}$ is defined as the Hilbert transformed even symmetric filter kernel.

Various AND combinations of these oriented features (see caption Fig. 6) are obtained by the gain-control mechanism described in Eq. (5). The integration over the whole domain results in *global* features $F_k := \int_{\mathbb{R}^2} g_k(x,y)\ d(x,y)$ which capture basic shape properties (Fig. 6).

4. Numerosity and Topology

One of the most fundamental and abstract ensemble properties is the number of elements of a set. Recent evidence (see Introduction) raised the question at which cortical level

the underlying computations are performed. In this processing, a high degree of invariance has to be achieved, since numerosity can be recognized largely independent of other properties like size, shape and positioning of elements. Models which address this question in a neurobiologically plausible fashion, starting from individual pixels or neural receptors instead of an abstract type of input, are rare. To our knowledge, the first approach in this direction has been made in [30]. A widely known model [31] has a shape-invariant mapping to number which is based on linear DOG filters of different sizes, which substantially limits the invariance properties. A more recent model is based on unsupervised learning but has only employed moderate shape variations [32]. In [30] we suggested that the necessary invariance properties may be obtained by use of a theorem which connects local measurements of the differential geometry of the image surface with global topological properties [30,33]. In the following we will build upon this concept.

The key factor of our approach is a relation between surface properties and a topological invariant as described by the famous Gauss-Bonnet theorem. In order to apply this to the image luminance function $l = l(x,y)$ we interpret this function as a surface $S := \{(x,y,z) \in \mathbb{R}^3 | (x,y) \in \Omega, z = l(x,y)\}$ in three-dimensional real space. We then apply the formula for the Gaussian curvature

$$K(x,y) = \frac{l_{xx}(x,y)l_{yy}(x,y) - l_{xy}(x,y)^2}{(1 + l_x(x,y)^2 + l_y(x,y)^2)^2}, \tag{6}$$

where subscript denotes the differentiation in the respective direction (e.g. $l_{xy} = \frac{\partial^2 l}{\partial x \partial y}$). The numerator of (6) can also be written as $D = \lambda_1 \lambda_2$ where $\lambda_{1,2}$ are the eigenvalues of the Hessian matrix of the luminance function $l(x,y)$ which represent the partial second derivatives in the principal directions. The values and signs of the eigenvalues give us the information about the shape of the luminance surface S in each point, whether it is elliptic, hyperbolic, parabolic, or planar. Since Gaussian curvature results from the multiplication of the second derivatives $\lambda_{1,2}$ it is zero for the latter two cases. It has been shown that this measure can be generalized in various ways, in particular towards the use of neurophysiologically realistic Gabor-like filters instead of the derivatives [27,30]. The crucial point, however, is the need for *AND combinations of oriented features* [27,30] which can be obtained as before by the neural mechanism of cortical gain control [28].

The following corollary from the Gauss-Bonnet theorem is the basis for the invariance properties in the context of numerosity.

Corollary 4.1 *Let $S \subset \mathbb{R}^3$ be a closed two-dimensional Riemannian manifold. Then*

$$\int_S K \, dA = 4\pi(1-g) \tag{7}$$

where K is the Gaussian curvature and g is the genus of the surface S.

We consider the special case where the luminance function consists of multiple objects (polyhedra with orthogonal corners) with constant luminance level. We compare the surface of this luminance function to the surface of a cuboid with holes that are shaped like the polyhedra. The trick is that the latter surface has a genus which is determined by the number of holes in the cuboid and which can be determined by the integration of the local curvature according to Eq. (7). If we can find the corresponding contributions of

the integral on the image surface, we can use this integral to count the number of objects. We assume the corners to be locally sufficiently smooth such that the surfaces are Riemannian manifolds. The Gaussian curvature K then is zero almost everywhere except on the corners. We hence have to consider only the contributions of the corners. It turns out that these contributions can be computed from the elliptic regions only if we use different signs for upwards and downwards oriented elliptic regions. We thus introduce the following operator which distinguishes the different types of ellipticity in the luminance function. Let $\lambda_1 \geq \lambda_2$, then the operator $N(x,y) := |\min(0,\lambda_1(x,y))| - |\max(0,\lambda_2(x,y))|$ is always zero if the surface is hyperbolic and has a positive sign for positive ellipticity and a negative one for negative ellipticity. We thus can calculate the numerosity feature which has the ability of counting objects in an image by counting the holes in an imaginary cuboid as follows:

$$F = \int_\Omega \frac{N(x,y)}{(1+l_x(x,y)^2+l_y(x,y)^2)^{\frac{3}{2}}}\, d(x,y). \tag{8}$$

The crucial feature of this measure are contributions of fixed size and with appropriate signs from the corners. The denominator can thus be replaced by a neural gain control mechanism and an appropriate renormalization. For the implementation here we use a shortcut which gives us straight access to the eigenvalues. The numerator $D(x,y)$ of (6) can be rewritten as

$$D(x,y) = l_{xx}l_{yy} - \frac{1}{4}(l_{uu}-l_{vv})^2 = \frac{1}{4}[(l_{xx}+l_{yy})^2 - \underbrace{((l_{xx}-l_{yy})^2+(l_{uu}-l_{vv})^2)}_{=:\varepsilon^2}] = \frac{1}{4}(\Delta l^2 - \varepsilon^2) \tag{9}$$

with $u := x\cos(\pi/4) + y\sin(\pi/4)$ and $v := -x\sin(\pi/4) + y\cos(\pi/4)$. The eigenvalues then are $\lambda_{1,2} = \frac{1}{2}(\Delta l \pm |\varepsilon|)$ and we can directly use them to compute $N(x,y)$. Application of this computation to a number of test images is shown in Fig. 7.

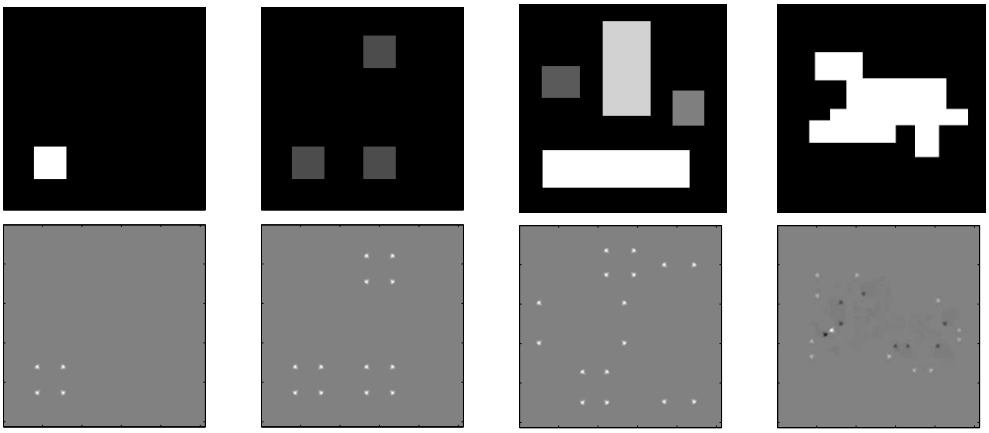


Figure 7. Based on a close relation to topological invariants the spatial integration of local curvature features can yield highly invariant numerosity estimates. The numerical values in the last row are the normalized integrals of the filter outputs (middle row).

5. Conclusion

Recent evidence shows that ensemble properties play an important role in perception and cognition. In this paper, we have investigated by which neural operations and on which processing level statistical ensemble properties can be computed by the cortex. Computation of a probability distribution requires *indicator functions* with different sensitivities, and our reinterpretation of cortical gain control suggests that this could be a basic function of this neural mechanism. The second potential of cortical gain control is the computation of *AND-like feature combinations*. Together with the linear summation capabilities of neurons this enables the computation of powerful invariants and summary features. We have repeatedly argued that AND-like feature combinations are essential for our understanding of the visual system [27,30,34,35,36,28]. The increased selectivity of nonlinear AND operators, as compared to their linear counterparts, is a prerequisite for the usefulness of integrals over the respective responses [30,28]. We have shown that such integrals of AND features are relevant for the understanding of texture perception [37], of numerosity estimation [30], and of invariance in general [28]. Recently, integrals over AND-like feature combinations in form of auto- and cross-correlation functions have been suggested for the understanding of peripheral vision [4,16,17].

A somewhat surprising point is that linear summation and cortical gain control, two widely accepted properties of cortical neurons, are the only requirements for the computation of ensemble properties. These functions are already available at early stages of the cortex, but also in other cortical areas [23]. The computation of ensemble properties may thus be an ubiquitous phenomenon in the cortex.

Acknowledgement

This work was supported by DFG, SFB/TR8 Spatial Cognition, project A5-[ActionSpace].

References

[1] S. C. Dakin and R. J. Watt. The computation of orientation statistics from visual texture. *Vision Res*, 37(22):3181–3192, 1997.

[2] D. Ariely. Seeing Sets: Representation by Statistical Properties. *Psychol Sci*, 12(2):157–162, 2001.

[3] Lin Chen. The topological approach to perceptual organization. *Visual Cognition*, 12(4):553–637, 2005.

[4] B. Balas, L. Nakano, and R. Rosenholtz. A summary-statistic representation in peripheral vision explains visual crowding. *J Vis*, 9(12):13.1–18, 2009.

[5] G. A. Alvarez. Representing multiple objects as an ensemble enhances visual cognition. *Trends Cog Sci*, 15(3):122–31, 2011.

[6] R. Rosenholtz, J. Huang, and K. Ehinger. Rethinking the role of top-down attention in vision: effects attributable to a lossy representation in peripheral vision. *Front Psychol*, 3:13, 2012.

[7] S. Thorpe, D. Fize, and C. Marlot. Speed of processing in the human visual system. *Nature*, 381(6582):520–522, 1996.

[8] A. Oliva and A. Torralba. Modeling the Shape of the Scene: A Holistic Representation of the Spatial Envelope. *International Journal of Computer Vision*, 42(3):145–175, 2001.

[9] E. M. Brannon. The representation of numerical magnitude. *Curr Opin Neurobiol*, 16(2):222–9, 2006.

[10] J. Hegde and D.J. Felleman. Reappraising the Functional Implications of the Primate Visual Anatomical Hierarchy. *The Neuroscientist*, 13(5):416–421, 2007.

[11] D. Navon. Forest before trees: The precedence of global features in visual perception. *Cognitive Psychology*, 9(3):353–383, 1977.

[12] M. R. Greene and A. Oliva. The Briefest of Glances. *Psychol Sci*, 20(4):464–472, 2009.

[13] J. Hegdé. Time course of visual perception: coarse-to-fine processing and beyond. *Prog Neurobiol*, 84(4):405–39, 2008.
[14] M. Fabre-Thorpe. The characteristics and limits of rapid visual categorization. *Front Psychol*, 2:243, 2011.
[15] M. J-M Macé, O. R. Joubert, J-L. Nespoulous, and M. Fabre-Thorpe. The time-course of visual categorizations: you spot the animal faster than the bird. *PloS one*, 4(6):e5927, 2009.
[16] J. Freeman and E. P. Simoncelli. Metamers of the ventral stream. *Nature neuroscience*, 14(9):1195–1201, 2011.
[17] H. Strasburger, I. Rentschler, and M. Jüttner. Peripheral vision and pattern recognition: a review. *J Vis*, 11(5):13, 2011.
[18] A. Nieder, D. J. Freedman, and E. K. Miller. Representation of the quantity of visual items in the primate prefrontal cortex. *Science*, 297(5587):1708–11, 2002.
[19] H. J. Gross, M. Pahl, A. Si, H. Zhu, J. Tautz, and S. Zhang. Number-based visual generalisation in the honeybee. *PloS one*, 4(1):e4263, 2009.
[20] J. D. Roitman, E. M. Brannon, and M. L. Platt. Monotonic coding of numerosity in macaque lateral intraparietal area. *PLoS biology*, 5(8):e208, 2007.
[21] J. Ross and D. C. Burr. Vision senses number directly. *Journal of vision*, 10(2):10.1–8, 2010.
[22] D. G. Albrecht and D. B. Hamilton. Striate cortex of monkey and cat: contrast response function. *J Neurophysiol*, 48(1):217–237, Jul 1982.
[23] M. Carandini and D. J. Heeger. Normalization as a canonical neural computation. *Nature Reviews Neurosci*, 13:51–62, Jul 2012.
[24] R. Rosenholtz, J. Huang, A. Raj, B. J. Balas, and L. Ilie. A summary statistic representation in peripheral vision explains visual search. *J Vis*, 12(4):1–17, 2012.
[25] H.L. Resnikoff and R.O. Wells. *Mathematics in Civilization*. Popular Science Series. Dover, 1984.
[26] E. H. Adelson and J. R. Bergen. Spatiotemporal energy models for the perception of motion. *J. Opt. Soc. Am. A*, 2(2):284–99, 1985.
[27] C. Zetzsche and E. Barth. Fundamental limits of linear filters in the visual processing of two-dimensional signals. *Vision Res*, 30(7):1111–1117, 1990.
[28] C. Zetzsche and U. Nuding. Nonlinear and higher-order approaches to the encoding of natural scenes. *Network*, 16(2–3):191–221, 2005.
[29] D.J. Heeger. Normalization of cell responses in cat striate cortex. *Visual Neurosci*, 9(2):181–198, 1992.
[30] C. Zetzsche and E. Barth. Image surface predicates and the neural encoding of two-dimensional signal variations. In B. E. Rogowitz and Jan P. A., editors, *Proc SPIE*, volume 1249, pages 160–177, 1990.
[31] S. Dehaene and J. P. Changeux. Development of elementary numerical abilities: a neuronal model. *J. Cogn. Neurosci.*, 5(4):390–407, 1993.
[32] I. Stoianov and M. Zorzi. Emergence of a 'visual number sense' in hierarchical generative models. *Nat Neurosci*, 15(2):194–6, 2012.
[33] M. Ferraro E. Barth and C. Zetzsche. Global topological properties of images derived from local curvature features. In L. P. Cordella C. Arcelli and G. Sanniti di Baja, editors, *Visual Form 2001. Lecture Notes in Computer Science*, pages 285–294, 2001.
[34] C. Zetzsche, E. Barth, and B. Wegmann. The importance of intrinsically two-dimensional image features in biological vision and picture coding. In A. B. Watson, editor, *Digital images and human vision*, pages 109–138. MIT Press, Cambridge, MA, 1993.
[35] G. Krieger and C. Zetzsche. Nonlinear image operators for the evaluation of local intrinsic dimensionality. *IEEE Transactions Image Processing*, 5:1026–1042, 1996.
[36] C. Zetzsche and G. Krieger. Nonlinear mechanisms and higher-order statistics in biological vision and electronic image processing: review and perspectives. *J Electronic Imaging*, 10(1):56–99, 2001.
[37] E. Barth, C. Zetzsche, and I. Rentschler. Intrinsic 2D features as textons. *J. Opt. Soc. Am. A*, 15(7):1723—1732, 1998.

Neural Computation of Statistical Image Properties in Peripheral Vision

Christoph Zetzsche[1], Ruth Rosenholtz[2], Noshaba Cheema[1], Konrad Gadzicki[1], Lex Fridman[2], and Kerstin Schill[1]

[1]Cognitive Neuroinformatics, University of Bremen

[2]Dept. of Brain and Cognitive Sciences CSAIL, Massachusetts Institute of Technology

In the peripheral field of view our visual system provides a much lower image quality than in the central region. This has often been attributed to a mere loss of spatial acuity, but recent investigations suggest that the system uses a more refined strategy. For lowering its data load it computes a statistical summary representation based on low-level image features. In a recent modeling approach the summary statistics refer to classical statistical measures, like auto- and cross- correlations, operating on wavelet-like filter outputs.

Here we investigate how such a statistical representation can be obtained in a neurobiologically plausible fashion. For this, we consider both the elementary neural operations and the architectural properties. For example, the neurobiological plausibility of multiplications which are an essential component of classic statistical operations, is unclear and often critically debated. Also, it remains to be determined how the characterization of a statistical distribution, classically achived by moments or histograms, can be achieved by neurobiological hardware. Furthermore, it is unclear which specific visual features are actually best suited for an efficient statistical summary representation.

We address these problems by considering how basic neural nonlinearities, like cortical gain control, can contribute to the computation of statistical properties, how basic neural selectivities, e.g. for intrinsically two-dimensional signals are related to statistical features, and how results from the promising domain of deep learning can help to understand the role of architectural properties in a statistical representation. We show that the visual cortex can provide a reliable statistical characterization of the visual environment, and we discuss which role this representation can play for different visual tasks, e.g., for object recognition, gist estimation, and localization.

Keywords: peripheral vision, visual crowding, neuro-biologically motivated statistics, deep networks, image compression, localization

Multimodal Convolutional Neural Networks for Human Activity Recognition*

Konrad Gadzicki, Razieh Khamsehashari and Christoph Zetzsche[1]

Abstract— **We investigated multimodal fusion with convolutional neural networks (CNN) for activity recognition. Out of the number of possible modalities, we have focused on RGB video, optical flow video and skeleton data. Our work here makes use of the "NTU RGB+D" dataset, as preparation for a later application to a large-scale database (project "EASE"). By combining the output layers of state of the art CNNs, we have implemented a late fusion approach. In addition to the fused CNN architecture, we have also investigated the performance of the individual CNNs in unimodal mode, and could improve the performance of skeleton classification on this dataset with regard to the literature.**

I. INTRODUCTION

Human activity recognition has become a prominent research area due to its potential application in video surveillance, human computer interfaces, ambient assisted living, human-robot-interaction, autonomous driving etc. Within the collaborative research center "EASE" (http://www.ease-crc.org), human activity recognition serves as a first step towards generating instructions for robots on how to perform actions. Activity recognition can be based on a variety of features and part of the work in "EASE" is centered around the acquisition of large volumes of high-dimensional biosignal data from humans performing everyday activities [32], [33], [31].

Along with the general development in artificial intelligence, deep learning techniques have also gained exceptional achievements in human activity recognition. Particularly, deep learning methods based on the Recurrent Neural Networks (RNN) and Convolutional Neural Networks (CNN) architectures have shown great performance in classification tasks by learning discriminant features from large amounts of data.

In the case of human activity recognition we are dealing with temporal data, e.g. video and audio streams or biosensor readings over time. In order to apply CNNs the usual 2D convolution used for images has to be expanded to three dimensions, making it a spatio-temporal convolution for videos.

While remarkable results have already been achieved by unimodal processing of RGB, skeleton, depth, audio, biosignals, etc., effective deep networks for fusion of multimodal data represent a promising research direction. A system utilizing different sources of data simultaneously has the potential to substantially improve the performance of current unimodal approaches. Our approach hence aims at the processing of such multimodal data with spatio-temporal convolutional neural networks.

*This work was supported by DFG (German Research Foundation) as part of Collaborative Research Center "EASE - Everyday activities Science and Engineering" (http://www.ease-crc.org)

[1]Authors are with the Faculty of Mathematics and Computer Science, University of Bremen, Bremen, Germany konrad.gadzicki@uni-bremen.de rkhamseh@uni-bremen.de zetzsche@uni-bremen.de

II. RELATED WORK

A. Data for Activity Recognition

Activity recognition can be performed on a wide variety of features and a large number of datasets have been provided (for review see [26], [54]). Recent approaches in acitivity recognition often work on RGB-D data. These consist of RGB video and accompanying depth maps and provide two useful modalities for human activity recognition. Skeleton data are a third modality of interest, and can be extracted from RGB-D data as well. The RGB channel provides information with regard to shape, color and texture from which rich features can be extracted. This includes, for example, the computation of optical flow. The depth channel on the other hand is rather invariant to changes in color, texture and illumination, thereby providing a certain robustness with regard to the perception of a scene. By providing 3D structural information it helps with segmentation, determination of shapes (e.g., a human silhouette) and in the computation of skeleton data. Such skeleton data are of particular interest for human activity recognition as they carry high level information in the form of abstracted 3D joint positions.

The acquisition of large volumes of high-dimensional biosignal data from humans performing everyday activities is an important goal of the collaborative research center "EASE" [32], [33], [31]. The EASE-Table-Setting-Dataset (EASE-TSD) is currently collected and is intended to benefit cognitive humanoid household robots. The final dataset is planned to consist of synchronously recorded biosignals from about 100 participants performing everyday activities while describing their task applying think-aloud protocols. Biosignals encompass multimodal multisensor streams of near and far speech and audio, video, marker-based and motion tracking, eyetracking, as well as EEG and EMG of humans performing everyday activities.

The EASE datatset has yet to be extended to larger size. An interesting large-scale dataset that is currently available is NTU RGB+D [41]. This dataset covers both a large number of subjects and of activity classes. It will hence be used for the initial phase of our investigations which we report in this paper.

B. Activity Recognition Models

The specific properties of the various modalities have led to different processing strategies. For instance, the RGB channel can be processed with a spatio-temporal CNN [48], [47], [22], [17], either on its own or together with derived optical flow [45], [5] through a two-stream CNN or as a multistage CNN [44], [57]. Another approach is to use RNNs for processing of RGB data [42], [16], [4], [8], [35]. The depth channel can be similarly processed with CNN [50], [38] or with a combination of CNN and RNN [43]. With regard to skeleton data, processing with CNN can be enabled by interpreting joint positions as image data [19], [10], [13], [26] or by using a Lie group representation [15]. There also exist RNN-based approaches [9], [11], [49], a Deep Boltzmann Machine (DBM) approach [39] and a Hidden Markov Model (HMM) with a deep network as a state probability predictor [55].

Apart from RGB-D data, bio-signals are significant and useful data in human activity recognition applications. For instance, [40] proposes two architectures consisting of a Deep Belief Network and a CNN, that recognize human activities in real time using multiple EEG sensors fusion in an unconstrained environment and selects a smaller sensor suite (similar performance) for a lean data collection system. For sEMG, deep learning approaches have been proposed as well [3], [12]. [2] presents a new transfer learning scheme employing a CNN to leverage inter-user data within the context of sEMG-based gesture recognition.

While these approaches work well for a specific modality, they do not necessarily work well for all modalities. Different modalities have their own particular properties, and the way of combining them with deep learning approaches is challenging.

III. MULTIMODAL FUSION

A. Multimodal Fusion Strategies

The general idea behind multimodal approaches to machine learning is to use several data sources as an input in order to increase the performance of the system in terms of robustness or recognition performance. Different data sources might help to remove ambiguity or to improve the data quality in case of noise. Multimodal machine learning has several challenges [34], one of which is the fusion of different modalities.

Existing approaches can be divided into early, late and hybrid fusion [7], based on where in the processing pipeline the fusion takes place. Early fusion [46] merges data sources right at the start. Usually a features extraction process, which operates on the unimodal data, takes place before merging, but it is also possible to fuse raw data. The features are then fused into a single representation, in the most simple case by concatenation. In this case we have to deal with challenges with regard to how well the individual sources can be organized so that a unified representation suitable for further processing is achieved. Early fusion has a great potential for increasing the overall performance by exploitation of cross-correlations between individual data sources.

Late fusion [46], on the other hand, merges data only after full unimodal processing. The unimodal processing part is done by individual models which opens the opportunity of using well established, sophisticated approaches for particular modalities. Fusion is performed after unimodal classification results have been generated, by merging them with strategies like averaging, majority voting etc. The major drawback of late fusion is that little exploitation of cross correlations is possible.

Hybrid approaches try to make usage of both fusion methods. This is achieved by one or more paths using early fusion, e.g. pairwise combination of modalities [23] which are processed together, as well as multiple paths for the unimodal processing of data. Late fusion is then used to merge the results from all paths. This approach allows for the exploitation of cross-correlations between modalities as well as for the usage of sophisticated models on individual modalities.

B. Multimodal Fusion for Convolutional Neural Networks

With the ongoing success of convolutional neural networks across various fields of application one can argue that the investigation of multimodal CNNs is a promising research direction. Again it is possible to use early fusion and combine features or raw data at an early level. The concatenation of raw data requires usually at least a minimum of preprocessing in order to generate spatio-temporally aligned samples, since sensory devices operate rarely with equal sampling frequencies. If the early fusion is moved to the feature level, the requirement for spatio-temporally alignment remains, but feature extraction can be achieved with several methods. In the classic case features are extracted from raw data with convolutional units which were trained from scratch. But it is also possible to bootstrap the convolutional units with weights already trained on a similar modality, e.g. using weights trained on Imagenet [6] data in order to bootstrap a CNN for video data [5]. Lastly one can also use designed convolutional units which would resemble applying a filter bank of parametrized filters. Late fusion can be performed after a classification result has been generated by unimodal networks. The structure of the individual networks does not need to be similar to each other, again making the fusion task trivial. In the most simple form, the individual dense layers which usually serve as an output layer can be summed and averaged. Hybrid approaches are possible as well, combining both early and late fusion as described above. Note that the hybrid approach is the computationally most expensive, since it requires a set of unimodal as well as fused networks.

We would like to investigate further ways to merge features at different levels of the network architecture. Early and late fusion are the two extremes of the processing pipeline, but a fusion of features at intermediate level can lead to different results than the combination of raw data/first features or final predictions. Typically we receive more complex and specific features the deeper we proceed in a network. The fusion of such complex features correlated for several

modalities can lead to better performance. Unfortunately due to combinatorial explosion of the number of features it is currently not feasible to perform such a fusion at higher levels. If we want to fully exploit cross-correlations between features, we have to perform the full combination of pairwise features resulting in $2^n - 1$ combination for n features [23]. On lower levels though this method might be feasible with current hardware.

IV. METHODS

A. Dataset

Since the EASE dataset is still in the making, we have opted for using the NTU RGB+D dataset [41]. We have selected this particular dataset because it offers several modalities (RGB video, depth video, IR video and skeleton data). Furthermore, it is the largest set offering these modalities with over 56k samples across 60 classes in three different categories: daily, mutual, and health-related actions. The actions were performed by 40 subjects, and recorded from three view points with a Microsoft Kinect v2. Sample frames of NTU RGB+D dataset are shown in Fig. 1.

Fig. 1. Sample frames of the NTU RGB+D dataset [41]. The images illustrate variety in subjects, camera views and intra-class variation.

B. Modelling

So far, we have investigated late fusion with CNNs, as well as the unimodal CNNs which we used for fusion. Here we have taken existing CNNs which offer state of the art performance while having an implementation in TensorFlow [1]. The modalities used for our work are RGB video, optical flow based on the RGB video and skeleton data. Our system has been implemented in TensorFlow.

The architecture of the late fusion system is shown in Fig. 2. The image and optical flow paths consist of "Kinetics I3D" [5] which uses "Inception 3D" units for spatio-temporal processing. The skeleton path consists of "Res-TCN" [19], a residual network for temporal convolution. The other modalities in Fig. 2 refer to potential further extensions. The dense layers of the individual networks are summed, generating the late fusion. We used sparse softmax cross entropy for loss calculation during training.

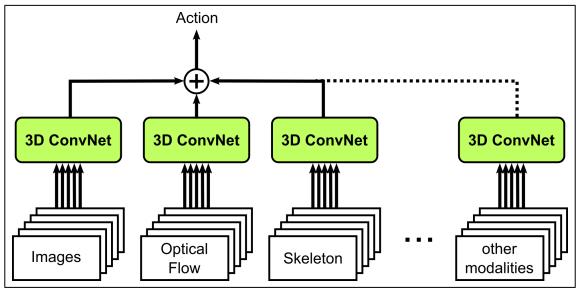


Fig. 2. Summation of dense layers of invididual CNNs as late fusion.

Based on the RGB videos, optical flow has been computed with "FlowNet2" [20]. The original full HD RGB videos were rescaled and cropped to 224x224 pixels, the same has been done with the optical flow videos based upon the RGB videos. The video duration was 10 seconds with 25 fps, with the videos being looped if they were shorter. The skeleton data was provided by the Kinect v2 and consist of the x,y,z coordinates for 25 joints per person. There were maximally two people tracked during the recording.

V. RESULTS

A. RGB Data

We started with the training of unimodal data streams. For the RGB stream the training was done with data from "NTU RGB+D" [41] and with preloaded weights for all layers from "Kinetics I3D" [5]. We retrained only the dense layer, which is a form of transfer learning [37]. The previously trained network has been taught to discriminate actions from a similar dataset and the task as such, activity recognition, remains the same. We achieved an accuracy of 51% which is slightly lower than reported by [30] who reached 56%. Since we only retrained the output layers, this leaves room for improvement by fine tuning the entire network.

B. Skeleton Data

In our experiments for skeleton data, the implementation consists of two models called EASE-1 and EASE-2. The same training and validation splits have been user for these two models. For validation, we apply two standard testing protocols. One is cross-subject, for which half of the subjects are used for training and the rest are used for testing. The other is cross-view, for which 67% of camera views are considered as training data and the rest as test data. The training on skeleton data provided by "NTU RGB+D" was

first performed with the Keras implementation of "Res-TCN" [19], with improved parameters, which we refer to as "EASE-1" in Table I. We use a weight decay of $1e^{-4}$ and stochastic gradient descent (SGD), and adopt the weight initialization and batch normalization [21], but with no dropout and a momentum value close to zero. This model is trained with a mini-batch size of 128 on one GPU. We start with a learning rate of 0.01, divide it by 10 when the testing loss plateaus for more than 10 epochs.

TABLE I
MODIFIED ARCHITECTURE OF RES-TCN WITH HYPERPARAMETER TUNING

	Res-TCN[19]	**EASE-1**	**EASE-2**
Optimizer	SGD	SGD	SGD
Nesterov acceleration	True	False	False
Momentum	0.9	0.01	0.01
L-1 Regulizer	1e-4	1e-4	False
Learning Rate	0.01	0.01	0.01
Batch size	128	128	128
Dropout	0.5	0.01	0.01
Epochs	200	300	200
Weight Initialization	Scratch	Scratch	Pretraining

For "EASE-2", the improved "Res-TCN" is implemented in the TensorFlow deep learning framework. Instead of training from scratch, we train the logits layer only and use the pretrained weights from "EASE-1". All hyperparameters are the same just apart from L-1 regularization in this case.

Table II compares the performance of our approach with published results on the "NTU-RGB+D" [41] dataset. The proposed method shows a considerable improvement on both cross-subject and cross-view settings.

TABLE II
ACCURACY (%) ON NTU RGB+D SKELETON DATASET

Method	Cross Subject	Cross View
HBRNN-L[9]	59.1	64.0
Dynamic Skeleton[14]	60.2	65.2
LieNet[15]	61.4	67.0
P-LSTM[41]	62.9	70.3
ST-LSTM[27]	69.2	77.7
Two-stream RNN[52]	71.3	79.5
Res-TCN[19]	74.3	83.1
Clips+CNN+MTLLN[18]	79.6	84.8
Skepxel[29]	81.3	89.2
ST-GCN[56]	81.5	88.3
EASE-1	79.3	86.1
EASE-2	82.7	–

Loss and accuracy curves for training and validation set are shown in Fig. 3 and 4.

C. Multimodal Processing

The multimodal late fusion has been done in two variants. The first combined two modalities, RGB video and skeleton, and the second combined RGB video, optical flow and skeleton with pretrained weights. Fig. 5 shows the accuracy curve for the training with three modalities so far. In both

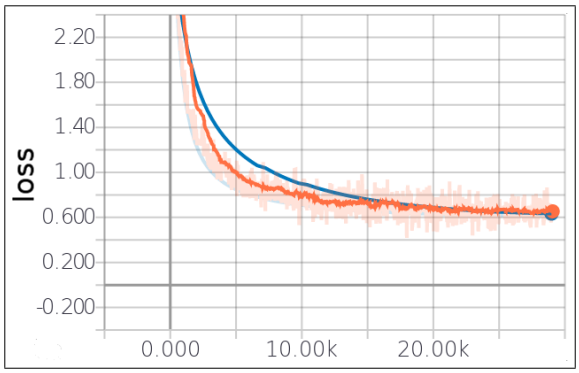


Fig. 3. Loss curve for "EASE-2" for training (orange) and validation set (blue) for skeleton data.

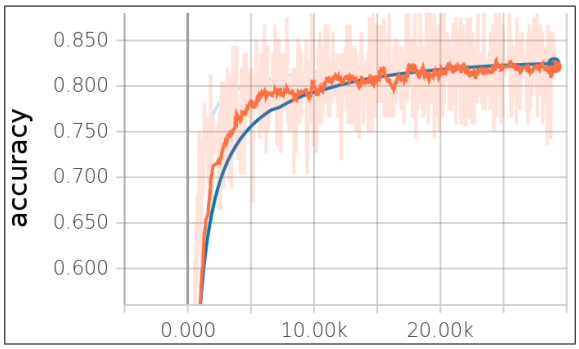


Fig. 4. Accuracy curve for "EASE-2" for training (orange) and validation set (blue) for skeleton data.

cases the resulting accuracy for the evaluation set did not exceed the accuracy of using the skeleton network alone. This suggests that for this particular dataset the skeleton data carry particularly informative features, while the additional RGB and optical flow data do not seem to contribute improvements. The skeleton data dominate the results.

VI. CONCLUSIONS

We study multimodal fusion architectures for convolutional neural networks. At this point, we have investigated a late fusion approach and individual CNNs for different modalities (video, optical flow and skeleton data) based on the "NTU RGB+D" [41] dataset. Our experimental investigations showed that modality-specific training of a CNN on skeleton data outperforms previous state-of-the-art skeleton based models on the standard large scale human activity recognition dataset. Our analysis of a late fusion approach revealed that if the modalities are only individually trained the result appears to be dominated by the contribution of the skeleton path. In the future, we will focus on a more effective training of 3D CNNs and on the fusion of other modalities.

ACKNOWLEDGMENT

The research reported in this paper has been supported by the German Research Foundation DFG, as part of Col-

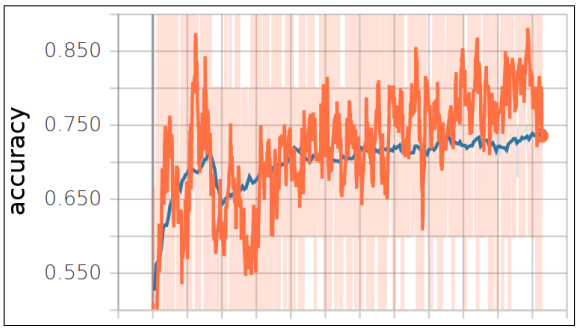


Fig. 5. Accuracy curve for multimodal late fusion for training (orange) and validation set (blue) for skeleton and RGB data.

laborative Research Center (Sonderforschungsbereich) 1320 “EASE - Everyday Activity Science and Engineering”, University of Bremen (http://www.ease-crc.org/). The research was conducted in subproject “H3 – Natural activity statistics”.

REFERENCES

[1] M Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, M. Kudlur, J. Levenberg, R. Monga, S. Moore, D. Murray, B. Steiner, P. Tucker, V. Vasudevan, P. Warden, M. Wicke, Y. Yu, X. Zheng, TensorFlow: A System for Large-scale Machine Learning, in Proceedings of the 12th USENIX Conference on Operating Systems Design and Implementation, pp. 265–283, 2016.

[2] U. C. Allard, F. Nougarou, C. L. Fall, P. Gigu‘ere, C. Gosselin, F. Lavio-lette, and B. Gosselin, Deep Learning for Electromyographic Hand Gesture Signal Classification Using Transfer Learning, Arxiv, 2018.

[3] M. Atzori, M. Cognolato, and H. Müller Deep learning with convolutional neural networks applied to electromyography data: A resource for the classification of movements for prosthetic hands, in Frontiers in neurorobotics, vol. 10, 2016.

[4] S. Buch, V. Escorcia, C. Shen, B. Ghanem, J. C. Niebles, Sst: Single-stream temporal action proposals, in CVPR, 2017.

[5] J. Carreira and A. Zisserman, Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset, in 2017 IEEE Conference on Computer Vision and Pattern, pp. 4724-4733, 2017.

[6] J. Deng, W. Dong, R. Socher, L. Li, Kai Li and Li Fei-Fei, ImageNet: A large-scale hierarchical image database, in 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, 2009, pp. 248-255. doi: 10.1109/CVPR.2009.5206848

[7] S. K. D’mello and J. Kory, A Review and Meta-Analysis of Multimodal Affect Detection Systems, ACM Computing Surveys, 2015.

[8] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, Long-term recurrent convolutional networks for visual recognition and description. In CVPR, pages 2625–2634, 2015.

[9] Y. Du, W. Wang, and L. Wang, Hierarchical recurrent neural network for skeleton based action recognition, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp.1110–1118.

[10] Y. Du, Y. Fu, L. Wang, Skeleton based action recognition with convolutional neural network, in: Pattern Recognition (ACPR), 2015 3rd IAPR Asian Conference on, IEEE, 2015, pp. 579–583.

[11] Y. Du, Y. Fu, L. Wang, Representation learning of temporal dynamics for skeleton-based action recognition, IEEE Transactions on Image Processing 25 (7) (2016) 3010–3022.

[12] Y. Du, W. Jin, W. Wei, Y. Hu, and W. Geng, Surface emg-based inter-session gesture recognition enhanced by deep domain adaptation, in Sensors, vol. 17, no. 3, p. 458, 2017.

[13] Y. Hou, Z. Li, P. Wang, W. Li, Skeleton optical spectra based action recognition using convolutional neural networks, in: Circuits and Systems for Video Technology, IEEE Transactions on, 2016, pp. 1–5.

[14] Hu, J.F., Zheng, W.S., Lai, J., Zhang, J., Jointly learning heterogeneous features for rgb-d activity recognition, in IEEE conference on Computer Vision and Pattern Recognition, 5344-5352, 2015.

[15] Huang, Z., Wan, C., Probst, T., Van Gool, L., Deep learning on lie groups for skeleton-based action recognition, in IEEE Conference on Computer Vision and Pattern Recognition, 2016.

[16] D.-A. Huang, L. Fei-Fei, J. C. Niebles, Connectionist temporal modeling for weakly supervised action labeling, in European Conference on Computer Vision, Springer, pp. 137–153, 2016.

[17] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, L. Fei-Fei, Large-scale video classification with convolutional neural networks, in Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1725–1732, 2014.

[18] Q. Ke, M. Bennamoun, S. An, F. Sohel, and F. Boussaid, A new representation of skeleton sequences for 3d action recognition, arXiv preprint arXiv:1703.03492, 2017.

[19] T. S. Kim and A. Reiter, Interpretable 3d human action analysis with temporal convolutional networks, in 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW).

[20] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy and T. Brox, FlowNet 2.0: Evolution of Optical Flow Estimation with Deep Networks, in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Jul 2017.

[21] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:1502.03167 , 2015.

[22] S. Ji, W. Xu, M. Yang, K. Yu, 3D convolutional neural networks for human action recognition, in IEEE Trans. Pattern Anal. Mach. Intell., vol. 35, no. 1, pp. 221-231, Jan. 2013.

[23] Z. Z. Lan, L. Bao, S. I. Yu, W. Liu, and A. G. Hauptmann, Multimedia classification and event detection using double fusion, in Multimedia Tools and Applications, 2014.

[24] Q. V. Le, W. Y. Zou, S. Y. Yeung, A. Y. Ng, Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis, in: Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on, IEEE, pp. 3361–3368, 2011.

[25] Y. Li, Q. Miao, K. Tian, Y. Fan, X. Xu, R. Li, J. Song, Large-scale gesture recognition with a fusion of rgb-d data based on the c3d model, in: Pattern Recognition (ICPR), 2016 23rd International Conference on, IEEE, 2016, pp. 25–30.

[26] C. Li, Y. Hou, P. Wang, W. Li, Joint distance maps based action recognition with convolutional neural networks, IEEE Signal Processing Letters 24 (5) (2017) 624–628.

[27] J. Liu, A. Shahroudy, D. Xu, and G. Wang, Spatio-temporal LSTM with trust gates for 3d human action recognition, arXiv preprint arXiv:1607.07043, 2016.

[28] C. Liu, Y. Hu, Y. Li, S. Song and J. Liu, PKU-MMD: A Large Scale Benchmark for Continuous Multi-Modal Human Action Understanding, CoRR, arXiv:1703.07475, 2017.

[29] J. Liu, N. Akhtar, A. Mian, Skepxels, Spatio-temporal Image Representation of Human Skeleton Joints for Action Recognition, arxiv, 2018.

[30] Z. Luo, B. Peng, D.A. Huang, A. Alahi, L. Fei-Fei, Unsupervised learning of long-term motion dynamics for videos, in: CVPR. 2017.

[31] J. L. Maldonado C., T. Kluss and C. Zetzsche, Exploring Human Kinematic Control for Robotics Applications: The Role of Afferent Sensory Information in a Precision Task, in IROS 2018: Towards Robots that Exhibit Manipulation Intelligence, Madrid, Spain, 2018 (accepted).

[32] C. Mason, M. Meier, F. Ahrens, F. Putze and T. Schultz, Human Activities Data Collection and Labeling using a Think-aloud Protocol in a Table Setting Scenario, IROS 2018: Workshop on Latest Advances in Big Activity Data Sources for Robotics and New Challenges, Madrid, Spain, 2018 (accepted).

[33] M. Meier, C. Mason, R. Porzel, F. Putze, T. Schultz, Synchronized Multimodal Recording of a Table Setting Dataset, IROS 2018: Workshop on Latest Advances in Big Activity Data Sources for Robotics and New Challenges, Madrid, Spain, 2018 (accepted).

[34] T. Baltrušaitis, C. Ahuja, and Louis-Philippe Morency, Multimodal Machine Learning: A Survey and Taxonomy, in CoRR, vol. abs/1705.09406, May 2017 (eprint arXiv:1705.09406).

[35] J. Y.-H. Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici, Beyond short snippets: Deep networks for video classification, in CVPR, pages 46944702, 2015.

[36] M. A. Oskoei and H. Hu, Myoelectric control systems a survey, in Biomedical Signal Processing and Control , vol. 2, no. 4, pp. 275–294, 2007.
[37] L. Y. Pratt, "Discriminability-based transfer between neural networks, in NIPS Conference: Advances in Neural Information Processing Systems 5, Morgan Kaufmann Publishers. pp. 204–211, 1993.
[38] H. Rahmani, A. Mian, 3d action recognition from novel viewpoints, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1506–1515, 2016.
[39] R. Salakhutdinov, J. B. Tenenbaum, A. Torralba, Learning with hierarchical-deep models, IEEE Transactions on Pattern Analysis and Machine Intelligence 35 (8) (2013) 1958–1971.
[40] S. Sarkar, K. Reddy, A. Dorgan, C. Fidopiastis, M. Giering, Wearable EEG-based Activity Recognition in PHM-related Service Environment via Deep Learning,IJPHM, vol. 7, no. 21, 2016.
[41] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang, NTU RGB+D: A large scale dataset for 3D human activity analysis, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016.
[42] Y. Shi, Y. Tian, Y. Wang, W. Zeng, T. Huang, Learning long-term dependencies for action recognition with a biologically-inspired deep network, in ICCV, pp. 716–725, 2017.
[43] Z. Shi, T.-K. Kim, Learning and refining of privileged information-based rnns for action recognition from depth sequences, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.
[44] Z. Shou, D. Wang, and S.-F. Chang. Temporal action localization in untrimmed videos via multi-stage CNNs, in CVPR, pages 10491058, 2016.
[45] K. Simonyan, A. Zisserman, Two-stream convolutional networks for action recognition in videos, in Proc. Adv. Neural Inf. Process. Syst. Conf., pp. 568-576, Dec. 2014.
[46] CGM. Snoek, M. Worringm, AWM. Smeulders, Early versus late fusion in semantic video analysis, in ACM international conference multimedia (MM'05), 2005.
[47] G. W. Taylor, R. Fergus, Y. LeCun, C. Bregler, Convolutional learning of spatio-temporal features, in Proc. Eur. Conf. Comput. Vis., pp. 140-153, Sep. 2010.
[48] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, Learning spatiotemporal features with 3D convolutional networks, in 2015 IEEE International Conference on Computer Vision (ICCV). IEEE, 2015, pp. 44894497.
[49] V. Veeriah, N. Zhuang, G.-J. Qi, Differential recurrent neural networks for action recognition, in ICCV, 2015.
[50] P. Wang, W. Li, Z. Gao, J. Zhang, C. Tang, P. Ogunbona, Action recognition from depth maps using deep convolutional neural networks, THMS 46 (4) pp. 498–509, 2016.
[51] L. Wang, Y. Qiao, and X. Tang, Action recognition with trajectory-pooled deep-convolutional descriptors, in CVPR,pages 43054314, 2015.
[52] H. Wang, L. Wang, Modeling Temporal Dynamics and Spatial Configurations of Actions Using Two-Stream Recurrent Neural Networks, in IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), 2017.
[53] P. Wang, S. Wang, Z. Gao, Y. Hou, W. Li, Structured images for rgb-d action recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1005–1014.
[54] P. Wang, W. Li, P. Ogunbona, J. Wan and S. Escalera, RGB-D-based Human Motion Recognition with Deep Learning: A Survey, CoRR, arXiv:1711.08362, 2017.
[55] D. Wu, L. Shao, Leveraging hierarchical parametric networks for skeletal joints based action segmentation and recognition, in 2014 IEEE Conference on Computer Vision and Pattern Recognition, pp. 724–731, 2014.
[56] S. Yan, Y. Xiong, D. Lin, Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition, arXiv, 2018.
[57] Y. Zhao, Y. Xiong, L. Wang, Z. Wu, D. Lin, X. Tang, Temporal action detection with structured segment networks, in ICCV, 2017.
[58] G. Zhu, L. Zhang, L. Mei, J. Shao, J. Song, P. Shen, Large-scale isolated gesture recognition using pyramidal 3d convolutional networks, in: Pattern Recognition (ICPR), in 2016 23rd International Conference on, IEEE, 2016, pp. 19–24.

Deep Residual Temporal Convolutional Networks for Skeleton-Based Human Action Recognition ⋆

R. Khamsehashari, K. Gadzicki, C. Zetzsche

Cognitive Neuroinformatics, University of Bremen, Germany
{rkhamseh,gadzicki}@uni-bremen.de, zetzsche@cs.uni-bremen.de

Abstract. Deep residual networks for action recognition based on skeleton data can avoid the degradation problem, and a 56-layer Res-Net has recently achieved good results. Since a much "shallower" 11-layer model (Res-TCN) with a temporal convolution network and a simplified residual unit achieved almost competitive performance, we investigate deep variants of Res-TCN and compare them to Res-Net architectures. Our results outperform the other approaches in this class of residual networks. Our investigation suggests that the resistance of deep residual networks to degradation is not only determined by the architecture but also by data and task properties.

Keywords: Deep residual networks · action recognition · degradation · hyperparameters.

1 INTRODUCTION

Human activity recognition has become a prominent research area due to its potential application in video surveillance, human computer interfaces, ambient assisted living, human-robot-interaction, etc. More recently it has also become important for understanding pedestrian behavior in autonomous driving [12]. As in other areas of artificial intelligence, deep learning techniques have also gained exceptional achievements in human activity recognition. Particularly, deep learning methods based on Recurrent Neural Networks (RNN) and Convolutional Neural Networks (CNN) architectures have shown great performance in classification tasks by learning discriminant features from large amounts of data.

Residual networks (e.g. Res-Net [1]) can avoid the degradation problem in deep CNN architectures. Originally developed for image recognition tasks, they have recently been extended to human activity recognition [7]. As to be expected for residual architectures, good performance levels could be obtained with quite

⋆ This work has been supported by the German Aerospace Center (DLR) with financial means of the German Federal Ministry for Economic Affairs and Energy (BMWi), project "OPA^{3}L" (grant No. 50 NA 1909) and by the German Research Foundation DFG, as part of CRC (Sonderforschungsbereich) 1320 "EASE - Everyday Activity Science and Engineering", University of Bremen (http://www.ease-crc.org/).

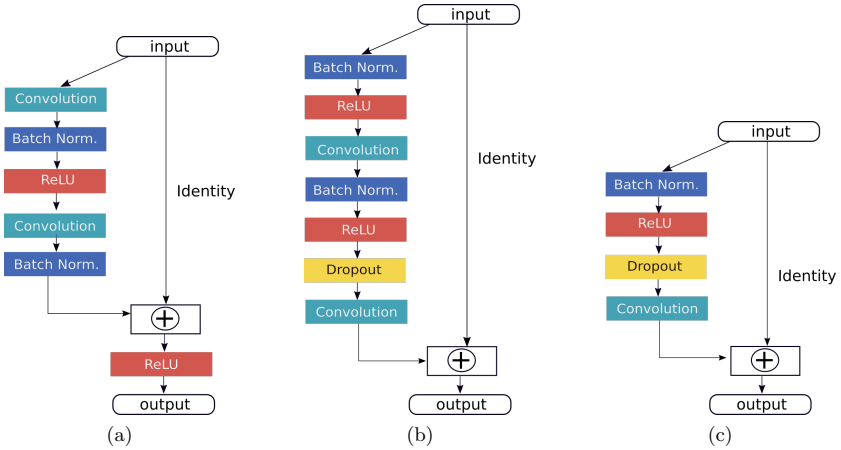


Fig. 1: The basic residual unit in different approaches. (a) original ResNet [1]; (b) improved ResNet [7]; (c) Res-TCN [4]

deep architectures, i.e. 56 to 110 layers [7]. Somewhat surprisingly, however, a recent approach which combined the residual network concept with a temporal convolutional network architecture (Res-TCN [4]) could achieve an almost comparable performance with a comparably "shallow" configuration of 11 layers. And this, although the basic building unit of this approach also is considerably simpler than that of the Res-Net architecture (cf. Figure 1), and offers in addition a much easier interpretability.

However, while the ResNet approaches have been systematically investigated with respect to the overall depth of the architecture (from 20 up to 110 layers [7]) the Res-TCN approach has only been tested in one single 11-layer variant. This prompted us to investigate whether very deep network architectures for activity recognition can profit from using the structurally simple residual unit of Res-TCN as basic building block, and a representation based on temporal convolutions. We introduce two variants of a deep learning architecture, Deep Res-TCN-3 and Deep Res-TCN-4, with depths ranging from 11 to 152 layers, to learn features from skeleton data and classify them into action classes.

2 RELATED WORK

Skeleton data are of particular interest for human activity recognition as they carry high level information in the form of abstracted 3D joint positions. They can be obtained by optical tracking of body markers, from depth videos (e.g. Microsoft Kinect), or from RGB video data with pose estimation methods [2].

We shortly review the central ideas behind the residual based architectures such as Res-Net [1] and Res-TCN [4]. Res-Net employs injected residual connections between processing streams to allow spatial-temporal interaction between them. Res-TCN redesigned the original TCN [5] by factoring out the deeper layers into additive residual terms that yielded both an interpretable hidden

representation and model parameters. Each unit in layer $L+1$ performs the following computations in Res-TCN and Res-Net, according to equations 1 and 3 respectively. In residual based networks, the traditional convolutional layers calculate a residual which is added to the input of the layers (see Figure 1). One of the differences of Res-TCN and Res-Net is the direct reference to the first convolution layer, which according to Equation 2 operates on the raw skeletal input and the created activation map, $X1$, is passed on to the subsequent layers.

$$X_L = X_1 + \sum_{i=2}^{L} W_i * max(0, X_{i-1}) \tag{1}$$

where

$$X_1 = W_1 * X_0 \tag{2}$$

$$X_{L+1} = W_L * max(0, X_L) + X_L. \tag{3}$$

X_L and X_{L+1} are input and output features of the L_{t_h} residual unit, respectively. The architecture of a residual based network is shown in Figure 2.

The basic residual unit of Res-TCN [4], as compared to that of the original ResNet [1], does not use ReLUs behind the element-wise additions $\oplus$ (see Figure 1a,c) and can thus provide readily interpretable representations. In addition, such units produce a direct path that enables the signal to be directly propagated in a forward pass through the entire network to any unit and also the gradients can be backwards propagated to any unit (cf. [7]). Finally, the Res-TCN building block is considerably shallower compared to both the original [1] and the improved ResNet [7] (cf. Figure 1).

3 METHODS

In this study we investigate different variants of deep residual architectures. Residual networks are assumed to cope with the degradation phenomenon by allowing for the fusion of all lower-level features of previous layers in the deeper layers, thus enabling more complex mappings to higher level feature maps. Approaches with residual units in both image processing [1] and action recognition [7] indicate that performance can be systematically improved by deeper architectures. For this study, we were particularly interested whether these advantages of deeper architectures can be successfully combined with the temporal representation and the simpler and more shallow nature of the basic Res-TCN building block [4], and how the performance of the resulting architectures compares to the more sophisticated approaches used in [1] and [4] (cf. also Figure 1).

For our investigation and performance comparisons we wanted to have a broad coverage of the depth dimension, reaching from the relative shallow depth of 11 layers, as used in the original Res-TCN approach [4], up to a quite high depth of 152 layers, the maximum depth used in the original Res-Net study [1].

Furthermore, we organized our investigations into three, more specific research questions: First, we wanted to find out whether it is possible to use the

Table 1: Specification of architectures. The values in the brackets state the filter length and number of features of each building block. The number of stacked building blocks is given after the brackets. Down sampling is performed within `conv1` of each block A to D with a stride of 2. The 11-layer variant corresponds to the original Res-TCN architecture.

(a) Deep Res-TCN-3

Layer Name	**Output Size**	**11-layer**	**18-layer**	**34-layer**	**60-layer**	**101-layer**	**152-layer**
Conv1	300	8,64					
Block A	300	[8,64]*3	[8,64]*5	[8,64]*10	[8,64]*16	[8,64]*33	[8,64]*50
Block B	150	[8,128]*3	[8,128]*5	[8,128]*11	[8,128]*16	[8,128]*33	[8,128]*50
Block C	75	[8,256]*3	[8,256]*6	[8,256]*11	[8,256]*16	[8,256]*33	[8,256]*50
Average pool, fc-60, softmax	1						

(b) Deep Res-TCN-4

Layer Name	**Output Size**	**11-layer**	**18-layer**	**34-layer**	**60-layer**	**101-layer**	**152-layer**
Conv1	300	8,64					
Block A	300	[8,64]*3	[8,64]*4	[8,64]*6	[8,64]*9	[8,64]*33	[8,64]*46
Block B	150	[8,128]*3	[8,128]*4	[8,128]*8	[8,128]*12	[8,128]*33	[8,128]*46
Block C	75	[8,256]*3	[8,256]*4	[8,256]*12	[8,256]*18	[8,256]*22	[8,256]*35
Block D	38	-	[8,512]*4	[8,512]*6	[8,512]*9	[8,512]*11	[8,512]*23
Average pool, fc-60, softmax	1						

simplified residual unit of Res-TCN for the design of deeper architectures, in order to obtain an improved classification performance. We stick with the setting of the original Res-TCN architecture [4], Res-TCN-3 in Table 1a, and varied the depth only, in order to have an as direct as possible comparison. We did not use the bottleneck architecture from [1], since we wanted to avoid a change in architecture along the depth dimension. Control tests of a few bottleneck variants of the networks also indicated no advantage in classification performance.

The second research question addresses in how far the simpler structure of the basic residual unit of the Res-TCN architecture [4] will possibly limit performance in comparison to the more sophisticated Res-Net architectures [1, 7] (cf. also Figure 1). Since those architectures make use of a 4-block design, we also designed a 4-block deep Res-TCN-4 architecture (see Table 1b).

The third research question addressed the improvement of the hyperparameters that we found to be optimal for training (see Table 2) which are different from the ones used in the original Res-TCN study [4]. In order to disentangle the influence of the hyperparameters from those of the other properties varied in this study we hence tested a further set of Res-TCN-4 networks which have been trained with the original Res-TCN parameters as described in [4].

We used the residual units of Res-TCN ([4], Figure 1c), and repeated them within one block multiple times. Between blocks the signal is down-sampled by a factor of 2 by convolution with stride 2 in the first element of each block. The convolutions are 1-dimensional with length 8 throughout the network. The number of features varies from 64 in the early stages up to 512 in the later ones. Figure 2 shows an example of our architecture.

Table 2: Modified architectures of Deep Res-TCN with hyperparameter tuning.

	Original Res-TCN[4]	**Deep Res-TCN-3** **Deep Res-TCN-4**
Optimizer	SGD	SGD
Nesterov acceleration	True	False
Momentum	0.9	0.01
L-1 Regulizer	1e-4	1e-4
Learning Rate	0.01	0.01
Batch size	128	128
Dropout	0.5	0.4
Epochs	200	200
Weight Initialization	Scratch	Scratch

We evaluate the architectures on the 3D skeleton based human activity recognition dataset NTU RGB+D [8], the currently largest set offering several modalities (RGB video, depth video, IR video and skeleton data) with more than 56k training videos across 60 action classes. The actions were performed by 40 distinct subjects, and recorded from three view points with a Microsoft Kinect v2. The dataset provides two different evaluation criteria: Cross-Subject (CS) and Cross-View (CV). The skeleton data used are the x, y, z coordinates for 25 joints per person. There are maximally two people tracked during the recording.

For validation, we apply two standard testing protocols. One is Cross-Subject, for which half of the subjects are used for training and the rest are used for testing. The other is Cross-View, for which 67% of camera views are considered as training data and the rest as test data. The training on skeleton data provided by NTU RGB+D was performed with the Keras implementation of Res-TCN [4]. The hyperparameters of the proposed architectures are tuned according to Table 2 with weight decay of $1e^{-4}$, stochastic gradient descent (SGD), and we adopt the weight initialization and batch normalization [3]. We set the mini-batch size of 128 on one GPU for all structures except for the 152-layer networks for which we used a batch size of 96. The training is started with rate of 0.01, divided by

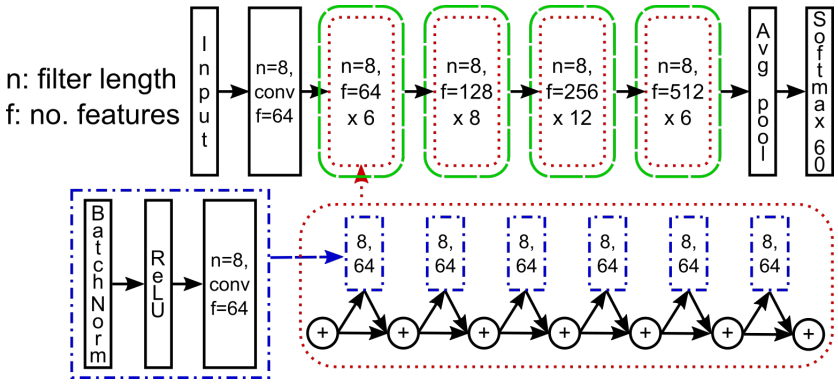


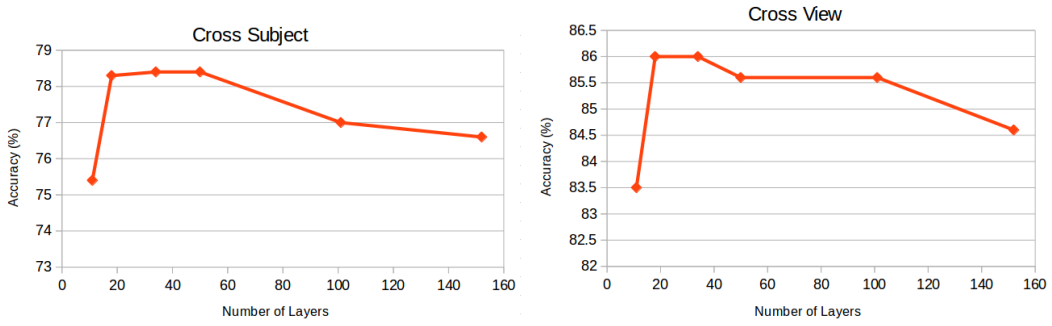
Fig. 2: Deep Res-TCN-4 architecture with 34 layers.

10 when the testing loss plateaus for more than 10 epochs. Finally, we use sparse softmax cross entropy for loss calculation during training. Evaluation of the loss curves of training and validation set did not show any indication for overfitting problems .

4 RESULTS

In the following we present graphs of the classification performance in dependence on network depth to answer the research questions of sect. 3. A detailed summary of all results in form of numerical values can be found in Table 3.

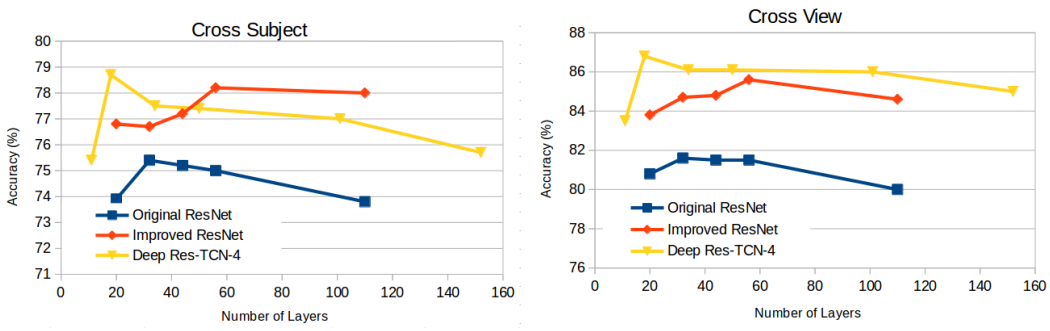
(1) *Can we use the simpler residual unit of Res-TCN to design deeper networks with improved performance?* As mentioned we want to keep this comparison straightforward and thus stick here to the 3-block design of the original Res-TCN [4]. Figure 3 shows the performance curve for this 11-layer Res-TCN and for our models with depths between 18 and 151 layers. The main result is that the deeper variants *all* provide a significantly improved classification accuracy in comparison to the original 11-layer Res-TCN architecture. However, the optimum occurs already at comparatively moderate depth levels of 34 and 18 layers for cross-subject and cross-view tests, respectively. But the optimum is a shallow one, and the decrease of accuracy for greater depths is quite moderate.



(a) (b)

Fig. 3: Influence of network depth on accuracy (Deep Res-TCN-3). The leftmost data point corresponds to the original 11-layer Res-TCN architecture.

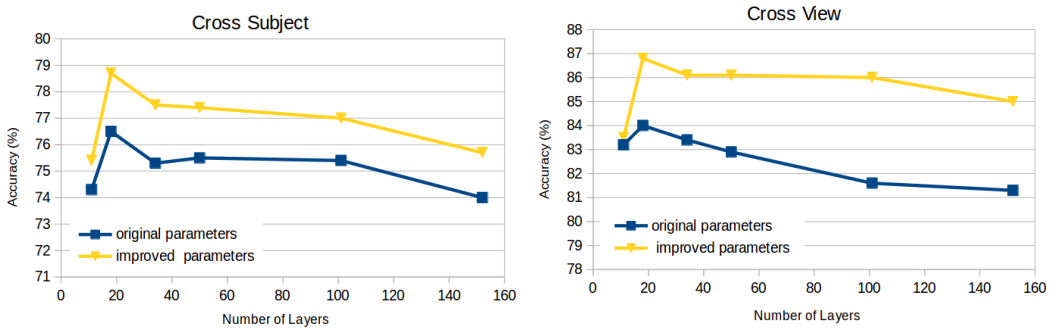
(2) *How does the simpler architecture of Res-TCN compare to the more sophisticated Res-Net architectures* [1, 7]? (*cf. also* Figure 1) Comparison is based on a 4-block design (Res-TCN-4), as used in the Res-Net architectures. Classification accuracy curves are shown in Figure 4. Although the Res-TCN-4 architecture is simpler than those of the two Res-Net variants its classification accuracy is similar or even superior. In particular it provides the best performance level (78.7% for cross-subject and 86.8% for cross-view tests). Again, this optimum is achieved for a relatively moderate depth of 18 layers (Res-TCN-4-18).



(a) (b)

Fig. 4: Accuracy curves of different architectures.

(3) *Influence of hyperparameters.* Figure 5 shows how the classification accuracy is improved by our new hyperparameters, particularly for cross-view.



(a) (b)

Fig. 5: Influence of hyperparameters.

Table 3 shows the accuracy results for the different deep Res-TCN variants we tested (and for the original Res-TCN-11). A comparison of our models with other ResNet-like architectures is shown in Table 4. The best Res-TCN-4-18 network trained with improved parameters achieved the overall best performance, an accuracy of 78.7% for cross-subject and of 86.8% for cross-iew.

5 DISCUSSION

In the previous section we demonstrated that the proposed deep Res-TCN architecture clearly outperforms the original "shallow" 11-layer Res-TCN [4]. In addition, it outperforms also alternative deep residual approaches of the Res-Net class [1, 7], using the same experimental setting and dataset.

A somewhat surprising result is a systematic pattern observed for all architectures (except the improved Res-Net [7]): a strong initial performance gain

Table 3: Recognition accuracy. The best results are highlighted in bold.

(a) Cross-Subject accuracy

Architectures	Original Parameters		Improved Parameters	
	Res-TCN	**Deep Res-TCN-4**	**Deep Res-TCN-3**	**Deep Res-TCN-4**
Original Res-TCN-11	74.3	–	75.4	–
Res-TCN-18	–	**76.5**	78.3	**78.7**
Res-TCN-34	–	75.3	**78.4**	77.5
Res-TCN-50	–	75.5	**78.4**	77.4
Res-TCN-101	–	75.4	77.0	77.0
Res-TCN-152	–	74.0	76.6	75.7

(b) Cross-View accuracy

Architectures	Original Parameters		Improved Parameters	
	Res-TCN	**Deep Res-TCN-4**	**Deep Res-TCN-3**	**Deep Res-TCN-4**
Original Res-TCN-11	83.2	–	83.5	–
Res-TCN-18	–	**84.0**	**86.0**	**86.8**
Res-TCN-34	–	83.4	**86.0**	86.1
Res-TCN-50	–	82.9	85.6	86.1
Res-TCN-101	–	81.6	85.6	86.0
Res-TCN-152	–	81.3	84.6	85.0

Table 4: Relationship between number of layers on ResNet and Res-TCN based architectures and their best performance on the NTU-RGB+D dataset. The numbers are in format cross-subject / cross-view.

Method	**11-layer**	**18-layer**	**32-layer**	**34-layer**	**50-layer**	**56-layer**
Res-TCN [4]	74.3 / 83.1					
Original ResNet [7]			75.4 / 81.6			
Improved ResNet [7]						78.2 / 85.6
Deep Res-TCN-3		– / 86.0		78.4 / 86.0	78.4 / –	
Deep Res-TCN-4		78.7 / 86.8				

due to the first step in depth, followed by a rapid leveling-off, or even a shallow decrease, for the deeper network variants. This effect is most expressed for the Res-TCN-4 architecture, where an increase from 11 to 18 layers yields the absolute optimum performance of all architectural variants considered in this study (78.7% and 86.8% accuracy for Cross-Subject and Cross-View, respectively, see also Figure 4). The same pattern arises also with the Res-TCN-3 architectures (Figure 3). The effect is not caused by the specific deep Res-TCN architectures

since the original Res-Net architecture shows a similar pattern both for Cross-Subject and, less expressed, for Cross-View [7] (Figure 4). However, the effect is also not solely caused by the dataset, since it does not occur with the improved Res-Net suggested by [7] which produces a systematic slow performance gain with increasing network depth, with an optimum at 56 layers (Figure 4).

What can we learn from these results? First, the systematic pattern observed for all architectures but the improved Res-Net indicates problems with degradation. Although residual architectures have the basic potential to cope with degradation, they here seem to fail at early network stages. In the case of the Res-TCN variants this may be attributed to the simplified structure of the residual unit with only one convolutional layer, which limits the power for nonlinear approximation. But why do we observe a similar behavior for the original Res-Net, an architecture that has proven to be able to cope with the degradation phenomenon for a variety of datasets [1]? The cause for this could be the format of the representation. The deep Res-TCN architectures are based on purely temporal convolutions, the interrelations between joints being represented by the filters. Although the order of joints has been carefully rearranged in [7], and has proven to be essential for the performance, we assume that the 2-D convolutions are not optimally suited for the representation. This idea is also supported by the fact that our Res-TCN-4 network (with limited power of the basic residual unit) can outperform the improved 56-layer Res-Net architecture by using only as few as 18 layers. Taken together this suggests that the resistance to degradation effects is not solely determined by the specific residual structure of a network but also by a non-trivial interaction of architecture and task/data properties.

It is worth mentioning that the influence of the hyperparameters is quite strong compared to the other effects (Figure 5). Since the main difference is the avoidance of momentum optimization this might indicate that this common default choice is not optimally suited for the relatively smooth and regular loss landscape of deep residual architectures [6], or will require additional measures to ensure full convergence.

Among the class of models which use only skeleton data and straight-forward deep residual network architectures the deep Res-TCN model suggested here provides the best performance. Other recent approaches exhibit even better performance levels, but these are achieved using quite specialized model structures, e.g. by making use of additional attention modules [9], multi-modal processing streams [11], view point adaptation [10] or other improvements e.g. [13–17]. It has to be expected that further gains can be obtained by a combination of our model with these more sophisticated approaches, for example by using our model as one component in a multimodal architecture.

References

1. K. He, X. Zhang,S. Ren and J. Sun. Deep Residual Learning for Image Recognition, IEEE Conference on Computer Vision and Pattern Recognition(CVPR) , 770778, 2016.

2. K. He, G. Gkioxari, P. Dollr and R. Girshick. Mask R-CNN, in 2017 IEEE International Conference on Computer Vision (ICCV), Venice, 2017, pp. 2980-2988. doi: 10.1109/ICCV.2017.322
3. S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:1502.03167 , 2015.
4. T. S. Kim and A. Reiter. Interpretable 3d human action analysis with temporal convolutional networks, in 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW).
5. C. Lea, M. D. Flynn, R. Vidal, A. Reiter, and G. D. Hager. Temporal convolutional networks for action segmentationand detection. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2017.
6. H. Li, Z. Xu, G. Taylor and T. Goldstein. Visualizing the Loss Landscape of Neural Nets, in: CoRR, 2017. arXiv:1712.09913 [cs.LG]
7. H.Pham, L. Khoudour, A. Crouzil, P. Zegers and S. Velastin. Exploiting deep residual networks for human action recognition from skeletal data, Computer Vision and Image Understanding (CVIU), 2018.
8. A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang. NTU RGB+D: A large scale dataset for 3D human activity analysis, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016.
9. Z. Yang, Y. Li, J. Yang and J. Luo. Action Recognition with Visual Attention on Skeleton Images, in: CoRR, 2018. arXiv:1804.07453 [cs.CV]
10. P. Zhang, C. Lan, J. Xing, W. Zeng, J. Xue, and N. Zheng. View Adaptive Neural Networks for High Performance Skeleton-based Human Action Recognition, in IEEE Transactions on Pattern Analysis and Machine Intelligence (to appear), 2019.
11. J. Zhu, W. Zou, L. Xu, Y. Hu, Z. Zhu, M. Chang, J. Huang, G. Huang and D Du. Action Machine: Rethinking Action Recognition in Trimmed Videos, in: CoRR, 2019. arXiv:1812.05770 [cs.CV]
12. A. Rasouli and J. K. Tsotsos. Joint Attention in Driver-Pedestrian Interaction: from Theory to Practice, in: CoRR, 2018. arXiv:1802.02522 [cs.RO]
13. M. Liu, L. Hong and C. Chen. Enhanced skeleton visualization for view invariant human action recognition. Pattern Recognition 68 (2017): 346-362.
14. C. Li, P. Wang, S. Wang, Y. Hou and W. Li. Skeleton-based action recognition using LSTM and CNN. 2017 IEEE International Conference on Multimedia & Expo Workshops (ICMEW). IEEE, 2017.
15. C. Li, Q. Zhong, D. Xie and S. Pu. Skeleton-based action recognition with convolutional neural networks. 2017 IEEE International Conference on Multimedia & Expo Workshops (ICMEW). IEEE, 2017.
16. Q. Ke, M. Bennamoun S. An, F. Sohel and F. Boussaid. A new representation of skeleton sequences for 3d action recognition. Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.
17. S. Yan, Y. Xiong and D. Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. Thirty-Second AAAI Conference on Artificial Intelligence. 2018.

Early vs Late Fusion in Multimodal Convolutional Neural Networks

1st Konrad Gadzicki
Cognitive Neuroinformatics
University of Bremen
Bremen, Germany
gadzicki@uni-bremen.de

2nd Razieh Khamsehashari
Cognitive Neuroinformatics
University of Bremen
Bremen, Germany
rkhamseh@uni-bremen.de

3rd Christoph Zetzsche
Cognitive Neuroinformatics
University of Bremen
Bremen, Germany
zetzsche@informatik.uni-bremen.de

***Abstract*—Combining machine learning in neural networks with multimodal fusion strategies offers an interesting potential for classification tasks but the optimum fusion strategies for many applications have yet to be determined. Here we address this issue in the context of human activity recognition, making use of a state-of-the-art convolutional network architecture (Inception I3D) and a huge dataset (NTU RGB+D). As modalities we consider RGB video, optical flow, and skeleton data. We determine whether the fusion of different modalities can provide an advantage as compared to uni-modal approaches, and whether a more complex early fusion strategy can outperform the simpler late-fusion strategy by making use of statistical correlations between the different modalities. Our results show a clear performance improvement by multi-modal fusion and a substantial advantage of an early fusion strategy.**

***Index Terms*—Multi-layer neural network, Activity recognition, Sensor fusion**

I. INTRODUCTION

The research of human activity recognition has gained attention over the years due to its utilization in various fields. The collaborative research center "EASE" (http://www.ease-crc.org) has the goal to develop robots capable of performing everyday activities. Examples from humans executing household activities like table setting, cooking etc. serve as an important information source for determining appropriate actions of the robot. The automated recognition of those activities is key for the access and analysis of data in a huge database of recorded human activities.

Deep learning had a tremendous impact on machine learning and pattern recognition, achieving results beyond the performance levels of classical approaches. Tasks like activity recognition profit substantially from deep learning, e.g. by utilizing the expressive power of Recurrent Neural Networks (RNN) and Convolutional Neural Networks (CNN). These approaches perform well on a large number of different sensory data relevant for activity recognition, e.g. image data, video, audio or skeleton data. Since activities are mostly recorded over time, the convolutional units of appropriate CNNs have to comprise the temporal dimension as well by applying some sort of spatio-temporal convolutions.

This work was supported by DFG (German Research Foundation) as part of Collaborative Research Center "EASE - Everyday activities Science and Engineering" (http://www.ease-crc.org)

Although CNNs featuring a single data source show already impressive performance for activity recognition, the fusion of several modalities could be a promising research direction for a further improvements of performance. Ambiguities of single data sources might be resolved and correlations between date sources could be exploited by integrating different modalities, thus improving the overall system performance.

Here we investigate the potential of multimodal fusion strategies in the context of spatio-temporal convolutional neural networks. We compare different multimodal architectures in relation to the unimodal variants without aiming for state-of-art performance on the dataset.

II. RELATED WORK

A. Data for Activity Recognition

Currently there is a large number of datasets for activity recognition available (for a review see [1]). They cover a wide range of modalities which are suitable for this task. RGB video data is often part of a dataset and recent approaches have focused on this modality, often together with additional modalities like depth maps. This combination of RGB+D is popular due to the frequent use of Microsoft Kinetic for recording of human activity. Since this device is also able to extract skeleton information from the RGB+D signal, skeleton data is often available as a third useful modality.

Every particular modality can provide different information, with its usefulness depending on the task at hand. Video data provide information about color, texture or shape of people and objects as well as about the whole scene in which an activity takes place. While this modality is very rich in information, the variation of illumination, color etc. in a real-world RGB channel makes it tricky to process.

Channels which are largely invariant to these variations might provide useful data which are easier to process. Depth maps, for instance, are rather invariant to illumination changes which might be very helpful for segmentation or for shape extraction. And there are further channels that can be derived from RGB or depth data. Optical flow is usually extracted from RGB and provides information about spatio-temporal changes in the scene. Skeleton data can be extracted from RGB or from depth, but can also be directly recorded with motion capturing. The information about skeleton points is especially

valuable for analyzing the human part of the scene in activity recognition.

The "NTU RGB+D" [2] which will be used in this paper, is a large-scale dataset providing RGB, depth and skeleton data. It offers large number of subjects and classes from different viewpoints.

B. Activity Recognition Models

The analysis of human activities has drawn significant attention, with special interest in action recognition from RGB video. In the last decade the success of CNN-based approaches in image-based classification has led to the application of these methods to video data. Video data can be treated as a series of 2D image, each processed with a 2D-CNN. While this is sufficient to extract the spatial features, the temporal dynamics need to be captured as well. [3] introduced multi-frame optical flow as an input to a 2D-CNN together with RGB frames. This sort of network is basically a 2D image classification CNN which has the advantage of being pre-trainable on Imagenet [4].

In recent years spatio-temporal 3D-CNN were introduced for processing multiple frames directly [5]–[8]. Two stream CNN [9] add optical flow, derived from RGB video, as a second modality to the network. Multistage CNN [10] and structured segment network [11] add the ability to detect actions in untrimmed videos by generating proposals for time slices with actions. The fusion of multiple modalities in CNNs has been investigated by [25] and [26].

Apart from convolutional neural networks, recurrent neural networks offer another way to process video data. Here the temporal dynamics between individual frames are captured by the recurrent structure of the network [12]–[14].

Skeleton data, representing the positions of joints of a human body over time, offer rich information with regard to human activity recognition. CNN-based approaches can treat the x, y, z-position of joints as separate time series and process them with a time convolutional network [15]. Another way is to transform the joint information into a 2D structure and use 2D-CNN for processing. [16] interpret the joint positions as 2D information and color code the temporal dynamics, [17] use one image dimension for coding the spatial structure of joints and the other for the temporal dynamics and [18] project the 3D positions onto four different 2D planes and encode the joint distances in those planes in images.

As for the combination of modalities, the approaches which work well for a certain modality, are not necessarily suitable or working equally well in a multimodal system.

C. Multi-Sensor Fusion

The fusion of multiple data sources is well established in literature. Bellot [19] identifies four gains which the fusion process might achieve:

- "gain in representation": the fused representation reaches a higher level of granularity or abstract level than the initial data sources.
- "gain in certainty": increase in the belief in the fused data.
- "gain in accuracy": the standard deviation on the data improves. Noise and errors are decreased.
- "gain in completeness": addition of new information make the view on the environment more complete.

With regard to the field of activity recognition, an overview of multi-sensor fusion can be found in [20] and [21]. Multi-sensor fusion is used when several sensors are placed in the environment [22], [23] or on the human body (wearable sensors) [24].

III. MULTIMODAL FUSION

A. Multimodal Fusion Strategies

Using multimodal approaches in a machine learning context is typically aimed at an improvement of the overall system performance with regard to recognition power or robustness. The idea is that individual data sources can provide different kinds of information which might resolve ambiguity, improve the overall quality of noisy data, or enable the exploitation of correlations.

One of the challenges of multimodal machine learning [27] lies in the methods for the fusion of the different modalities. The respective approaches can be broadly categorized as early and late fusion [28], depending on the position of the fusion within the processing chain. Hybrid fusion approaches try to combine the properties of the two basic methods [28].

Late fusion [29] is the simplest and most commonly used fusion method. It merges data after a separate full processing in different unimodal streams. The individual modalities can be processed by powerful targeted approaches, tailored to the specific properties of the particular modality. After a full chain of unimodal processing, typically after predicting labels in a recognition task, the results are merged, in the most simple case by summation or averaging. Late fusion has a major drawback which is the very limited potential for the exploitation of cross correlations between the different unimodal data.

Early fusion [29] is more powerful since it merges data sources in the beginning of the processing. Raw data can be fused directly without any pre-processing, but usually certain features are initially extracted. These basically unimodal features are then fused by concatenating the individual data into a joint representation. The unified representation has to make sure that the data is properly aligned, thus being suitable for further joint processing. If the data is properly aligned, cross-correlations between data items may be exploited, thereby providing an opportunity to increase the performance of the system. [25] argue that those fused low-level features might be irrelevant for the task, thus decreasing the fusion power.

Between late and early fusion as the extremes, it is also possible to use a halfway fusion [25] or middle fusion [26]. Here the fusion point somewhere in the middle of the network.

In this paper, we want to investigate whether the fusion of different modalities can provide an advantage as compared

to uni-modal approaches, and whether a more complex early fusion strategy can outperform the standard late-fusion strategy by making use of statistical correlations between the different modalities. We address this issue in the context of human activity recognition. To ensure a meaningful comparison we avoid special solutions but use one state-of-the-art convolutional network architecture (Inception I3D) for all different settings. Furthermore, we perform the tests on a sufficiently large and general dataset (NTU RGB+D). As modalities we consider RGB video, optical flow, and skeleton data.

B. Multimodal Fusion for Convolutional Neural Networks

Convolutional neural networks have reached remarkable success in a variety of applications. Combining multimodal fusion and CNNs thus appears to be a promising direction for future research. In particular, the possible fusion methods described above can be applied to CNNs.

In the case of early fusion, one can combine raw data or early features. Since raw data from different data sources are rarely spatio-temporally aligned due to different resolutions or sampling frequencies, they require a certain amount of pre-processing before being concatenated for processing by a CNN. If one moves one step further in the network and starts fusion at an early features level, the requirement for spatio-temporal alignment remains, but there are several ways how to extract features. The most simple case is to use convolutional units for feature extraction and train them from scratch, or pre-train on a different dataset which offers the same modality. One could also bootstrap those units with weights learned on a similar data source. For instance one can train on Imagenet [4] and initialize a CNN with these pretrained weights. If the dimensionality changes to 3D as with video data, 2D weights can still be used for bootstrapping [9]. A last possibility is to use classical approaches for features extraction, e.g. a filter bank of parameterized filters.

For late fusion several unimodal networks are used as the basis for the fused architecture. The individual networks can be heterogeneous, fitting only the modality they are designed for. The actual fusion is then trivial requiring only the merging of the individual results of each network, i.e. the predicted labels in a recognition task. In order to achieve the fusion, the dense layers which usually form the output layer of a network need to be merged by summing or averaging.

The fusion of raw data resp. first features in the early fusion case and the fusion of final predictions in the late fusion case are the extreme variants. Apart from these two there are many more potential fusion points within a deep CNN. With increasing number of layers, the complexity of individual features typically rises. Fusing such correlated complex features for multiple modalities might result in increased performance.

IV. METHODS

A. Dataset

The reasons for using the "NTU RGB+D" dataset [2] are the size of the dataset (over 56k samples, ca. 40k for training and 16k for validation) and the different modalities (RGB video, depth video, IR video and skeleton data) it offers. There are 60 classes, 40 subjects performing the activities and three different view points. The data have been recorded with a Microsoft Kinect v2. Figure 1 shows sample frames from the dataset.

Fig. 1: Sample frames of the NTU RGB+D dataset [2].

B. Modeling

For our investigation we have used early and late multi-modal convolutional neural networks as well as the respective unimodal CNNs. Our basic architecture uses an established architecture from literature, "Inception-v1 I3D" [9]. This architecture uses "Inception" modules (see Figure 2) which introduce parallel pathways for processing of a given input, concatenating the results of each pathway as an output of the module. These "Inception" blocks are repeated several times with `MaxPooling` operations and down-sampling in between at certain points as shown in Figure 3.

We use RGB video, optical flow based on the RGB and skeleton data in our work. The CNNs have been implemented in TensorFlow [30].

The structure of the early and late fusion systems are shown in Fig. 4. The early fusion variant does not fuse the raw data, but the one of the first convolutional features. Architectures like "Inception I3D" are designed to have a set of convolutional layers before processing the data with a series of "inception modules" (see Figure 2) respectively. The first layers, leading to the first "inception" block, are called the stem of the network. We apply the term early fusion, if the fusion takes place within the stem.

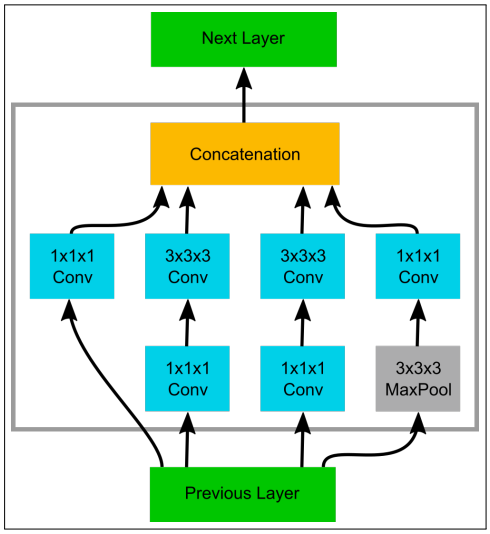


Fig. 2: Structure of an Inception-v1 block (adapted from [9])

The early and mid fusion are performed by concatenating the outputs of the partial networks for the given modalities. For instance, the early fusion of the first convolutional layers (stem layers 1[1]) concatenates the outputs for these layers which have been calculated for each modality, e.g. RGB and Optical Flow, separately. Afterwards the resulting tensor is fed into the remaining network without changing the architecture any further.

In the late fusion variant, the individual modalities are process on their own up to the dense layer at the top which are summed in the end (see Figure 4b).

The fusion of RGB video with Optical Flow is straight forward due to the same dimensionality of these modalities. The processing of skeleton data within the early fusion architecture would require a transformation of the input data, i.e. we transform the 1D skeleton data for a particular time step to 2D. Therefore, we tested the late fusion variant only with skeleton data which does not required any further transformation of the data. Here we can use an existing an existing CNN, build for 1D data, and sum its results with the outputs of the other branch. We use "Res-TCN" [15] for the late fusion variants. It is a residual network with temporal convolutions, thus the convolutions are 1D in nature.

The loss function is sparse softmax cross entropy. We use a Momentum optimizer with 0.9 momentum and a learning rate of 0.001

Based on the RGB videos, optical flow has been computed with "FlowNet2" [31]. The original RGB videos were downscaled to 256x256 and randomly cropped to 224x224 pixels around the center. The optical flow data has been processed in the same manner with the cropping positions being aligned

[1]For "Inception I3D": after 'conv3d_1a_7x7' in the implementation from `https://github.com/deepmind/kinetics-i3d`.

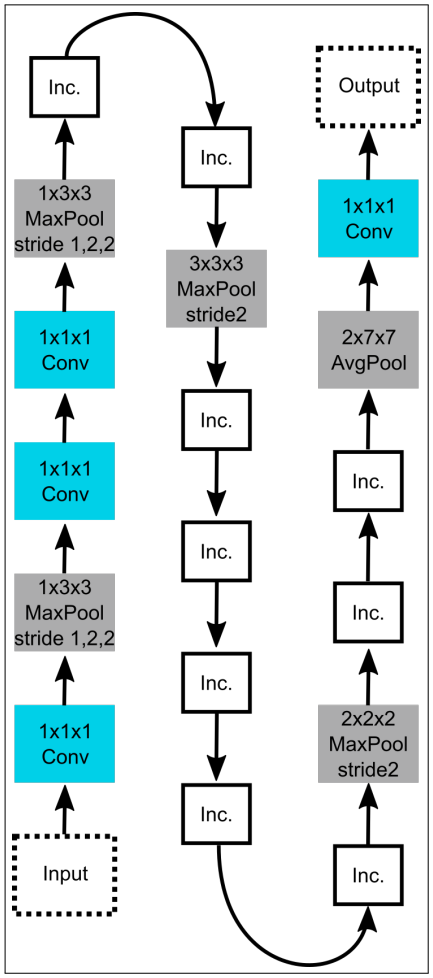


Fig. 3: Layout of the Inception-v1 I3D CNN (adapted from [9])

to the RGB video. The time slices were set to 10 seconds and looped if the data was shorter. The FPS was 25. The skeleton data consists of the x,y,z coordinates for 25 joints per person and was computed with the Kinect v2. There were maximally two people tracked during the recording.

V. RESULTS

We have used the "NTU RGB+D" dataset with the cross-subject split provided by [15] which provides 40320 samples for training. The unimodal and multimodal variants of the network are based on the "Inception I3D" architecture. The results involving skeleton data (unimodal and multimodal late fusion) were obtained with "Res-TCN" [15] for the skeleton part. Figure 5 shows an exemplary plot of the training of a multimodal network. There are no signs indicative of over-training.

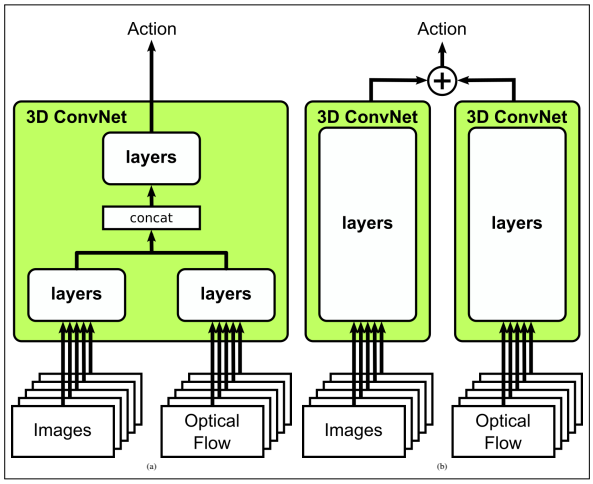


Fig. 4: CNN with (a) early fusion and (b) late fusion. The modalities shown here are RGB images and Optical Flow. For the early fusion the action label is directly output by the logits layer of the fused network. For late fusion the outputs of the logits layers of the individual CNNs for each modality are summed.

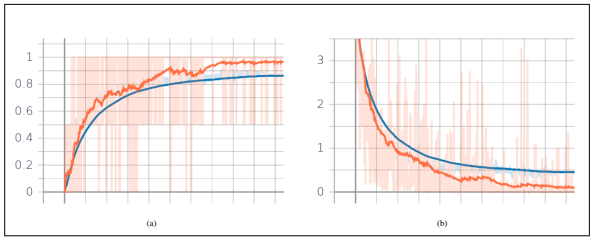


Fig. 5: Accuracy (a) and loss (b) for early fusion with RGB and optical flow with training set (orange) and validation set (blue).

Table I shows a summary of the results in terms of recognition performance. We measure the recognition accuracy by number of correct predictions divided by number of samples. For the unimodal versions of the CNNs classification performance ranges from 66.4% to 78.6%. Optical flow as a single modality provides a relatively poor performance of 66.4%. The other two modalities enable a significantly better recognition, with the best result of 78.6% being obtained on basis of the skeleton data.

The multimodal versions all show an improved performance in comparison to their unimodal counterparts. Somewhat surprising, the smallest improvement to a level of 82.3% is obtained by a late fusion of the RGB channel with the unimodally best performing skeleton channel. Nevertheless, the multimodal performance is superior to that of the individual channels alone (76.8% and 78.6%). The best multimodal performance of 86.7% is obtained with a network which makes use of an early fusion architecture which integrates RGB and optical flow.

TABLE I: Multimodal fusion results of "Kinetics I3D" [9] on NTU Dataset [2] for all results but skeleton data (marked with *) which used "Res-TCN" [15].

	Fusion	**Trained Layers**	**Modalities**	**Accuracy**
Unimodal	RGB video	all layers	RGB	76.8%
	Optical Flow	all layers	Optical Flow	66.4%
	*Skeleton	all layers	Skeleton	78.6%
Multimodal	Early Fusion	all layers	RGB, Op. Flow	86.7%
	Late Fusion	all layers	RGB, Op. Flow	82.9%
	*Late Fusion	last layer	RGB, Skeleton	82.3%

VI. DISCUSSION

In this paper our investigation was focused on the question whether the fusion of information from several data sources is helpful for the task of human activity recognition by convolutional neural networks. Our results show that any sort of fusion will improve the performance. This is valid irrespective of whether the fusion is performed early or late, and irrespective of which modalities are combined.

On a detailed level, our investigations show a clear superiority of an early fusion strategy over a late combination (86.7% for early as opposed to 82.3% and 82.9% for late). This lends support to the hypothesis that a multimodal convolutional network architecture in which the information from different modalities can be combined and recombined across processing stages is able to exploit the multivariate correlational structure of the data sources.

It is interesting to note that in our setting the specific modalities used for the combination seem to have less relevance than the fact that a combination is used at all. Although the unimodal skeleton channel as such yields a much higher performance than unimodal optic flow (78.6% vs. 66.4%), a late fusion of this skeleton channel with the RGB channel cannot provide a better performance than the fusion of the seemingly inferior optic flow channel with the RGB channel (82.3% vs. 82.9%).

Future work can explore several directions for multimodal information fusion with convolutional networks. One direction is the full integration of skeleton data into an early fusion architecture. For this we have to bring the image raster data and the skeleton data into a format suitable for combination. Another direction is to investigate halfway fusion. [25] achieved best results by fusing in the middle of the network. On the other hand [26] reported worse performance for middle fusion under certain conditions.

A further direction is to consider hybrid approaches which try to combine the advantages of both early and late fusion methods. For example one could combine highly specialized processing architectures for unimodal data streams with more

general architectures for early fusion, e.g. for pairwise combination of modalities [32]. These pathways can finally be merged by late fusion, allowing to exploit potential cross-correlations residing in the different data streams, and at the same time use sophisticated models for each data stream.

In conclusion, our results yield further support for the general idea that fusion of and within convolutional network architectures could be a promising research direction for human activity recognition. The greatest potential, in our view, should be sought in an early integration of a variety of information sources.

REFERENCES

[1] P. Wang, W. Li, P. Ogunbona, J. Wan, and S. Escalera, "Rgb-d-based human motion recognition with deep learning: A survey," *CoRR*, vol. abs/1711.08362, 2017. [Online]. Available: http://arxiv.org/abs/1711.08362

[2] A. Shahroudy, J. Liu, T. Ng, and G. Wang, "NTU RGB+D: A large scale dataset for 3d human activity analysis," *CoRR*, vol. abs/1604.02808, 2016. [Online]. Available: http://arxiv.org/abs/1604.02808

[3] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 1*, ser. NIPS'14. Cambridge, MA, USA: MIT Press, 2014, pp. 568–576. [Online]. Available: http://dl.acm.org/citation.cfm?id=2968826.2968890

[4] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and F.-F. Li, "Imagenet: A large-scale hierarchical image database," *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009.

[5] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, ser. ICCV '15. Washington, DC, USA: IEEE Computer Society, 2015, pp. 4489–4497. [Online]. Available: http://dx.doi.org/10.1109/ICCV.2015.510

[6] G. W. Taylor, R. Fergus, Y. LeCun, and C. Bregler, "Convolutional learning of spatio-temporal features," in *Computer Vision – ECCV 2010*, K. Daniilidis, P. Maragos, and N. Paragios, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 140–153.

[7] S. Ji, W. Xu, M. Yang, and K. Yu, "3d convolutional neural networks for human action recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, pp. 221–231, Jan 2013.

[8] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition*, ser. CVPR '14. Washington, DC, USA: IEEE Computer Society, 2014, pp. 1725–1732. [Online]. Available: https://doi.org/10.1109/CVPR.2014.223

[9] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017, pp. 4724–4733.

[10] Z. Shou, D. Wang, and S.-F. Chang, "Temporal action localization in untrimmed videos via multi-stage cnns," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[11] Y. Zhao, Y. Xiong, L. Wang, Z. Wu, X. Tang, and D. Lin, "Temporal action detection with structured segment networks," in *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[12] Y. Shi, Y. Tian, Y. Wang, W. Zeng, and T. Huang, "Learning long-term dependencies for action recognition with a biologically-inspired deep network," in *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[13] D.-A. Huang, L. Fei-Fei, and J. C. Niebles, "Connectionist temporal modeling for weakly supervised action labeling," in *Computer Vision – ECCV 2016*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham: Springer International Publishing, 2016, pp. 137–153.

[14] S. Buch, V. Escorcia, C. Shen, B. Ghanem, and J. Carlos Niebles, "Sst: Single-stream temporal action proposals," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[15] T. S. Kim and A. Reiter, "Interpretable 3d human action analysis with temporal convolutional networks," in *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, July 2017, pp. 1623–1631.

[16] Y. Hou, Z. Li, P. Wang, and W. Li, "Skeleton optical spectra-based action recognition using convolutional neural networks," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 3, pp. 807–811, March 2018.

[17] Y. Du, Y. Fu, and L. Wang, "Skeleton based action recognition with convolutional neural network," in *2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR)*, Nov 2015, pp. 579–583.

[18] C. Li, Y. Hou, P. Wang, and W. Li, "Joint distance maps based action recognition with convolutional neural networks," *IEEE Signal Processing Letters*, vol. 24, no. 5, pp. 624–628, May 2017.

[19] D. Bellot, A. Boyer, and F. Charpillet, "A new definition of qualified gain in a data fusion process: application to telemedicine," in *Proceedings of the Fifth International Conference on Information Fusion. FUSION 2002. (IEEE Cat.No.02EX5997)*, vol. 2, 2002, pp. 865–872 vol.2.

[20] A. A. Aguileta, R. F. Brena, O. Mayora, E. Molino-Minero-Re, and L. A. Trejo, "Multi-sensor fusion for activity recognition—a survey," *Sensors*, vol. 19, no. 17, 2019. [Online]. Available: https://www.mdpi.com/1424-8220/19/17/3808

[21] B. Fu, N. Damer, F. Kirchbuchner, and A. Kuijper, "Sensing technology for human activity recognition: A comprehensive survey," *IEEE Access*, vol. 8, pp. 83 791–83 820, 2020.

[22] M. Tenorth, J. Bandouch, and M. Beetz, "The TUM Kitchen Data Set of Everyday Manipulation Activities for Motion Tracking and Action Recognition," in *IEEE International Workshop on Tracking Humans for the Evaluation of their Motion in Image Sequences (THEMIS), in conjunction with ICCV2009*, 2009.

[23] R. S. Ransing and M. Rajput, "Smart home for elderly care, based on wireless sensor network," in *2015 International Conference on Nascent Technologies in the Engineering Field (ICNTE)*, 2015, pp. 1–5.

[24] C. Zhu and W. Sheng, "Human daily activity recognition in robot-assisted living using multi-sensor fusion," in *2009 IEEE International Conference on Robotics and Automation*, 2009, pp. 2154–2159.

[25] J. Liu, S. Zhang, S. Wang, and D. N. Metaxas, "Multispectral deep neural networks for pedestrian detection," *CoRR*, vol. abs/1611.02644, 2016. [Online]. Available: http://arxiv.org/abs/1611.02644

[26] N. Damer, K. Dimitrov, A. Braun, and A. Kuijper, "On learning joint multi-biometric representations by deep fusion," in *Proceedings of the IEEE 10th International Conference on Biometrics Theory, Applications and Systems (BTAS 2019)*, Tampa, FL, USA, 10 2019.

[27] T. Baltrušaitis, C. Ahuja, and L. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 2, pp. 423–443, Feb 2019.

[28] S. K. D'mello and J. Kory, "A review and meta-analysis of multimodal affect detection systems," *ACM Comput. Surv.*, vol. 47, no. 3, Feb. 2015. [Online]. Available: https://doi.org/10.1145/2682899

[29] C. G. M. Snoek, M. Worring, and A. W. M. Smeulders, "Early versus late fusion in semantic video analysis," in *Proceedings of the 13th Annual ACM International Conference on Multimedia*, ser. MULTIMEDIA '05. New York, NY, USA: Association for Computing Machinery, 2005, p. 399–402. [Online]. Available: https://doi.org/10.1145/1101149.1101236

[30] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, and et al., "Tensorflow: A system for large-scale machine learning," in *Proceedings of the 12th USENIX Conference on Operating Systems Design and Implementation*, ser. OSDI'16. USA: USENIX Association, 2016, p. 265–283.

[31] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox, "Flownet 2.0: Evolution of optical flow estimation with deep networks," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. [Online]. Available: http://lmb.informatik.uni-freiburg.de/Publications/2017/IMSKDB17

[32] Z.-Z. Lan, L. Bao, S.-I. Yu, W. Liu, and A. G. Hauptmann, "Multimedia classification and event detection using double fusion," *Multimedia Tools Appl.*, vol. 71, no. 1, p. 333–347, Jul. 2014. [Online]. Available: https://doi.org/10.1007/s11042-013-1391-2

From Human to Robot Everyday Activity

Celeste Mason[1], Konrad Gadzicki[2], Moritz Meier[1], Florian Ahrens[3], Thorsten Kluss[2], Jaime Maldonado[2], Felix Putze[1], Thorsten Fehr[3], Christoph Zetzsche[2], Manfred Herrmann[3], Kerstin Schill[2], Tanja Schultz[1]

***Abstract*— The Everyday Activities Science and Engineering (EASE) Collaborative Research Consortium's mission to enhance the performance of cognition-enabled robots establishes its foundation in the EASE Human Activities Data Analysis Pipeline. Through collection of diverse human activity information resources, enrichment with contextually relevant annotations, and subsequent multimodal analysis of the combined data sources, the pipeline described will provide a rich resource for robot planning researchers, through incorporation in the OpenEASE cloud platform.**

I. INTRODUCTION

Currently, robots have displayed remarkable feats that would suggest they will soon be able to take over many of our more onerous daily activities (cleaning, cooking, feeding the dog etc), leaving us free to focus our energies elsewhere (eating, petting the dog etc). However, the underlying truth remains – these robots display such sophisticated abilities due to their creators' contextually precise, expertly crafted planning algorithms. For this everyday robotic revolution to occur, these agents will need to be able to react to vague instructions and changing context, in a manner that more closely adheres to human behaviors and abilities. So, how may we identify and, more importantly, collect and describe the missing pieces of this puzzle that would enable cognitive robots to perform actions that approach the aplomb with which humans are able to interact everyday, through habit, common sense, intuition, and problem solving approaches seemingly effortlessly developed throughout their lifetimes?

In this paper we present a novel data processing pipeline for human activity recognition (HAR). To our knowledge, our pipeline is the first to combine multimodal data collection, hierarchical and semantic annotations, and ontological reasoning to enhance cognitive robots with human-like reasoning capabilities derived from everyday human activities. The collaborative research center EASE ("Everyday Science and Engineering," http://ease-crc.org) has facilitation of robotic mastery of everyday activities as its unifying mission. The subprojects concerned with human activities data collection have the goal of providing so-called narrative-enabled episodic memories (NEEMs). These data structures store recorded observations, experiences and activities compiled as a single coherent item. Another goal is the derivation of pragmatic everyday activity manifolds (PEAMs), which will form the basis for robot agent enhancement by enabling real-time interaction similar to how humans function.

[1]Cognitive Systems Lab, University of Bremen, Germany
[2]Cognitive Neuroinformatics, University of Bremen, Germany
[3]Neuropsychology and Behavioral Neurobiology, University of Bremen, Germany

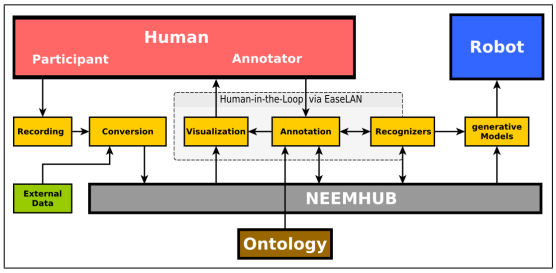


Fig. 1: The EASE human activities data analysis pipeline

We start with recording of human activities in a kitchen setting, preprocessing, and (optionally) supplementing with outside data sources, before storing data in the openEASE robotics knowledge base platform. Annotations, based on the EASE-Ontology, are designed for cognitive robots. The automatic annotators use different modalities and approaches, e.g. multimodal activity recognition, speech recognition or object tracking, thus complementing each other. Data produced from further processing through manual and automatic annotation, and subsequent analyses through a variety of machine learning techniques, can then be queried in openEASE. Based on performance in robotic activity scenarios, the annotation schema can then be improved further. The results–raw and processed multimodal data recordings, together with annotations and data derived from analyses–are stored in openEASE, a framework for knowledge representation and reasoning, as shown in figure 1.

The NEEMs derived from research projects in EASE subproject area H (Human Activities Data Collection) provide unique and critical contextual background for robots, based on human activities, perceptions, and feedback. Analyses of biosignals derived from brain, muscle, or skin signals may provide insight into diverse aspects of human behavior required for humans to masterfully perform everyday activities with little effort or attention. Through the integration of analyses from a wide array of data sources, through a multitude of complementary methods, we endeavor to build an extensive, contextually dense reserve of activity experience and problem solving approach methods derived from human behaviors to transfer effectively to robot systems. The following examples are among those being employed within the EASE-CRC at this time.

Brain activity measurements allow evaluation of attentional focus while performing tasks, adaptation to ambiguous or conflicting situations or physical obstacles, decision

making processes, and how motor imagery when viewing performance of activities compares with in-situ motor execution. Skin and muscle activity sensors can indicate overall mental state, and information about manual manipulation interactions with objects, such as the force used. Full body motion capture provides motion in an environment and object interactions. Assessment of small scale hand movements (including e.g. forces, velocities, trajectories, etc.) using the PHANToM haptic interface Scene video from many perspectives allows tracking of objects and the order of interactions, insight into efficient movement within a space while performing tasks. First person video provides understanding of scene aspects people may focus on while planning and executing tasks. Important information for a robot might include attention (internal vs external) and visual search strategies for objects or positions based on contexts such as meal type, formality, or number of diners. Microphones record scene audio, speech and non-speech vocalizations. Through audio recordings of what a person thinks-aloud while they perform tasks, we gain a rich description of the scene as the performer sees it, obstacles encountered, reasoning and problem solving approaches, frustration or enjoyment, and the task process as a whole.

openEASE [1] is an online knowledge representation and reasoning framework for robots. It provides the infrastructure to store and access nonhomogeneous experience data from robots and humans, and comes with software tools which enable researchers to analyze and visualize the data.

EASE subprojects record human activity datasets (HAD) in a range of scenarios. Data collection efforts focus on contexts involving "Activities of Daily Living" (ADLs) in the kitchen, such as setting and clearing the table or doing dishes. From these experiments, we are producing the multimodal EASE Table Setting Dataset (EASE-TSD), a dataset featuring brain measurements using functional magnetic resonance imaging (fMRI) and electroencephalography (EEG) during table setting related tasks, and the EASE Manipulation Adaptivity Dataset (EASE-MAD), that focuses on sensorimotor regulation during individual (atomic) actions or short sequences, in detail. These experimental contexts and data recordings are described in later sections.

Seamless integration of these datasets, annotations, and derived models into the openEASE framework will provide the solid foundation for robotic researchers to expedite development of robotic agents that are more robust to unexpected variations in task requirements and context, taking human behaviour as inspiration.

II. RELATED WORK

A. Data for Activity Recognition

Activity recognition can be performed on a wide variety of features and a large number of datasets have been provided (for a review see [2]).

Recent approaches in activity recognition often work on RGB-D data. These consist of RGB video and accompanying depth maps and provide two useful modalities for human activity recognition. Skeleton data can be recorded with motion capturing or extracted from RGB-D data as in the case of a Kinect. Other modalities, e.g. optical flow, can be extracted from RGB data as well.

There are several other datasets featuring kitchen related activities. "EPIC-Kitchens" [3] features head-mounted camera video taken during kitchen activities performed by 32 participants in their homes, annotated with 125 verb classes and 352 noun classes in varied languages. "50 Salads" [4] uses 3rd-person camera (including RGB-D cameras) and accelerometer-mounted objects to record meal preparation sequences. "MPII Cooking Activities Dataset" [5] uses video recordings of participants performing 65 kitchen activities for pose-based activity recognition. The "TUM Kitchen Dataset" [6] features video, full-body motion capture, RFID tag readings and magnetic sensor data taken during activities, processed with manual motion tracker labels and automatic semantic segmentation.

B. Activity Recognition Models with Neural Networks

The specific properties of the various modalities have led to different processing strategies. For instance, the RGB channel can be processed with a spatio-temporal convolutional neural network (CNN) [7], [8], [9], [10], either on its own or together with derived optical flow [11], [12] through a two-stream CNN or as a multistage CNN [13], [14]. Another approach is to use recurrent neural networks (RNNs) for processing of RGB data [15], [16], [17], [18], [19]. The depth channel can be similarly processed with a CNN [20], [21] or with a combination of CNN and RNN [22]. With regard to skeleton data, processing with CNN can be enabled by interpreting joint positions as image data [23], [24], [25], [26]. There also exist RNN-based approaches [27], [28], [29], a Deep Boltzmann Machine (DBM) approach [30] and a Hidden Markov Model (HMM) with a deep network as a state probability predictor [31].

C. Semantic, Multimodal Activity Recognition

In [32], semantic hierarchically structured actions are recognized within the kitchen-related context of pancake making, sandwich making, and setting the table in order to transfer task-related skills to humanoid robots. Their ontologically-associated knowledge representations of the observed behavioral data, recorded as video, of people during interactions with objects during such tasks is defined at varying levels of abstraction. Semantic activity recognition of kitchen ADLs (such as making pasta or taking medicine) in the form of multimodal sensor data [33] has also been performed, supported by a Semantic Sensor Network ontology for worn and environment sensor information.

III. DATA RECORDING

A. Human Activities in a Pseudo-natural Setting

The EASE Table Setting Dataset (EASE-TSD) is composed of multimodal biosignal data recorded during experimental observations of various table setting tasks performed by participants in our Biosignals Acquisition Space (EASE-BASE), as described in [34]. All signals are recorded

synchronously using Lab Streaming Layer (LSL) [35]. The recorded sensor modalities include: full-body motion-tracking, audio (from a scene mic and head-worn mic for speech), video from 7 mounted webcams and one head-mounted eyetracker, and biosignals from muscle and brain activity) from participants performing everyday activities while describing their task through use of think-aloud protocols during the task (concurrently) and after the task is completed (retrospectively) [36], as shown in Figure 2a-b. For the EASE-TSD experimental recordings, 70 sessions have been recorded, composed of six or more trials each, totaling 470 concurrent and 405 retrospective think-aloud trial variants. Over 37,400 transcribed words of the think-aloud speech during these trials have been created, with think-aloud encoding annotations underway. Over 16,600 action annotation labels, broken down into between 2 and 12 category sets, for these trials have been performed at varying levels of granularity. Annotations and transcriptions on numerous levels continue to be created as analyses progress. Once the planned recordings with 100 participants are finished and preliminary analysis is completed, the data set will be made available to the public.

Fig. 2: Experimental data recording of participants a) performing concurrent think-aloud trial tasks, b) performing retrospective think-aloud trial tasks c) performing tasks during fMRI, and d) performing tasks during stationary EEG.

B. Human Activities in a Controlled Setting

Table setting videos recorded from the first-person perspective are used in neuroimaging studies, using a 3-Tesla MRI-Scanner as well as high-density multi-channel EEG system, situated in an electromagnetically shielded room. Study participants are tasked with actively imagining themselves acting out the presented situations, thus employing motor imagery [37], while their brain activity is measured.

While fMRI offers unrivaled spatial resolution and the ability to accurately measure whole brain volumes, the residential EEG-System provides high temporal resolution and, in comparison to mobile solutions, offers the advantage of less data contamination caused body movement and electromagnetic emission sources such as cameras, movement detection systems and so forth. Due to its high number of acquisition channels, it also allows for detailed source localization of brain activity.

C. Adaptivity of Human Activities

The Manipulation Adaptivity Dataset (EASE-MAD) includes data from different sources that were assessed in controlled VR settings. This approach allows for deeper insights into the *sensorimotor loop (SML)*, which is a model concept for describing the integration of sensory and motor systems that is the basis of continuous modification of motor commands in response to sensory inputs. Whereas basic sensorimotor loops have been successfully modeled in several variants using control engineering approaches [38], they cannot explain the effortless precision and vast flexibility found in human voluntary actions.

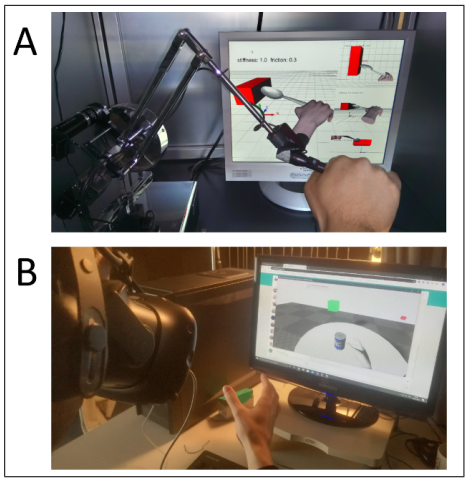


Fig. 3: Data recording during adaptivity testing a) using the PHANToM haptic interface that allows for force rendering to interact with objects and tools in VR, and b) using optical tracking to control a full hand model in VR. The head mounted display is equipped with an eyetracker.

It should be noted, however, that observation of real-world everyday activities only allows us to capture the sensorimotor loop from the outside (i.e. analyze its *outer* PEAMs without being able to directly address its inner laws. We seek more direct access by closing the sensorimotor loop in *virtual reality* (VR), in which we are, unlike in real world experiments, in full control of the individual parameters of the environment and actions. Thus, our main research paradigm will be to intervene in the sensorimotor loop at different points of the control chain [39], [40]). VR as experimental setting enables systematic intervention beyond the physical limitations of real world studies. This allows analysis of how cognitive systems adjust to changing and ambiguous environmental conditions and a systematic modeling of both the inner and outer PEAMs of everyday activities.

Figure 3 shows experimental setups to record multimodal data including e.g. grasping trajectories, hand pose and finger positions during an action, or applied forces (assessed with an PHANToM haptic interface (e.g. [41], [42]). For a detailed description of data acquisition methods for the EASE-MAD

Dataset and the underlying approach to investigate human sensorimotor adaptivity see [43], [44], [45].

IV. DATA ANNOTATION

A. Annotation Schema and Ontology Integration

Annotation and transcription schema developed for the EASE-TSD to describe aspects of everyday activities pertinent to robotic planning algorithm improvement, and are therefore aligned to the EASE Ontology, are used to annotate video and transcribe audio from speech recorded during the EASE-TSD trials. The annotation schema for video-based recordings are hierarchically-structured semantic descriptions of events at increasingly fine-grained levels of detail. The highest level is the task phase (planning, object retrieval, etc.). Below that are specific recurring action types (picking up objects, searching for places to set them on the table). Actions are further broken down into motions (picking = reach, grasp, lift, retract arm) for each hand or other differentiating criteria. When multiple actions occur simultaneously, multiple annotation tiers are required.

B. Annotation Process

Annotation of various modalities is performed in accordance with the requirements for each type of data, first manually then through automated processing. For the EASE-TSD, the annotation and transcription processes are primarily performed in ELAN, as seen in Figure 4. For each trial, annotation or transcription is performed by one person, then checked by a second. As more data is collected and annotated, additional annotations will continually be performed by additional annotators on previously annotated data, then followed by inter-rater reliability scoring.

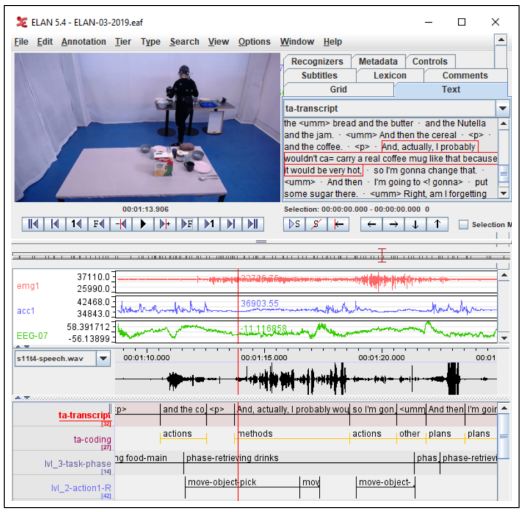


Fig. 4: ELAN is used to create transcriptions and annotations from audiovisual and biosignal data.

Video from multiple angles is used to label time segments where a person is performing specific actions. Frames from these videos are used to obtain information about the number, layout, and positioning of objects within the scene.

Transcription of speech is performed for both concurrent and retrospective think-aloud protocols, with the final goal of transcription by at least two transcribers. These think aloud trials are then coded based on an utterance level schema, to describe the types of thought processes and topics each participant thought relevant to the task at hand at the time.

V. ANALYSIS

A. Human Activity Recognition with Multimodal CNNs

Within our pipeline, we would like to automatically annotate our data. For annotation of video, we used a multimodal fusion approach based on CNNs.

Multimodal fusion is a popular approach to increase the performance of a machine learning system by using several data jointly. It comes in different flavors, with early, late and hybrid fusion being the primary distinctive types [46], with the main difference being where in the processing chain the fusion takes place. All those different types come with different advantages and challenges. The most simple fusion is probably late fusion [47]. Here each modality is processed separately and results are fused afterwards. It allows for maximum flexibility in choosing the processing method for each modality, so that one could use sophisticated unimodal systems (e.g. classifiers) and combine their outputs by i.e. summation, averaging or majority vote. It lacks the potential to exploit possible cross-correlations which may exist between the different data. Early fusion [47] offers a way to exploit those. Here, either raw data or data produced by feature extraction are fused in the beginning of the processing, in the most simple case by concatenation. Afterward the combined data are processed together. This approach requires that the input data are aligned, which might not be trivial when one has to deal with different dimensionality, sampling rate etc. Furthermore there is no choice of specialized approaches for separate modalities; the chosen approach has to fit the joint data.

We have developed a system which allows for the fusion at arbitrary layers. We define a splitting point within the network, up to which the different modalities are processed separately. Afterwards the merged layers are processed by the remaining network. The underlying architecture of our CNN is a “Kinetics I3D” [12] which uses “Inception 3D” units for spatio-temporal processing.

We have used an early fusion CNN approach for this work since in [48] we could show that for activity recognition early fusion performs better than late fusion. For early fusion the individual modalities are processed by the first convolutional layer separately. Its results are fused by concatenation for further processing. Figure 5 shows our architecture for early fusion. Apart from the fusion step, it is a standard “Kinetics I3D” implementation in TensorFlow [49] with sparse softmax cross entropy for loss calculation during training.

Based on the RGB videos, optical flow has been computed with “FlowNet2” [50]. The original full HD RGB videos were rescaled and cropped to 224x224 pixels, the same has

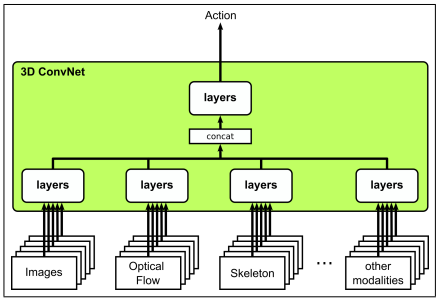


Fig. 5: Early fusion of multiple modalities in a CNN. The modalities are processed up to a specific point in individual paths, then fused by concatenation for further processing.

been done with the optical flow videos based on the RGB videos. See [48] for details of the multimodal network.

The EASE-TSD videos are typically several minutes long. To process them, the data has been chunked into slices during pre-processing. For each slice, we extract the associated ground truth labels which are present during the time slice. Since there can be more than one active label in the time span of a time slice, we have enabled our system to produce multi-label outputs as well as single labels. The multi-label training uses binary cross-entropy as the loss function while the single label variant uses categorical cross-entropy.

Depending on the granularity of the activities within the hierarchy, different time slices might be most useful. For low level actions like `reach` or `pick`, a brief time window of $250ms$ might be sufficient, while a higher level composed action like `pick & place` might not be properly recognized, requiring a longer slice of up to $2s$.

B. Speech Processing

Alongside the human-data table setting recordings, verbal reports of the performed actions and thought processes are recorded. As soon as a new speech recording is present, human transcribers are assigned to a transcription process, which is specified by explicit transcription rules. Automatic speech recognition with pretrained Kaldi acoustic models trained on the GlobalPhone corpus together with a language model enhanced by the previous transcripts is employed to aid the transcription process. A custom ELAN plugin generates a transcript with fine grained segments.

C. Multimodal Biosignal Action Recognition

The analysis of fMRI-data is based on different methodological approaches. Statistical models such as the General Linear Model (GLM) and Independent Component Analysis (ICA) allow contrastive analysis of differences in spatiotemporal patterns of brain activity related to annotated semantic episodes within a perceived point in time during video presentation (e.g. pick up, place, carry), thus leading to detection of distinct neuronal networks correlating with ontological categories. In particular, focus will be on the analysis of the level of neuronal network complexity during planning and execution of complex everyday activities. NEEMs and PEAMs generated from these categorized episodes of brain activity will then be contributed to openEASE.

Building on this knowledge of brain areas that are closely correlated in their activity to ontological categories, further research will also aim at developing algorithms that predict stimuli and semantic episodes on different levels of complexity. Thus, a semi-automatic scene recognition approach will be developed which can feed information into the planned process of automatic activity recognition and its annotation.

Furthermore, the combined use of multi-channel EEG and fMRI allows for a detailed examination of spatiotemporal characteristics of event-related brain activity [46] by using the spatial activation patterns derived from the analysis of the fMRI-data as seed regions for fMRI-constrained source analyses of EEG data [47]. EEG data will be first examined via Fourier analyses (FFT) and band-pass filtered according to oscillatory specificities of ontologically different time periods identified by topographical signal space analyses. Source analyses techniques will then be applied to determine characteristics of the spatial distribution and the spatiotemporal complexity of different periods of table setting action.

For EASE-TSD biosignal-based action recognition, we work toward a model to decode arm movements involved in object manipulation during typical table setting tasks from brain and muscle activity signals, captured by mobile electroencephalography (EEG) and electromyography (EMG) sensors. A subset of 50 EASE-TSD trials performed by 15 participants, manually annotated at the lowest level of arm motion, were used as the basis for multi-class classification using a convolutional and long-short term memory (CLSTM) model on spatially and temporally extracted features derived from EEG and EMG data. This subset of data was recorded using mobile EEG using 16 channels on the scalp as well as 4 EMG sensors each placed on the forearms. After undergoing channel selection, preprocessing, and then statistical and spatiotemporal feature extraction, this became the basis for classification of pick and place activities. For this experimental scenario from the EASE-TSD, we used recordings from experiments performed by 15 right handed participants, 8 male, with age ranging from 20 to 30 as described in [51].

At the lowest level, data from sensors placed at four positions on each arm (e.g., on muscles controlling hand activity of the right forearm) and scalp (e.g., motor regions on the left hemisphere) is used to classify hand movements. EEG data is further filtered to the frequency bands typically corresponding to motor imagery or motor execution–the alpha and beta bands from 8-12 Hz and 12-30 Hz, respectively.

Initially, manually labeled segments such as ‘reach’, ‘grasp’, ‘release’, and ‘retract’ were used for leave-one-out session-independent classification in a supervised manner. To classify these actions using EEG and EMG data, we use a combined CNN/LSTM approach as described in [51]. This analysis will provide the basis for additional custom ELAN recognizer plugin development, to generate activity annotations based on multimodal biosignals.

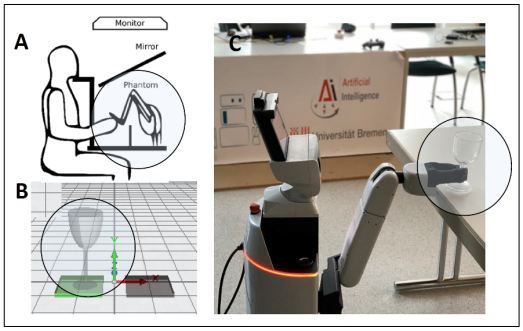


Fig. 6: Transfer of human data to robot control. (A) Data assessment using the PHANToM haptic interface is (B) combined with VR presentation, to (C) control a robot.

VI. Iterative Execution and Improvement through Integration with ELAN and openEASE

The collection of data ultimately serves the purpose of improving the performance of robot systems and enabling them to execute certain actions within the given context and environment. To make the collected data available to the openEASE robotics platform, the detected activities and objects must be translated into a format known to the robot–high-level action plans, such as `move to position Z` or `place object X on surface Y`.

As depicted in figure 6, a pilot demonstration has shown that human data from the EASE-MAD can be successfully transferred to robot actions. In this use case, the task was to place a delicate object, such as a fragile wine glass, on a table. The data were assessed in VR using the PHANToM haptic device in order to present the subjects with realistically rendered forces during placing actions along with the visual sensory feedback. The idea was to transfer the skill of a fast movement with force control to the robot. The resulting end effector variables were suitable to enable the robot to perform the action in an appropriate fashion, i.e. in a real world application it would have been able to lift and to place the glass without breaking it. This approach only comprises a limited range of variables for a short sequence of actions. More complex plans, even though rather abstract in nature, can be executed by the CRAM framework [52].

VII. Results

A. Human Activity Recognition with Multimodal CNNs

We could show that for activity recognition an early fusion approach is better suited than the classic late fusion [48]. For the evaluation in this work, we have used an early fusion architecture with RGB video and optical flow as modalities. We have evaluated the performance on a cross-subject split of the EASE Table Setting Dataset where one recording session (session 17) featuring a specific subject was used as the validation set while seven other sessions featuring other subjects were used as training set. The time slice was set to $0.53s$ (16 frames with 30 fps) for both training and evaluation. With this setting there 29872 data items for training and 8107 for validation. We have achieved an accuracy of 87.8% for the multi-label task and 80.6% for the single-label task on the validation set. Figure 7 shows a plot from one of the training sessions.

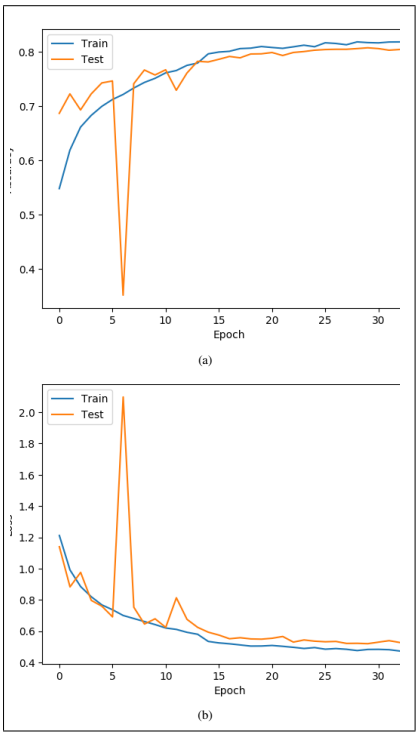


Fig. 7: (a) Accuracy and (b) loss plot for training (blue) and validation (orange) set.

B. Multimodal Biosignal Action Recognition

Brain activity of 30 participants was measured in an EEG-study, consisting of four 1st-person Videos. The videos were annotated according to EASE-ontology, resulting in 312 distinct episodes of various categories and complexity levels. An fMRI-study with 30 participants consisting of ten 1st-person videos was recently finished, with an overall number of 1461 annotated episodes. Preliminary ICA results from this study point out brain areas that discriminate between object interaction events and episodes of no object interaction during the presentation of the videos, as illustrated exemplarily in figure 8. These will later serve as seed regions for the analysis of EEG data.

For the person-independent multi-class motion classification of EASE-TSD trials using a convolutional and long-short term memory (CLSTM) model on EEG and EMG data, the results indicate that EMG features alone provided a better basis for classification at this level of activity. While the low level segments were too brief to extract meaningful information from the EEG sensor data, classification performance

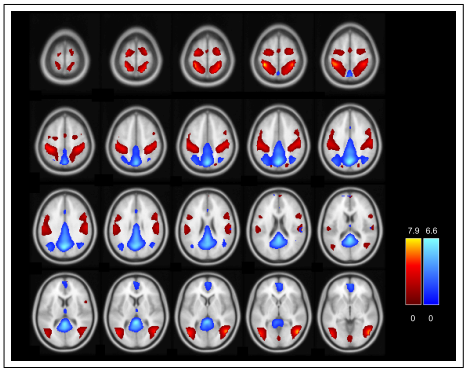


Fig. 8: Brain areas susceptible to stimuli of object interaction events (red) and events with no discernible interaction (blue) during the presentation of a table setting video.

on features derived from EMG sensor data reached 59% accuracy for the right hand movements (MR) and 61.3% accuracy for the left hand movements (ML). Precision for ML was 0.97 vs 1.00 for MR features, recall for ML was 0.95 vs 0.92 for MR, and f1-scores for ML were 0.97 vs 0.95 for MR. Confusion matrices for all combined ML and MR runs are shown in figure 9.

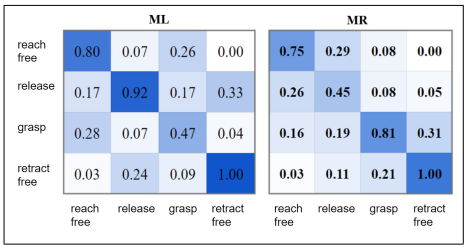


Fig. 9: Confusion matrices for classifications of motions performed with the left and right hand for 4 classes.

VIII. CONCLUSION

Through large-scale collection of human activities of daily living data, annotation with contextually relevant and ontologically linked labeling schema, analysis with diverse multimodal methods for a wide range of sensor modalities, and ultimately, incorporation into the OpenEASE robotics cloud platform, the EASE human activities data analysis pipeline provides the rich groundwork on which to build cognitively-enhanced robotics for use in everyday scenarios.

ACKNOWLEDGMENT

The research reported in this paper has been supported by the German Research Foundation DFG, as part of Collaborative Research Center (Sonderforschungsbereich) 1320 "EASE - Everyday Activity Science and Engineering", University of Bremen (http://www.ease-crc.org/). The research was conducted in sub-projects H01 *Acquiring activity models by situating people in virtual environments* and H03 *Descriptive models of human everyday activity*.

REFERENCES

[1] M. Beetz, M. Tenorth, and J. Winkler, "Open-EASE – a knowledge processing service for robots and robotics/ai researchers," in *IEEE International Conference on Robotics and Automation (ICRA)*, Seattle, Washington, USA, 2015, finalist for the Best Cognitive Robotics Paper Award.

[2] P. Wang, W. Li, P. Ogunbona, J. Wan, and S. Escalera, "Rgb-d-based human motion recognition with deep learning: A survey," *CoRR*, vol. abs/1711.08362, 2017. [Online]. Available: http://arxiv.org/abs/1711.08362

[3] D. Damen, H. Doughty, G. M. Farinella, S. Fidler, A. Furnari, E. Kazakos, D. Moltisanti, J. Munro, T. Perrett, W. Price, and M. Wray, "Scaling egocentric vision: The EPIC-KITCHENS dataset," *CoRR*, vol. abs/1804.02748, 2018. [Online]. Available: http://arxiv.org/abs/1804.02748

[4] S. Stein and S. J. McKenna, "Combining embedded accelerometers with computer vision for recognizing food preparation activities," in *Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, ser. UbiComp '13. New York, NY, USA: ACM, 2013, pp. 729–738. [Online]. Available: http://doi.acm.org/10.1145/2493432.2493482

[5] M. Rohrbach, S. Amin, M. Andriluka, and B. Schiele, "A database for fine grained activity detection of cooking activities," 06 2012, pp. 1194–1201.

[6] M. Tenorth, J. Bandouch, and M. Beetz, "The tum kitchen data set of everyday manipulation activities for motion tracking and action recognition," in *Computer Vision Workshops (ICCV Workshops), 2009 IEEE 12th International Conference on*. IEEE, 2009, pp. 1089–1096.

[7] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, ser. ICCV '15. Washington, DC, USA: IEEE Computer Society, 2015, pp. 4489–4497. [Online]. Available: http://dx.doi.org/10.1109/ICCV.2015.510

[8] G. W. Taylor, R. Fergus, Y. LeCun, and C. Bregler, "Convolutional learning of spatio-temporal features," in *Computer Vision – ECCV 2010*, K. Daniilidis, P. Maragos, and N. Paragios, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 140–153.

[9] S. Ji, W. Xu, M. Yang, and K. Yu, "3d convolutional neural networks for human action recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, pp. 221–231, Jan 2013.

[10] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition*, ser. CVPR '14. Washington, DC, USA: IEEE Computer Society, 2014, pp. 1725–1732. [Online]. Available: https://doi.org/10.1109/CVPR.2014.223

[11] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 1*, ser. NIPS'14. Cambridge, MA, USA: MIT Press, 2014, pp. 568–576. [Online]. Available: http://dl.acm.org/citation.cfm?id=2968826.2968890

[12] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017, pp. 4724–4733.

[13] Z. Shou, D. Wang, and S.-F. Chang, "Temporal action localization in untrimmed videos via multi-stage cnns," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[14] Y. Zhao, Y. Xiong, L. Wang, Z. Wu, X. Tang, and D. Lin, "Temporal action detection with structured segment networks," in *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[15] Y. Shi, Y. Tian, Y. Wang, W. Zeng, and T. Huang, "Learning long-term dependencies for action recognition with a biologically-inspired deep network," in *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[16] D.-A. Huang, L. Fei-Fei, and J. C. Niebles, "Connectionist temporal modeling for weakly supervised action labeling," in *Computer Vision – ECCV 2016*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham: Springer International Publishing, 2016, pp. 137–153.

[17] S. Buch, V. Escorcia, C. Shen, B. Ghanem, and J. Carlos Niebles, "Sst: Single-stream temporal action proposals," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[18] J. Donahue, L. A. Hendricks, M. Rohrbach, S. Venugopalan, S. Guadarrama, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 677–691, Apr. 2017. [Online]. Available: https://doi.org/10.1109/TPAMI.2016.2599174

[19] J. Yue-Hei Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici, "Beyond short snippets: Deep networks for video classification," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[20] P. Wang, W. Li, Z. Gao, J. Zhang, C. Tang, and P. Ogunbona, "Action recognition from depth maps using deep convolutional neural networks," *IEEE Transactions on Human-Machine Systems*, vol. 46, pp. 498–509, 2016.

[21] H. Rahmani and A. Mian, "3d action recognition from novel viewpoints," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Los Alamitos, CA, USA: IEEE Computer Society, jun 2016, pp. 1506–1515. [Online]. Available: https://doi.ieeecomputersociety.org/10.1109/CVPR.2016.167

[22] Z. Shi and T.-K. Kim, "Learning and refining of privileged information-based rnns for action recognition from depth sequences," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[23] T. S. Kim and A. Reiter, "Interpretable 3d human action analysis with temporal convolutional networks," in *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, July 2017, pp. 1623–1631.

[24] Y. Du, Y. Fu, and L. Wang, "Skeleton based action recognition with convolutional neural network," in *2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR)*, Nov 2015, pp. 579–583.

[25] Y. Hou, Z. Li, P. Wang, and W. Li, "Skeleton optical spectra-based action recognition using convolutional neural networks," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 3, pp. 807–811, March 2018.

[26] C. Li, Y. Hou, P. Wang, and W. Li, "Joint distance maps based action recognition with convolutional neural networks," *IEEE Signal Processing Letters*, vol. 24, no. 5, pp. 624–628, May 2017.

[27] Y. Du, W. Wang, and L. Wang, "Hierarchical recurrent neural network for skeleton based action recognition," *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1110–1118, 2015.

[28] Y. Du, Y. Fu, and L. Wang, "Representation learning of temporal dynamics for skeleton-based action recognition," *IEEE Transactions on Image Processing*, vol. 25, pp. 3010–3022, 2016.

[29] V. Veeriah, N. Zhuang, and G.-J. Qi, "Differential recurrent neural networks for action recognition," in *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, ser. ICCV '15. Washington, DC, USA: IEEE Computer Society, 2015, pp. 4041–4049. [Online]. Available: http://dx.doi.org/10.1109/ICCV.2015.460

[30] R. Salakhutdinov, J. B. Tenenbaum, and A. Torralba, "Learning with hierarchical-deep models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1958–1971, 2013. [Online]. Available: https://doi.org/10.1109/TPAMI.2012.269

[31] D. Wu and L. Shao, "Leveraging hierarchical parametric networks for skeletal joints based action segmentation and recognition," in *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition*, ser. CVPR '14. Washington, DC, USA: IEEE Computer Society, 2014, pp. 724–731. [Online]. Available: https://doi.org/10.1109/CVPR.2014.98

[32] K. Ramirez-Amaro, M. Beetz, and G. Cheng, "Transferring skills to humanoid robots by extracting semantic representations from observations of human activities," *Artificial Intelligence*, vol. 247, pp. 95 – 118, 2017, special Issue on AI and Robotics. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0004370215001320

[33] D. Triboan, L. Chen, F. Chen, and Z. Wang, "A semantics-based approach to sensor data segmentation in real-time activity recognition," *Future Generation Computer Systems*, vol. 93, pp. 224 – 236, 2019. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167739X18303947

[34] C. Mason, M. Meier, F. Ahrens, T. Fehr, M. Herrmann, F. Putze, and T. Schultz, "Human activities data collection and labeling using a think-aloud protocol in a table setting scenario," in *IROS 2018: Workshop on Latest Advances in Big Activity Data Sources for Robotics & New Challenges, Madrid, Spain*, 2018.

[35] M. Meier, C. Mason, F. Putze, and T. Schultz, "Comparative analysis of think-aloud methods for everyday activities in the context of cognitive robotics," *20th Annual Conference of the International Speech Communication Association*, vol. 9, p. 10, 2019.

[36] K. A. Ericsson and A. S. Herbert, "Verbal reports as data," vol. 87, no. 3, pp. 215–251, 1980.

[37] M. Jeannerod, "Mental imagery in the motor context," *Neuropsychologia*, vol. 33, no. 11, pp. 1419–1432, nov 1995. [Online]. Available: https://doi.org/10.1016/0028-3932(95)00073-C

[38] R. Der, G. Martius, and R. Pfeifer, *The Playful Machine: Theoretical Foundation and Practical Realization of Self-Organizing Robots*. Springer Science & Business Media, 2012, vol. 15.

[39] C. Zetzsche, J. Wolter, C. Galbraith, and K. Schill, "Representation of space: Image-like or sensorimotor?" *Spatial Vision*, vol. 22, no. 5, pp. 409–424, 2009.

[40] T. Kluss, N. Schult, K. Schill, C. Zetzsche, and M. Fahle, "Spatial alignment of the senses: The role of audition in eye-hand-coordination," *i-Perception*, vol. 2, no. 8, pp. 939–939, 2011.

[41] T. H. Massie, J. K. Salisbury, *et al.*, "The phantom haptic interface: A device for probing virtual objects," in *Proceedings of the ASME winter annual meeting, symposium on haptic interfaces for virtual environment and teleoperator systems*, vol. 55, no. 1. Chicago, IL, 1994, pp. 295–300.

[42] M. C. Çavuşoğlu, D. Feygin, and F. Tendick, "A critical study of the mechanical and electrical properties of the phantom haptic interface and improvements for highperformance control," *Presence: Teleoperators & Virtual Environments*, vol. 11, no. 6, pp. 555–568, 2002.

[43] J. L. Maldonado Cañon, T. Kluss, and C. Zetzsche, "Adaptivity of end effector motor control under different sensory conditions: Experiments with humans in virtual reality and robotic applications," *Frontiers in Robotics and AI*, vol. 6, p. 63, 2019.

[44] J. L. Maldonado Cañon, T. Kluss, and C. Zetzsche, "Pre-contact kinematic features for the categorization of contact events as intended or unintended," in *2020 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, 2020, pp. 764–765.

[45] J. L. Maldonado Cañon, T. Kluss, and C. Zetzsche, "Categorization of contact events as intended or unintended using pre-contact kinematic features," in *2020 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, 2020, pp. 46–51.

[46] S. K. D'mello and J. Kory, "A review and meta-analysis of multimodal affect detection systems," *ACM Comput. Surv.*, vol. 47, no. 3, Feb. 2015. [Online]. Available: https://doi.org/10.1145/2682899

[47] C. G. M. Snoek, M. Worring, and A. W. M. Smeulders, "Early versus late fusion in semantic video analysis," in *Proceedings of the 13th Annual ACM International Conference on Multimedia*, ser. MULTIMEDIA '05. New York, NY, USA: Association for Computing Machinery, 2005, p. 399–402. [Online]. Available: https://doi.org/10.1145/1101149.1101236

[48] K. Gadzicki, R. Khamsehashari, and C. Zetzsche, "Early vs late fusion in multimodal convolutional neural networks," in *Proceedings of the 23rd International Conference on Information Fusion*. IEEE, July 2020.

[49] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, and et al., "Tensorflow: A system for large-scale machine learning," in *Proceedings of the 12th USENIX Conference on Operating Systems Design and Implementation*, ser. OSDI'16. USA: USENIX Association, 2016, p. 265–283.

[50] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox, "Flownet 2.0: Evolution of optical flow estimation with deep networks," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. [Online]. Available: http://lmb.informatik.uni-freiburg.de/Publications/2017/IMSKDB17

[51] A. Kondinska, "Comparison and classification of multimodal biosignals during and prior to hand motor executions involved in activities of daily living," M.S. thesis, University of Bremen, Germany, 2020.

[52] M. Beetz, L. Mösenlechner, and M. Tenorth, "CRAM – A Cognitive Robot Abstract Machine for Everyday Manipulation in Human Environments," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, Taipei, Taiwan, October 18-22 2010, pp. 1012–1017.

Hierarchical Clustering of Sensorimotor Features

Konrad Gadzicki

University of Bremen
28359 Bremen, Germany
konny@informatik.uni-bremen.de

Abstract In this paper a method for clustering patterns represented by sets of sensorimotor features is introduced. Sensorimotor features as a biologically inspired representation have proofed to be working for the recognition task, but a method for unsupervised learning of classes from a set of patterns has been missing yet. By utilization of Self-Organizing Maps as a intermediate step, a hierarchy can be build with standard agglomerative clustering methods.

1 Motivation

The task of unsupervised discovering of meaningfull structures in data sets – formally known as clustering – is probably among the most covered in computer science. Without having any (or much) knowledge about the underlying structure of the data, one hopes that partition of class can be extracted from the similarity of patterns.

The goal of this work is investigate in how far semantic information can be found in pattern represented by sensorimotor features. Sensorimotor features are defined as two distinct points associated with some sensory data and connected by some relation. In the scope of this work, it means a saccade-like representation, but it is also possible to use it for instance in the spatial domain with two locations in space being connected by some motor actions of an agent[1,2].

The target output is a hierarchy of classes. The goal is thus not only to partition the space meaningfully but also to obtain a memory structure for further recognition tasks. The usage of a hierarchical memory is not only usefull from a computational point of view, but also agrees with cognitive memory structures. Psychological experiments give evidence for the existence of hierarchical propositional networks [3], sequence planning and execution [4], cognitive maps [5], memorizing of sequences of symbols [6] or hierarchical mental imagery[7].

In this work, the clustering of sensorimotor features will be performed on image sets.

The following section 2 will explain how images are represented by sensorimotor features and what those features actually consists of. Furthermore, Self-Organizing Maps and Agglomerative Hierarchical Clustering will be explained briefly in general since both are used in the overall clustering task. Section 3 presents the actual approach to clustering sensorimotor features and section 4 shows the results.

2 Methods

2.1 Image Representation

Images are represented by a set of "eye movements" [8], mimicking the biological way of recognition of views by performing saccades. The *foveal spot* in the human eye – the only part with a high optical resolution – covers only a very narrow part of the view. Still humans are able to perceive the environment in detail by performing rapid eye movements, thus shifting the fixation location of the *fovea* around.

The data structure used to store an "eye movement" is called a sensorimotor feature. It consists of a triple $< feature, action, feature > (FAF)$ where the *features* contain some sensory data at an image location and the *action* stores the relative position change from the first to the second feature. The features are locally limited, resembling the narrow angle of the view field covered by the *foveal spot* during a fixation. The action corresponds with the relative shift of gaze.

Feature Extraction and Representation. The extraction of interesting fixation location is based on the concept of intrinsic dimensionality. This concept relates the degrees of freedom provided by a domain to the degrees of freedom actually used by a signal and states that the least redundant (the most informative) features in images are intrinsically two-dimensional signals (i2D-signals) [9,10].

For the actual extraction, nonlinear i2D-selective operators are applied to the image to find the fixation locations. Afterwards the feature vectors describing the local characteristics are generated by combining the outputs of linear orientation selective filters [8]. As a result, the foveal feature data structure stores the local opening angle, the orientation of the angle opening with respect to the image and the color.

The action structure stores the relation between two foveal features, containing the distance, the difference angle (the difference between the opening angles of those features) and the relation angle.

An image as a whole is represented as a set of such sensorimotor features. For a given number of fixation locations, each pair of *foveal features* can be connected by two *actions* since a sensorimotor feature is directional. In terms of graph theory this makes a fully connected directional graph with a cardinality of $|E| = |V| \cdot (|V| - 1)$ where an edge $|E|$ is an action and a vertex $|V|$ is a fixation location. The number of sensorimotor features in total is equivalent to the number of edges.[1]

2.2 Self-Organizing Maps

A Self-Organizing Map (SOM)[11] is a competitive neural network, developed by Teuvo Kohonen in the 1980s [12].

[1] In practice, an 512x512 image has roughly 40–50 extracted foveal features.

A SOM realizes a mapping from a higher-dimensional input space to a two-dimensional grid while preserving the original topological information of the input space. The inspiration for these networks comes from topological structures in the human brain which are spatially organized according to the sensory input [13] (see [14,15] for an overview of research results).

Working Principle. The basic principle of the SOM is to adopt a neuron and its local area in order to make it fit better to a specific input, thus specializing individual nodes to particular inputs. The map as a whole converges to a state where certain areas of the map are specialized to certain parts of input space, so that the topological relations of the input space are preserved in the output space.

The training algorithm for the map is a iterative process during which the best-matching-unit (neuron with minimum distance to current input pattern) is found first. Afterwards the unit and its surrounding neurons are adapted to the current input by changing the weight vector of a neuron according to (1)

$$\mathbf{m}_n(t+1) = \mathbf{m}_n(t) + \alpha(t)\ h_{c_n}[\mathbf{x}(t) - \mathbf{m}_n(t)] \tag{1}$$

where t denotes time. $\mathbf{x}(t)$ is an input vector at time t, the neighborhood kernel around the best-matching unit c is given as h_{c_n} and finally, the learning rate $\alpha(t)$ at a specific time.

2.3 Hierarchical Clustering

Classically clustering algorithms can be divided into hierarchical and partitioning ones. While the partitioning ones produce one, flat set of partitions, hierarchical construct nested partitions, resulting in a dendrogram. In a dendrogram each node represents a cluster; the original patterns are the leafs of the tree and thus singleton clusters.

Agglomerative hierarchical algorithms start with each pattern in a singleton cluster and merge them iteratively until only one cluster is left. The merging process is basically driven by linkage rules which define which two clusters will be merged in each step, and the similarity measure which is calculated between individual patterns that populate clusters. Both, linkage and similarity have a high impact on cluster quality.

Similarity Measures. Similarity is expressed within the range $[0 \ldots 1]$ where 1 states equality and 0 nothing in common at all.

Among the possible similarity measures, *Euclidean distance*-based measures seem to have the highest popularity. It can be derived from the more general L_p-norm (or *Minkowski* norm)

$$d_p(x, y) = \left(\sum_{i=1}^{n} |x_i - y_i|^p \right)^{\frac{1}{p}} \quad . \tag{2}$$

With $p = 2$ the *Minkowski* distance results in *Euclidean*.

The cosine measure is especially popular in document clustering. The similarity is expressed by the cosine of the angle between two vectors and measures similarity directly. It is given by

$$s_{cosine}(x_i, x_j) = \frac{x_i \cdot x_j}{\|x_i\|_2 \cdot \|x_j\|_2} \tag{3}$$

where $x_i \cdot x_j$ is the *dot*-product and $\|x\|_2$ is the Euclidean norm (p-norm with $p = 2$) of the vector.

The *Tanimoto* similarity captures the degree of overlap between two sets and is computed as the number of shared attributes.

$$s_{Tanimoto}(x_i, x_j) = \frac{x_i \cdot x_j}{\|x_i\|^2 + \|x_j\|^2 - x_i \cdot x_j} \tag{4}$$

with $x_i \cdot x_j$ being the *dot*-product and $\|x\|_2$ is the *Euclidean* norm of the vector. This measure is also referred to as *Tanimoto coefficient* (TC).

[16] gives a comparison about the behavior of the three above-mentioned measures.

Linkage Rules. Out of the number of available linkage rules, single and complete linkage are two extremes. Single linkage defines the distance between two clusters as the minimum distance between the patterns of these clusters. In contrast complete linkage takes the maximum distance between patterns. Both produce clusters of different shape: single linkage produces chain-like clusters which is usefull for filamentary data sets; complete linkage produces sphere-like clusters which works well with compact data.

In practice data sets often are not structured in a way suitable for the beforementioned linkage rules. Average based linkage rules incorporate all patterns from a pair of clusters in the distance calculation. There are several variants like "Average Linkage Between Groups" which takes the average distance between all pairs of patterns or "Average Linkage Within Group" which takes the average distance between all possible pairs of patterns if they formed a single cluster. "Ward's method"[17] aims at minimizing the intra-cluster variance by forming a hypothetical cluster and summing the squares of the within-cluster distances to a centroid. The average linkage rules, especially "Ward's" linkage work pretty robust on arbitrary data.

3 Hierarchical Image Clustering

3.1 Comparison of Sensorimotor Features and Images

The problem of comparing pairs of sensorimotor features brings up some problems. In [8], where the task was image recognition, those features were treated as symbols with a particular *FAF* giving evidence for a set of scenes. The initial

sensory measurements – angles, distances and colors – are numerical values which can be measured rather precisely. Still 1:1 comparison of such values turned out to be not very robust to slight transformations. The mapping of the original values to intervals leads to a more robust representation with slightly different values being mapped to the same interval.

The calculation of similarity between images with features treated as symbols allows only the usage of usage of similarity measures based on set operations. Initial tests with such similarity measures produced only mediocre results and led to the idea of a numerical representation. Similar to approaches from document clustering [18] a vector representation for an image was used. Such a representation allows for the usage of numerical similarity measures like those mentioned above.

3.2 Obtaining a Numerical Image Representation.

For obtaining the numerical representation, firstly the sensorimotor features were extracted from a given set of images. A Self-Organizing Map was then trained with the entire set of features. That way the system learned how to group sets of features for a given output size (the map size) which corresponds with the number of components of the vector used for image representation.

The actual representation for a particular image is obtained by presenting the sensorimotor features associated with the image to the SOM. For each feature, a particular map node will be activated, being the Best-Matching-Unit. By counting the activations of map nodes, a histogram is generated for each image which serves as the vector for similarity measurement (fig. 1).

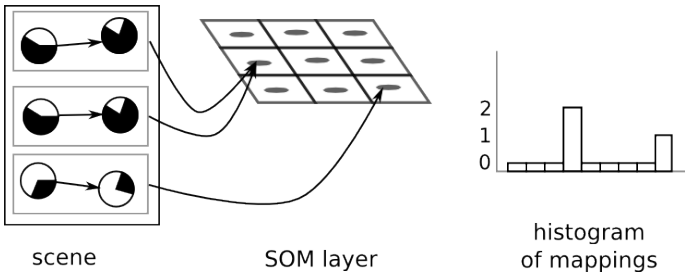


Figure 1. From sensorimotor representation of image to histogram. The features representing an image are mapped to a previously trained SOM. A histogram representation is obtained by counting those mappings.

3.3 Hierarchical Clustering and Quality Assessment.

The actual generation of the hierarchy was with agglomerative hierarchical clustering (see 2.3), with all combinations of the five linkage rules and the three similarity measures.

For the assessment of the quality of the produced hierarchy, the *f-measure* is a suitable measure[19]. It is computed from two other measures, *precision* and *recall*. *Precision* states the fraction of a cluster that belongs to a particular class and *recall* expresses the fraction of objects from a particular class found in a particular cluster out of all objects of that class.

The *f-measure* is a combination of precision and recall. It states to which extend a cluster x contains only and all objects of class c and is given by

$$F(x,c) = \frac{2 \cdot precision(x,c) \cdot recall(x,c)}{precision(x,c) + recall(x,c)} \,. \tag{5}$$

The values produced by the measure are in the range of $[0 \ldots 1]$. A value of 1 means that a cluster is entirely populated by objects from one class and no other. If computed for a hierarchy the *f-measure* is computed for every node of the tree for a given class and the maximum is returned.

The interpretation of a value of 1 is that there is a particular subtree of the hierarchy with objects of class c only and no other. The overall *f-measure* of a hierarchy is then basically the sum of weighted *f-measures* for all classes

$$F = \sum_c \frac{n_c}{n} F(c) = \sum_c \frac{n_c}{n} \max_x (F(x,c)) \tag{6}$$

with n_c being the number of objects in cluster c and n is total number objects.

4 Results and Discussion

The performance has been tested with the "Columbia Object Image Library" (COIL-20)[20]. This image database consists of 20 objects photographed under stable conditions from 72 different perspectives, each view rotated by 5 °.

The initial tests were performed on small subsets of the database in order to see whether the system is able at all to discover semantic information. Objects were selected randomly and from the selected objects, views were picked randomly as well.

Table 1 shows that a *f-measure* value of 1.000 can be achieved with certain linkage rules and similarity measures which means that image were separated according to their inherent class.

Figure 2 shows the hierarchy generated

When tested with the full COIL-20 set, consisting of 1440 image, the figures drop significantly. With *Tanimoto* similarity measure and *Ward's* linkage rule, a *f-measure* of 0.358 can be obtained. Inspecting the generated hierarchy shows that the rough similarities have been captured. For instance, round-shaped patterns populate a particular subtree, but patterns from different classes are mixed.

Based on the results above, you can say that this method is able to capture semantic information in small scale, but on a large scale, the separation does not work well enough.

Table 1. *F-measure* for small COIL-20 subset for combinations of linkage rules and similarity measures

	Euclidean	Cosine	Tanimoto
Single	0.656	0.794	0.861
Complete	0.675	0.905	0.915
Average between groups	0.656	0.915	1.000
Average within group	0.656	0.915	0.915
Ward	0.675	1.000	1.000

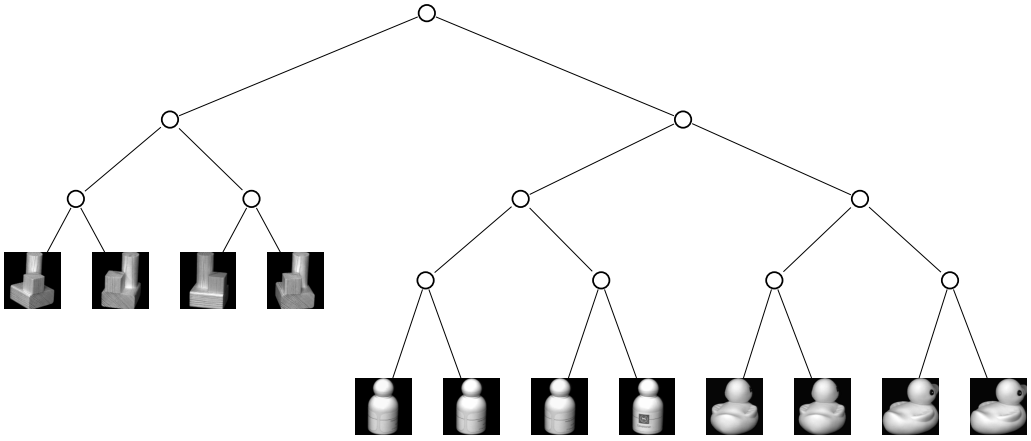


Figure 2. Generated hierarchy with Tanimoto distance and Ward's linkage.

5 Summary

The hierarchical clustering approach introduced works on patterns represented by sets of sensorimotor features. By mapping the sensorimotor features to a Self-Organizing Map, a fixed-size vector representation of patterns is produced. The mapping of the set of sensorimotor features of a particular pattern to the output layer of the SOM generates a histogram of SOM activations which serves as a representation of the pattern. Being a numerical representation, it can be further processed with standard agglomerative hierarchical clustering methods in order to produced the hierarchy.

Acknowledgments

Supported by SFB.

References

1. C. Zetzsche, J. Wolter and K. Schill: Sensorimotor representation and knowledge-based reasoning for spatial exploration and localisation. Cognitive Processing 9, 283–297 (2008)

2. T. Reineking: Active Vision-based Localization using Dempster-Shafer Theory. Mastersthesis, University of Bremen (2008)
3. A.M. Collins and M.R. Quillian: Retrieval Time from Semantic Memory. In: *Journal of Verbal Learning and Verbal Behaviour*, 8, 240–247 (1969)
4. R. Oesterreich: Handlungsregulation und Kontrolle. München, Urban & Schwarzberger (1981)
5. A. Stevens and P. Coupe: Distortions in Judged Spatial Relations. In: *Cognitive Psychology* 10, 422–437 (1978)
6. N.F. Johnson: The Role of Chunking and Organization in the Process of Recall. In: G. Bower, editor, *Psychology of Language and Motivation* 4 (1970)
7. S.K. Reed: Structural Descriptions and the Limitations of Visual Images. In: *Memory and Cognition* 2, 329–336 (1974)
8. K. Schill, E. Umkehrer, S. Beinlich, G. Krieger and C. Zetzsche: Scene analysis with saccadic eye movements: Top-down and bottom-up modeling. Journal of Electronic Imaging 10, 152–160 (2001)
9. C. Zetzsche and U. Nuding: Nonlinear and higher-order approaches to the encoding of natural scenes. Network: Computation in Neural Systems 16, 191–221 (2005)
10. C. Zetzsche and G. Krieger: Intrinsic dimensionality: nonlinear image operators and higher-order statistics. In: *Nonlinear image processing*, Academic Press, Orlando, 403–441 (2001)
11. T. Kohonen: Self-Organizing Maps. Springer: Berlin (2001)
12. T. Kohonen: Self Organized Formation of Topologically Correct Feature Maps. Biological Cybernetics 43, 59–69 (1982)
13. J. Kangas: On the Analysis of Pattern Sequences by Self-Organizing Maps. PhD thesis, Helsinki University of Technology, 1994. URL http://nucleus.hut.fi/ jari/papers/thesis94.ps.Z
14. E. I. Knudsen, S. du Lac, and S. D. Esterly: Computational maps in the brain. Annu Rev Neurosci 10, 41–65 (1987)
15. J. A. Anderson, A. Pellionisz, and E. Rosenfeld: Neurocomputing (vol. 2): directions for research. MIT Press: Cambridge, US (1990)
16. J. Ghosh and A. Strehl: Similarity-Based Text Clustering: A Comparative Study. In: J. Kogan, C. Nicholas and M. Teboulle, editors: *Grouping Multidimensional Data*, Springer, Berlin (2006)
17. J. H. Ward: Hierarchical Grouping to Optimize an Objective Function. Journal of the American Statistical Association 58, 236–244 (1963)
18. G. Salton, A. Wong, and C. S. Yang: A Vector Space Model for Automatic Indexing. Communications of the ACM 18, 613–620 (1975)
19. B. Larsen and C. Aone: Fast and effective text mining using linear-time document clustering. In: *KDD '99: Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, 16–22 (1999)
20. S. A. Nene, S. K. Nayar and H. Murase: Columbia Object Image Library (COIL-20). Technical Report CUCS-005-96 (1996)

Bio-inspired Architecture for Active Sensorimotor Localization

Thomas Reineking*, Johannes Wolter,
Konrad Gadzicki, and Christoph Zetzsche

Cognitive Neuroinformatics, University of Bremen,
P.O. Box 330 440, 28334 Bremen, Germany
trking@informatik.uni-bremen.de

Abstract. Determining one's position within the environment is a basic feature of spatial behavior and spatial cognition. This task is of inherently sensorimotor nature in that it results from a combination of sensory features and motor actions, where the latter comprise exploratory movements to different positions in the environment. Biological agents achieve this in a robust and effortless fashion, which prompted us to investigate a bio-inspired architecture to study the localization process of an artificial agent which operates in virtual spatial environments. The spatial representation in this architecture is based on sensorimotor features that comprise sensory sensory features as well as motor actions. It is hierarchically organized and its structure can be learned in an unsupervised fashion by an appropriate clustering rule. In addition, the architecture has a temporal belief update mechanism which explicitly utilizes the statistical correlations of actions and locations. The architecture is hybrid in integrating bottom-up processing of sensorimotor features with top-down reasoning which is able to select optimal motor actions based on the principle of maximum information gain. The architecture operates on two sensorimotor levels, a macro-level, which controls the movements of the agent in space, and on a micro-level, which controls its eye movements. As a result, the virtual mobile agent is able to localize itself within an environment using a minimum number of exploratory actions.

1 Introduction

In spite of substantial advances in the design of artificial intelligent systems, biological systems still represent the desirable ideal in many contexts. This is also true regarding a basic competence in spatial cognition: the ability to determine one's own location within the environment. In this paper our aim is to investigate how we can make use of results and concepts from psychology and neurobiology in the design of a bio-inspired architecture for vision-based localization.

For this we have to consider several factors: First, a basic prerequisite is an adequate representation of the environment. The notion of a *cognitive map* is often seen as an abstract copy of the physical layout, which may, for example, resemble

* Corresponding author.

an annotated cartographic representation, which can be substantially distorted with respect to the true metrical and geometrical properties, but which is basically similar to a two-dimensional image-like entity. This concept of a "map in the head" has been repeatedly criticized as potentially misleading, e.g., [17,42]. In our opinion, the most critical shortcoming of this concept is the absent or indirect role of motor actions in the representational model. In most cases, there is a clear separation between configurations of spatial entities and the actions that can be performed on/with it. This is in stark contrast to recent developments in perceptual psychology and neurobiology. Starting with the affordance concept of Gibson [10] over the common coding theory of Prinz [26,14] to the concept of sensorimotor contingencies of O'Regan [23], many psychological theories argue that the strict separation of sensory and motor components in the representational concepts is no longer tenable. The well-studied primate vision also brought up evidence for a stronger coupling of sensory and motor processes. For example, the system of canonical and mirror neurons revealed an intricate coupling of perception and motor control [30], and the postulated dorsal-ventral stream also shows representations of space used for the organization of actions [31]. Likewise, the firing of visual neurons in the ventral path, which were historically ascribed to early and solely sensory processing, turned out to be directly related to eye movements [20].

In understanding how a representation is organized it is also important to consider how it is established under natural behavioral conditions. Typically, this is a dynamic process in which motor actions play an essential role. Mobile agents move within the environment and produce a sequence of motor actions, and each action changes the relation between the agent and the environment. From the static perspective on a spatial representation, this is a disaster, but research in active perception has revealed that these motor actions actually simplify the development of a reliable representation of the environment [1,2]. It should be noted that motor actions also play an important role in the developmental landmark-route-survey (LRS) concept of spatial representation proposed in [37], although the final stage, the survey representation, again represents an image-like cognitive map. Finally, our own experiments with physically "impossible" virtual environments provide strong evidence against the concept of an image-like cognitive map [50,48], which is in line with a number of other studies which also found evidence against a cognitive map in the sense of an enduring allocentric representation [11,43,7]. Taken together, this indicates that a biologically plausible spatial representation should also comprise motor information, and preferably not as a simple add-on but in an integrated combination with sensory information. This leads us to make use of a sensorimotor representation of the spatial environment in our architecture.

A second important point in the design of a biologically plausible architecture is efficiency. Biological system often achieve their goals with a minimization of both effort and use of resources. Regarding information processing, this can be formalized as information-theoretic optimization. For example, the neural processing in the visual system can be successfully described as a result of such an

information theoretic optimization (e.g., [49]). For a biologically plausible architecture it would thus be desirable to obtain a maximum amount of information about an environment with a minimum number of motor actions [34]. As a last point in our design considerations, we have to take into account that biological representations are typically not established and used in a purely bottom-up fashion, but are part of an action-perception cycle, which involves bottom-up processing as well as top-down control.

Here, we approach the aforementioned ideas by the design of an artificial system, the Sensori-Motor Explorer (SMX), which is a virtual mobile agent that uses sensorimotor features as basic representational elements for exploratory localization in virtual spatial environments. The system presented here results from an integration of our sensorimotor representation [34,51] with a temporal belief update mechanism [27] and a learning component which allows for the unsupervised learning of the underlying hierarchical sensorimotor representations [9]. Central to the system is the use of the principle of maximum information gain to compute and execute the most informative actions. As a result, the SMX can localize itself in its environment using a minimum of exploratory steps.

The paper is organized as follows. In section 2, we provide a brief overview of the system properties of SMX, of its micro-level and macro-level exploration behavior, and of the generic hybrid architecture that is used to control both levels. The individual components of the system are then explained in more detail in section 3. This section contains also descriptions of the learning mechanism for the generation of sensorimotor hierarchies and of the temporal update of the belief distribution in response to spatial context changes. The resulting system behavior is described section 4. The paper concludes with a discussion of the major achievements.

2 System Architecture

The SMX performs exploration and localization in a VR environment. We use virtual reality and simulation in our research because this provides us with simple and complete control over all properties of both the environment and the agent. In particular, we can easily investigate the influence of features, objects and spatial arrangements on the performance of the system. In the current state, we use indoor environments consisting of rooms which are populated by typical objects like chairs, bookshelves, etc. The objects and the room walls have uniform or simple textures, typically with static lighting conditions.

The agent is characterized by two major features: first, it operates on two behavioral levels with different sensorimotor granularity, and, second, both behavioral levels are controlled by a single hybrid architecture, which integrates bottom-up sensory processing with a top-down uncertainty minimization strategy. The two sensorimotor levels are illustrated in Fig. 1: at the micro-level, a local view of the environment is explored in a detailed analysis by saccadic eye movements. At the macro-level, the agent performs exploratory movements within the spatial environment. In the bottom-up stage, features are extracted

from the environment and combined with motor data. The resulting sensorimotor features are the basic representational elements of the system. At the micro-level, these features comprise local image features and the motor data for saccadic eye movements while, at the macro-level, the sensorimotor features combine information about local views with the motor data for changing the agent's spatial location in the environment. At both levels, an estimation of the current state is performed based on the observed sensorimotor features. States at the macro-level are distinctive regions in space (in this particular case rooms) while states at the micro-level correspond to single views, which, in turn, form features at the macro-level. The resulting hypothesis spaces have a tree-like structure where nodes higher-up in the hierarchy represent disjunctions of states. They are obtained by hierarchically clustering a large number of sampled sensorimotor features [9]. Such a hierarchical structure has in fact been identified as a key property of spatial representations [13,45], and the formalization of the resulting estimation problem naturally lends itself to a belief function representation in Dempster-Shafer framework. In particular, this allows the agent to remain agnostic with respect to the belief distribution over the leaf nodes if the sensorimotor evidence does not support specific leaf nodes.

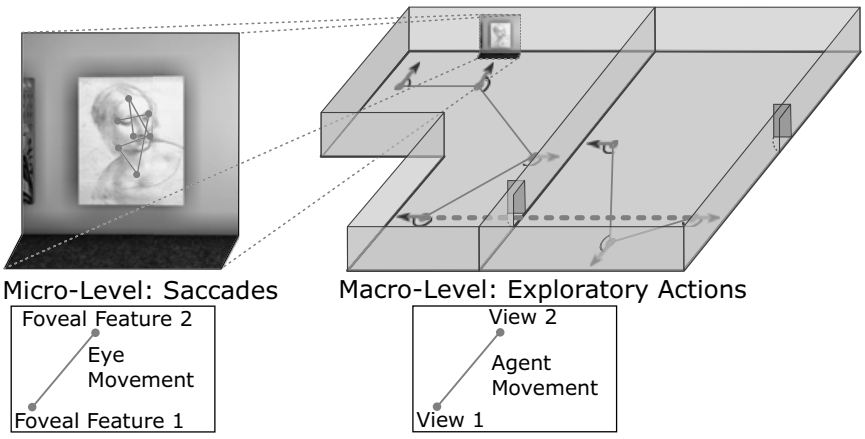


Fig. 1. Two levels of sensorimotor granularity. The micro-level scene analysis (left) is based on saccadic eye movements on a single view. At the macro-level (right), the agent moves between different locations in the environment.

The architecture used at both behavioral levels is shown in Fig. 2. In each action-perception cycle, a new sensorimotor feature is extracted and the belief distribution over the hierarchy of the corresponding level is updated based on a statistical model of state-feature co-occurrences which are learned in an initial training phase. The top-down component uses the updated belief in order to compute the expected information gain associated with subsequent features and their corresponding actions [34]. The sensorimotor levels operate in an interleaved fashion, with results from the micro-level passed as input to the macro-level. Together, they enable the agent to perform a statistically optimal sequence

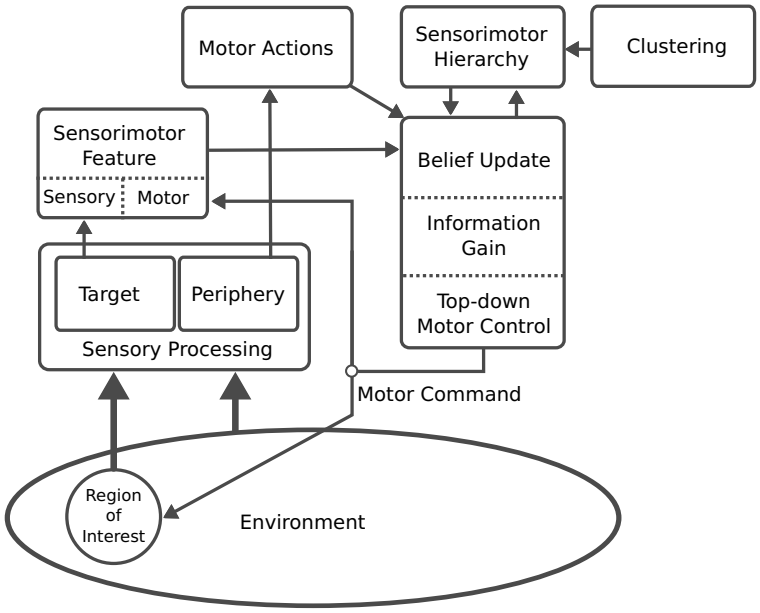


Fig. 2. Generic hybrid architecture. The same type of architecture is used for the micro- and the macro-level control. Sensory input and motor data are combined in the bottom-up processing to obtain a sensorimotor feature, which is used to update the belief distribution over the hierarchy at the corresponding level. From this, the top-down strategy selects a new action by minimizing the expected uncertainty.

of exploratory actions to gain information about the environment. The different components of the system, the internal sequence of operations and the resulting behavior are described in more detail in the following section.

3 Components of the System

3.1 Sensorimotor Features

A characteristic property of the system is its representation of the spatial environment via sensorimotor features, which form the basic representational elements for both the macro-level and the micro-level. A sensorimotor feature is a triple of the form $f = [v_1, a, v_2]$ where v_1 is the sensory feature vector obtained prior to action a and v_2 is the sensory feature vector obtained after executing a. By making actions an explicit part of the representation, they become an additional source of information for the state estimation because one can make use of their correlation with states—a feature not present in classical localization approaches where one is typically agnostic with respect to the question of where certain actions are likely to occur, e.g., [40]. The continuous-valued sensorimotor vector f is mapped to the closest element f_i of a finite set of prototype vectors. At the micro-level, the motor component of each feature is a saccadic eye movement, and the sensory components are derived by a biologically motivated vision system by use of a local wavelet analysis that is applied to the pre- and post-saccadic fixation points.

At the macro-level, the motor component is a movement of the complete agent in space, and the sensory components are labels (for the local views) that the agent registers before and after the movement. The label for a local view is the result of the micro-level analysis of the local view by saccadic eye movements, and this label is then passed as sensory input to the macro-level analysis. Each level makes use of a discrete set of sensorimotor micro- and macro-level features, and they are acquired in an initial exploration process in which the association of sensorimotor features with states in the environment is established by supervised learning.

An action at the macro-level consists of two rotations and one translation. The first rotation turns the agent at the starting location in the direction of the target location, and the following straight movement gets the agent to this location. Here a second rotation aligns the agent to the orientation of the target view. At the macro-level there is a distinction between the handling of intra- and inter-room sensorimotor features. An intra-room feature belongs to a single room and the updating for this case is described in 3.4. Inter-room features, on the other hand, are those where the pre-action part belongs to one room and the post-action part belongs to another. Here, the resulting belief has to be transferred to the corresponding destination, which is described in 3.5.

3.2 Saccadic Eye Movements

An essential problem of processing a visual scene is the detection of the most informative visual regions (e.g., those parts of an object, which are most informative for its identification). Information about the image structure at these few locations is usually sufficient to draw reliable conclusions about the local scene. Biological vision systems have developed an efficient design in which the pattern recognition capabilities are concentrated in a small region of the visual field, the central fovea, whereas the periphery has only limited optical resolution and processing power. With a static eye, one can hence only see a small spot of the environment with satisfactory quality, but this spot can be rapidly moved with fast saccadic eye movements of up to $700°/s$ towards all the "relevant" regions of a scene. This selection process is determined by bottom-up processes on the input scene as well as by top-down processes determined by the memory, internal states and current tasks [47].

In order to enable an efficient selective "sampling" of a local scene by saccadic eye movements, we have integrated bottom-up and top-down processing into a hybrid architecture (cf. Fig. 2). With respect to saccadic scene analysis, the sensory processing stage in this architecture has two functions. On the one hand, the pre-processing stage has to identify highly informative candidate locations within the scene, which can be the target of saccadic fixations, and, on the other hand, it has to provide detailed information about the fixated local pattern. It consists of a wavelet-like image decomposition by size- and orientation-specific filters and by nonlinear saliency operators based on the concept of intrinsic dimensionality. A detailed description of the filters and the non-linear operators can be found in [49] and [34]. In addition, we process each scene by a ratio of Gaussians first in order to increase luminance invariance. The extracted visual

features are combined with the motor information necessary for shifting the focus to the next fixation point, thus forming a micro-level sensorimotor feature which is then used to update the belief about the current scene.

3.3 Generating Sensorimotor Hierarchies

The hypotheses used by the SMX are represented in a hierarchical manner for efficiency as well as for coping with non-specific evidence. The macro- and the micro-level both have their individual hierarchical representation. Each node $H \in \mathbf{H}$ in the respective hierarchy $\mathbf{H}$ represents a set of singleton hypotheses, and the leaf nodes represent hypotheses about individual items. The leaf hypotheses are currently pre-defined while the higher-level nodes representing sets of views or rooms are generated in an unsupervised clustering process. Fig. 3 illustrates this structure for both, the micro- and macro-level.

The hierarchical structure in our system is organized according to the similarity of the sensorimotor information associated with different states, i.e., views and rooms sharing similar features are grouped together in the clustering process. In the past, we have investigated alternative grouping principles, in particular

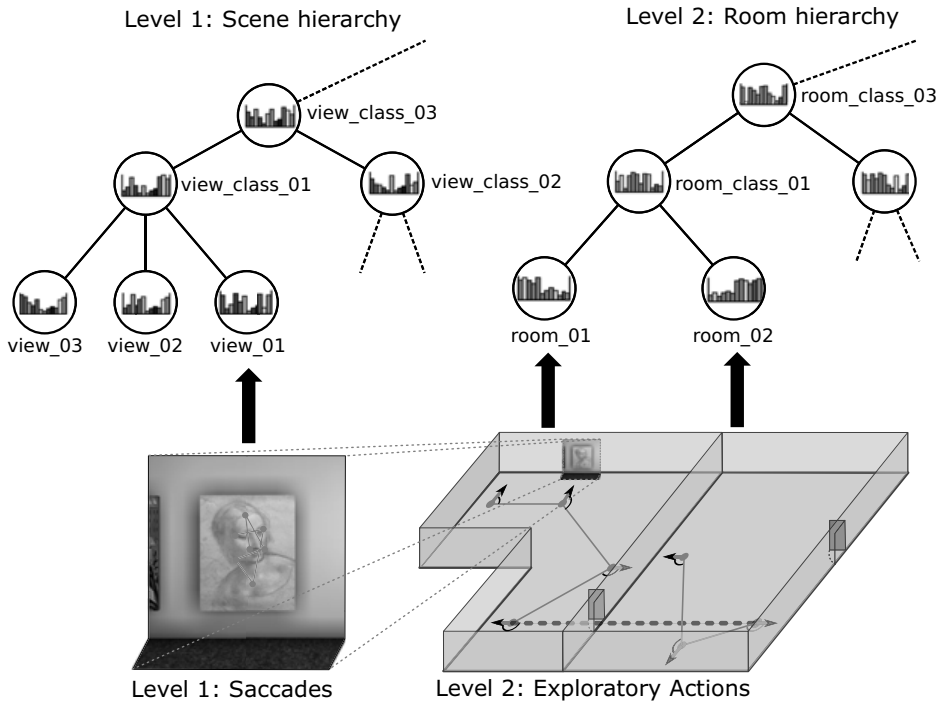


Fig. 3. Two levels of hierarchical sensorimotor representations. At the micro-level (left) scenes are organized in a hierarchical manner with individual scenes as leafs. A histogram of quantized saccadic information is stored in each node as a description of the scene/class of scenes. At the macro-level (right) an equivalent representation is maintained for spatial information with leaf nodes representing single rooms. A room description consists of a histogram of quantized macro-level sensorimotor features (views with motor action).

"spatial similarity" [28], where each node represents a connected region in space, and semantic similarity [29] based on an ontological model. The most suited principle depends on the task, and we use sensorimotor similarity here, because the resulting clusters are the most relevant for scene analysis/localization[1] while a region-based organization might be more appropriate for a task like navigation.

In a previous version of the system [51], these hierarchies were constructed manually for the domain. In order to automate this process, we developed an unsupervised learning method that generates a hierarchy by performing an agglomerative clustering on distributions of sensorimotor features associated with a each hypothesis H. For this, one first needs to measure the similarity of two sensorimotor features. Treating them as symbols and testing for equality worked well for the recognition task described in [34], but it restricts the measure to set-based operations. Initial tests with such similarity measures produced only mediocre results and led us to consider a numerical representation. Similar to approaches from document clustering [32], we thus decided to use a numerical representation which allows for the usage of more suited similarity measures.

For obtaining the numerical representation, sensorimotor features are extracted from a large set of samples and a Self-Organizing Map (SOM) [16] is trained with the entire sample set. That way the system learns how to group sets of features for a given output size (the map size) which corresponds to the number of components of the vector used for the representation of an instance. The actual representation for a particular instance is obtained by processing all sensorimotor features associated with a instance by the SOM. For each feature, a particular map node will be maximally activated. By counting the activations of map nodes for all features, a histogram is generated for each instance, which can then be compared by the similarity measure. Based on this vector representation, an agglomerative clustering algorithm generates a dendrogram which results in the desired hierarchy. Starting with each pattern in a singleton cluster, clusters are merged iteratively until only one is left. The merging process is driven by linkage rules which define which two clusters are merged in each step and by the similarity measure, which is calculated between individual patterns populating a cluster. Empirical tests on the COIL-20 image set [21] using different combinations of linkage rules and similarity measures led to the conclusion that Ward's linkage rule [44] and the Tanimoto coefficient [39] as a similarity measure produce the most robust and suitable results.

3.4 Belief Update

The agent has to cope with the typical incompleteness, ambiguity and inconsistency in input data from realistic environments. We hence have developed an inference component, which uses Dempster-Shafer theory for uncertain reasoning [36]. This theory can be considered as a generalization of probability theory and distinguishes between conflicting evidence and a lack of knowledge. A basic concept of this theory is the frame of discernment Θ, which is the set of

[1] One can measure the suitability of different clustering principles by the loss of information that the resulting hierarchies introduce during the belief update.

all possible singleton hypotheses in the domain. The resulting hypothesis space 2^{Θ} comprises all possible subsets $H \subseteq \Theta$. Since we are using the hierarchical representation $\mathbf{H}$ described above, the number of relevant subsets $H \in \mathbf{H}$ is substantially reduced. The belief induced by a piece of evidence can be expressed by a mass function $m : \mathbf{H} \rightarrow [0, 1]$ that assigns values to all hypotheses H such that $\sum_{H \subseteq \mathbf{H}} m(H) = 1$. A mass distribution can be interpreted as an underspecified probability distribution that preserves the unspecificity of the underlying evidence (e.g., a piece of evidence could support multiple hypotheses without committing to a specific probability distribution over these hypotheses). In addition, a belief can be equivalently described by a plausibility function pl defined as $pl(H) = \sum_{H' \cap H \neq \emptyset} m(H')$, which is sometimes more convenient.

Given all collected sensorimotor features $f_{0:t} = f_0, \ldots, f_t$, the mass distribution $m(H_t|f_{0:t})$ over the hierarchy can be recursively computed by the generalized Bayesian theorem [38,4]:

$$m(H_t|f_{0:t}) = \eta \prod_{h_t \in H_t} pl(f_{0:t}|h_t) \prod_{h_t \in H_t^C} (1 - pl(f_{0:t}|h_t)) \,, \tag{1}$$

$$pl(f_{0:t}|h_t) = pl(f_t|h_t) \sum_{H'_t \ni h_t} m(H'_t|f_{0:t-1}) \,. \tag{2}$$

Here, η is a normalization constant ensuring that the resulting mass values sum up to 1, and H_t^C denotes the complement of H_t. The plausibility $pl(f_t|h_t)$ of the new feature f_t given a hypothesis h_t (i.e., a room at the macro-level and a scene at the micro-level) is given by the relative frequency with which f_t was observed together with h_t during the training phase. Since we are using the hierarchical representation $\mathbf{H}$ instead of the full hypothesis space 2^{Θ}, the computational complexity of the update is reduced from exponential to linear [12,24]. This restriction introduces a certain error, however, due to the grouping of hypotheses sharing similar features in the clustering process, this error is largely negligible in practice.

3.5 Context Change

Compared to an earlier version, we extended the architecture by incorporating inter-room sensorimotor features in the update process using a Dempster-Shafer filter algorithm, which performs a prediction step similar to that in classical Bayes filters [15,40]. For features that indicate a state transition, the update in (1) is preceded by the following transition update. It consists of a conjunctive combination of the prior belief over H_{t-1} with the newly induced belief over H_t (under a first-order Markov assumption) by summing over all prior states H_{t-1} [5]:

$$m(H_t|f_{0:t}) = \eta \sum_{H_{t-1}} m(H_t|H_{t-1}, f_t)\, m(H_{t-1}|f_{0:t-1}) \tag{3}$$

The transition belief $m(H_t|H_{t-1}, f_t)$ can be further simplified by applying the disjunctive rule of combination $\textcircled{\cup}$ [38] in order to make the belief depend only on singletons h_{t-1} instead of aggregated states H_{t-1}:

$$m(H_t|H_{t-1}, f_t) = \bigcircledcup_{h_{t-1} \in H_{t-1}} m(H_t|h_{t-1}, f_t)\,, \tag{4}$$

$$(m_1 \circledcup m_2)(H) = \sum_{H_1 \cup H_2 = H} m_1(H_1)\, m_2(H_2)\,. \tag{5}$$

Each $m(H_t|h_{t-1}, f_t)$ is estimated in the training phase mentioned above. What is interesting here is that, in contrast to the prediction step in Bayes filters, state changes actually provide additional information which leads to a refinement of the localization belief. This is due to the fact that actions do not occur in arbitrary states but rather follow distinct sensorimotor patterns, which usually ties the act of moving in a certain way to a distinct set of locations in the environment (e.g., turning is more likely at corners).

3.6 Top-Down Uncertainty Minimization

The uncertainty minimization strategy used by the agent is based on the IBIG algorithm (*inference by information gain* [33]) and lets the agent perform actions that reduce the overall amount of uncertainty, both at the micro- and at the macro-level. Its basic principle is to determine the action a^* exhibiting the highest expected information gain with respect to the current belief distribution. The expected uncertainty is computed for each potential sensorimotor feature $\hat{f}_a = (v_1, a, v_2)$ that is compatible with the current state, i.e., v_1 and a match the current sensory input and a possible action while v_2 is integrated out. First, (1) and (3) are applied for each potential sensorimotor feature $\hat{s}_a$ corresponding to an action a in order to obtain the updated belief $m(H_t|f_{0:t}, \hat{f}_a)$. Next, the local conflict uncertainty measure I [25] is used to select the action a^* yielding the lowest expected uncertainty:

$$I(m) = \sum_H m(H) \log \frac{|H|}{m(H)}\,, \tag{6}$$

$$a^* = \arg\min_a E\left[I(m(\cdot|f_{0:t}, \hat{f}_a))\right]\,. \tag{7}$$

After executing the selected action (an eye movement at the micro-level or a change in location at the macro-level) the belief is updated with the actually observed sensorimotor feature and additional exploration steps are performed if ambiguity persists.

4 System Behavior

The following is a short description of how the components of the SMX interact in order to explore the environment and localize the agent in it. During the initial exploration phase the agents moves in the environment in order to build

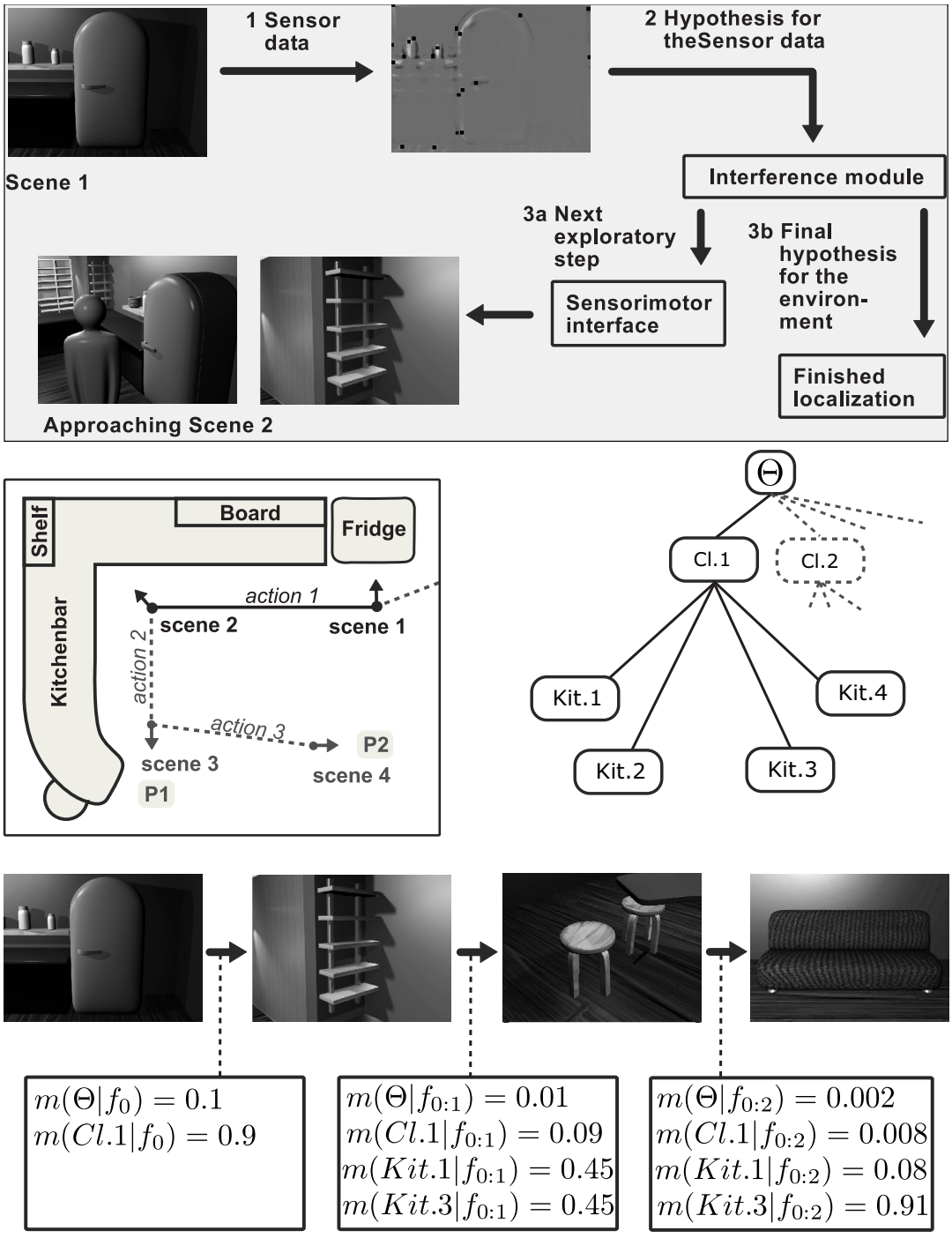


Fig. 4. Illustration of how the SMX localizes itself in the environment. An single update cycle at the macro-level is shown at the top. Parts of the environment and its hierarchical representation are shown in the middle. The lower third shows the updated beliefs after processing a sequence of sensorimotor features.

up a representation for both, the micro and macro level. Due to the way sensorimotor features are represented, salient points need to be extracted from the environment, serving as starting and destination points for actions. At the micro level, an image filter sensitive to intrinsic 2-dimensional features extracts the salient points from views, while, at the macro level, we currently use pre-defined locations in the environment. From the set of all possible sensorimotor features a large number of samples is randomly generated for both levels. Based on these samples, the hierarchical representations are build by agglomerative clustering.

The macro-level cycle starts with the agent facing a local scene. The micro-level subsystem is used to analyze this local scene by saccadic eye movements. The new sensory information returned from the micro-level exploration is a scene label which, together with the macro-level motor data and the sensory information obtained at the previous location, forms the new macro-level sensorimotor feature f_t. This feature is then used to update the current belief distribution over the hierarchy using (1) for an intra-room feature and additionally (3) for an inter-room feature (indicating a context change). Based on this, the next action is selected according to the minimum expected uncertainty as defined by (7). The agent executes this action by first rotating and then moving towards the target location. At the new location, it rotates again, if necessary, and starts a new micro-exploration by the saccadic eye movement subsystem. The cycle is repeated until a sufficient belief threshold for one of the macro-level hypotheses is reached.

A small example of a complete localization run consisting of three exploration steps is shown in Fig. 4. After processing the first sensorimotor feature, the agent has a strong belief for cluster 1, which consists of four kitchens. After processing the next feature, the evidence equally supports two of the kitchens and only the final feature completely resolves this ambiguity. A detailed quantitative analysis of the system's localization performance and of the efficiency of the action selection strategy was conducted in [51] using different virtual environments. The number of exploration steps required for sufficiently reducing localization uncertainty is considerably lower compared to the baseline of randomly performing actions (see Fig. 5), in particular for environments that exhibit high degrees of perceptual aliasing.

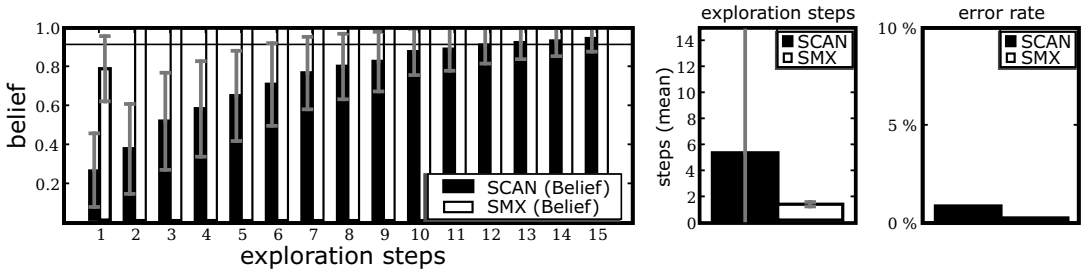


Fig. 5. Performance comparison of the IBIG exploration strategy with random action selection based on the number of required actions for reaching a belief level of 0.9 (left) and on the error rate (right). Figure adapted from [51].

5 Discussion

We developed a mobile virtual agent, SMX, which localizes itself in an indoor environment via exploratory actions at two levels of sensorimotor granularity. Compared to a previous version of the agent [51], we extended it by a belief update across different rooms in the environment and by a method for obtaining the underlying hierarchical representations in an unsupervised fashion. The former is based on a new approach for updating Dempster-Shafer belief distributions over time [27] and allows the agent to distinguish identical looking rooms and reduce localization uncertainty more quickly. Beyond that, the agent is characterized by three main properties: first, the spatial environment is represented in terms of sensorimotor features. This is motivated by doubts about the biological plausibility of map-like representations, and by psychological and neurobiological evidence suggesting a joint contribution of sensory and motor information to perception and representation. In particular, the sensorimotor representation enables the utilization of actions as an additional source of information due to their correlation with states. Second, the system operates in a loop of bottom-up processing and tow-down reasoning governed by the principle of information gain. This is achieved by a hybrid architecture in which a top-down strategy selects those exploratory actions providing the highest expected information gain with respect to the current belief. Third, the same generic hybrid architecture and information-gain strategy are used at two levels of sensorimotor granularity. This results in active localization behavior with location changes at the macro-level and saccadic eye movements at the micro-level, which mimic the way humans analyze visual scenes.

The combination of these components yields a psychologically and neurobiologically plausible system that acquires a maximum amount of information about its environments using a minimum number actions. This uncertainty minimization principle is particularly important for environments exhibiting a high degree of perceptual aliasing, e.g., rooms consisting of many similar or identical objects. The tests conducted in [51] furthermore indicate that the performance is not degraded by minor distortions of image features or of object configurations.

As mentioned, the SMX shows some differences to commonly used representations of spatial environments. The greatest difference exists with respect to those approaches that use grid-like or image-like two-dimensional maps since these do not include any explicit information about potential motor actions [6,41]. With respect to topological representations, the relation is dependent on the interpretation. If the key concept of a topological representation is seen in the abstraction from metrical properties, there is a clear difference to our approach since the motor actions in our sensorimotor representation are encoded in association with metrical information (e.g., the translation vector of an eye movement). Topological representations are also different from our approach if they are simply seen as less restricted variants of conventional spatial maps. However, it is also possible to interpret the edges in a topological graph in the sense of actions that are required to move from one node to the other, and under this perspective there is a much closer relation to the sensorimotor representation (e.g., [3,19,11,18]).

In the future development of SMX, we will apply the generic architecture to additional granularity levels of sensorimotor features. Furthermore, we will investigate the suitability of different clustering principles at these levels, e.g., by incorporating semantics of clusters [35] and by using spatial structuring principles that are better suited for problems like large-scale navigation [46,28]. This design will be guided by an ongoing evaluation based on comparisons with empirical results. On the behavioral side, this will be actual eye movements and macro-levels actions of human subjects, both in realistic and in virtual environments. On the neurobiological side, the most interesting entity for our future system development is the place cell [22]. Although there is currently no module in our system that is intended as a model of place cells, it is interesting to note that the units in our hierarchy are both influenced by the properties of the local environment and by the history of how the agent arrived at the current position, a non-trivial property that has also been observed in hippocampal neurons [8].

Acknowledgements

This work was supported by the DFG (SFB/TR 8 Spatial Cognition, project A5-[ActionSpace]).

References

1. Aloimonos, Y. (ed.): Special Issue: Purposive, Qualitative and Active Vision, Image Understanding, vol. 56 (1992)
2. Ballard, D.: Animate vision. Artificial Intelligence 48, 57–86 (1991)
3. Byun, Y.T., Kuipers, B.: A robot exploration and mapping strategy based on a semantic hierarchy of spatial representations. IEEE Journal of Robotics and Autonomous Systems 8, 47–63 (1991)
4. Delmotte, F., Smets, P.: Target identification based on the transferable belief model interpretation of Dempster-Shafer model. IEEE Transactions on Systems, Man, and Cybernetics, Part A: Systems and Humans 34(4), 457–471 (2004)
5. Dubois, D., Prade, H.: On the unicity of Dempster's rule of combination. International Journal of Intelligent Systems 1(2), 133–142 (1986)
6. Elfes, A.: Sonar-based real-world mapping and navigation. IEEE Journal of robotics and automation 3(3), 249–265 (1987)
7. Foo, P., Warren, W.H., Duchon, A., Tarr, M.J.: Do humans integrate routes into a cognitive map? map- versus landmark-based navigation of novel shortcuts. Journal of Experimental Psychology 31(2), 195–215 (2005)
8. Frank, L., Brown, E., Wilson, M.: Trajectory encoding in the hippocampus and entorhinal cortex. Neuron 27(1), 169–178 (2000)
9. Gadzicki, K.: Hierarchical clustering of sensorimotor features. In: Mertsching, B., Hund, M., Aziz, Z. (eds.) KI 2009. LNCS (LNAI), vol. 5803, pp. 331–338. Springer, Heidelberg (2009)
10. Gibson, J.J.: The ecological approach to visual perception. Houghton Mifflin, Boston (1979)
11. Gillner, S., Mallot, H.A.: Navigation and acquisition of spatial knowledge in a virtual maze. Journal of Cognitive Neuroscience 10(4), 445–463 (1998)

12. Gordon, J., Shortliffe, E.H.: A method for managing evidential reasoning in a hierarchical hypothesis space. Artif. Intell. 26(3), 323–357 (1985)
13. Hirtle, S.C., Jonides, J.: Evidence of hierarchies in cognitive maps. Memory and Cognition 13(3), 208–217 (1985)
14. Hommel, B., Muesseler, J., Aschersleben, G., Prinz, W.: The theory of event coding (tec): A framework for perception and action planning. Behavioral and Brain Sciences 24, 849–878 (2001)
15. Kalman, R.: A new approach to linear filtering and prediction problems. Journal of Basic Engineering 82(1), 35–45 (1960)
16. Kohonen, T.: Self-organizing maps. Springer series in information sciences, 3rd edn., vol. 30. Springer, Heidelberg (2001)
17. Kuipers, B.: The map in the head metaphor. Environment and Behavior 14(2), 202–220 (1982)
18. Kuipers, B.: The spatial semantic hierarchy. Artificial Intelligence 119, 191–233 (2000)
19. Mataric, M.: Integration of representation into goal-driven behavior-basedrobots. IEEE Transactions on Robotics and Automation 8(3), 304–312 (1992)
20. Moore, T.: Shape representations and visual guidance of saccadic eye movements. Science 285(5435), 1914 (1999)
21. Nene, S., Nayar, S., Murase, H.: Columbia object image library (COIL-20). Tech. rep., Dept. Comput. Sci., Columbia Univ., New York (1996)
22. O'Keefe, J., Nadel, L.: The hippocampus as a cognitive map. Clarendon Press, Oxford (1978)
23. O'Regan, J.K., Noë, A.: A sensorimotor account of vision and visual consciousness. Behavioral and Brain Sciences 24, 939–973 (2001)
24. Orponen, P.: Dempster's rule of combination is #P-complete. Artificial Intelligence 44(1-2), 245–253 (1990)
25. Pal, N., Bezdek, J., Hemasinha, R.: Uncertainty measures for evidential reasoning II: A new measure of total uncertainty. International Journal of Approximate Reasoning 8(1), 1–16 (1993)
26. Prinz, W.: A common coding approach to perception and action, relationships between perception and action: current approaches edn., pp. 167–203. Springer, Berlin (1990)
27. Reineking, T.: Particle filtering in the Dempster-Shafer theory. International Journal of Approximate Reasoning (2010) (in revision) (submitted) (Febuary 17, 2009)
28. Reineking, T., Kohlhagen, C., Zetzsche, C.: Efficient wayfinding in hierarchically regionalized spatial environments. In: Freksa, C. (ed.) Spatial Cognition VI. LNCS (LNAI), vol. 5248, pp. 56–70. Springer, Heidelberg (2008)
29. Reineking, T., Schult, N., Hois, J.: Combining statistical and symbolic reasoning for active scene categorization. In: Knowledge Discovery, Knowledge Engineering and Knowledge Management, IC3K 2009, Revised Selected Papers. Communications in Computer and Information Science. Springer, Heidelberg (2010) (in press)
30. Rizzolatti, G., Craighero, L.: The mirror-neuron system 27, 169–192 (2004)
31. Rizzolatti, G., Matelli, M.: Two different streams form the dorsal visual system: anatomy and functions. Experimental Brain Research 153(2), 146–157 (2003)
32. Salton, G.: The SMART Retrieval System—Experiments in Automatic Document Processing. Prentice-Hall, Inc., Upper Saddle River (1971)
33. Schill, K.: Decision Support Systems with Adaptive Reasoning Strategies. In: Freksa, C., Jantzen, M., Valk, R. (eds.) Foundations of Computer Science. LNCS, vol. 1337, pp. 417–427. Springer, Heidelberg (1997)

34. Schill, K., Umkehrer, E., Beinlich, S., Krieger, G., Zetzsche, C.: Scene analysis with saccadic eye movements: Top-down and bottom-up modeling. Journal of Electronic Imaging 10(1), 152–160 (2001)
35. Schill, K., Zetzsche, C., Hois, J.: A belief-based architecture for scene analysis: From sensorimotor features to knowledge and ontology. Fuzzy Sets and Systems 160(10), 1507–1516 (2009)
36. Shafer, G.: A Mathematical Theory of Evidence. Princeton University Press, Princeton (1976)
37. Siegel, A., White, S.: The development of spatial representations of large-scale environments. Advances in child development and behavior 10, 9 (1975)
38. Smets, P.: Belief functions: The disjunctive rule of combination and the generalized Bayesian theorem. International Journal of Approximate Reasoning 9, 1–35 (1993)
39. Tanimoto, T.: IBM interal report. Tech. rep., IBM (November 1957)
40. Thrun, S., Fox, D., Burgard, W., Dellaert, F.: Robust Monte Carlo localization for mobile robots. Artificial Intelligence 128(1-2), 99–141 (2001)
41. Thrun, S.: Learning occupancy grids with forward sensor models. Autonomous Robots 15, 111–127 (2003)
42. Tversky, B.: Distortions in cognitive maps. Geoforum 23(2), 131–138 (1992)
43. Wang, R.F., Spelke, E.S.: Updating egocentric representations in human navigation. Cognition 77, 215–250 (2000)
44. Ward, J.H.: Hierarchical grouping to optimize an objective function. Journal of the American Statistical Association 58(301), 236–244 (1963)
45. Wiener, J., Mallot, H.: 'Fine-to-coarse' route planning and navigation in regionalized environments. Spatial Cognition and Computation 3(4), 331–358 (2003)
46. Wiener, J., Mallot, H.: 'Fine-to-coarse' route planning and navigation in regionalized environments. Spatial Cognition and Computation 3(4), 331–358 (2003)
47. Yarbus, A.L.: Eye Movements and Vision. Plenum Press, New York (1967)
48. Zetzsche, C., Galbraith, C., Wolter, J., Schill, K.: Navigation based on a sensorimotor representation: a virtual reality study. In: Rogowitz, B.E., Pappas, T.N., Daly, S.J. (eds.) Proceedings of SPIE. Human Vision and Electronic Imaging XII, February 2007, vol. 6492 (2007)
49. Zetzsche, C., Krieger, G.: Nonlinear operators and higher-order statistics in image processing and analysis. In: Proc. ISPA 2001 - 2nd International Symposium on Image and Signal Processing and Analysis, pp. 119–124 (2001)
50. Zetzsche, C., Wolter, J., Galbraith, C., Schill, K.: Representation of space: image-like or sensorimotor. Spatial Vision 22(5), 409–424 (2009)
51. Zetzsche, C., Wolter, J., Schill, K.: Sensorimotor representation and knowledge-based reasoning for spatial exploration and localisation. Cognitive Processing 9, 283–297 (2008)

IAC-15-A3.IP.15

KANARIA: IDENTIFYING THE CHALLENGES FOR COGNITIVE AUTONOMOUS NAVIGATION AND GUIDANCE FOR MISSIONS TO SMALL PLANETARY BODIES

Alena Probst[*], Graciela González Peytaví[†], David Nakath[‡], Anne Schattel[§], Carsten Rachuy[**], Patrick Lange[††], Joachim Clemens[‡], Mitja Echim[§], Verena Schwarting[**], Abhishek Srinivas[†], Konrad Gadzicki[**], Roger Förstner[*], Bernd Eissfeller[†], Kerstin Schill[‡,], Christof Büskens[§], Gabriel Zachmann[††]**

[*] Institute of Space Technology & Applications, Space Technology Dept., Bundeswehr University Munich, Germany, A.Probst@unibw.de

[†] Institute of Space Technology & Applications, Navigation Dept., Bundeswehr University Munich, Germany, graciela.gonzalez@unibw.de

[‡] Cognitive Neuroinformatics, Sensor Fusion Group, University of Bremen, Germany, dnakath@informatik.uni-bremen.de

[§] Working group Optimization and Optimal Control, University of Bremen, Germany, ascha@math.uni-bremen.de

[**] Cognitive Neuroinformatics, Autonomy Group, University of Bremen, Germany, rachuy@informatik.uni-bremen.de

[††] Computer Graphics and Virtual Reality, University of Bremen, Germany, lange@informatik.uni-bremen.de

With the rapid evolution of space technologies and increasing thirst for knowledge about the origin of life and the universe, the need for deep space missions as well as for autonomous solutions for complex, time-critical mission operations becomes urgent. Within this context, the project KaNaRiA aims at technology development tailored to the ambitious task of space resource mining on small planetary bodies using increased autonomy for on-board mission planning, navigation and guidance.

This paper focuses on the specific challenges as well as first solutions and results corresponding to the KaNaRiA mission phases (1) interplanetary cruise, (2) target identification and characterization and (3) proximity operations.

Based on the KaNaRiA asteroid mining mission objectives, initially, a mission reference scenario as well as a reference mission architecture are described in this paper. KaNaRiA has been proposed as a multi-spacecraft mission to the asteroid main belt. Composed of a flock of prospective scout spacecraft, a mother ship carrying the mining payload and several service modules placed on a 2.8 AU parking orbit around the Sun, KaNaRiA intends to characterize main belt asteroid properties, identify targets for mining and perform a soft-landing for in-situ characterization and mining.

Subsequently, the autonomous navigation system design of KaNaRiA for the interplanetary cruise is presented. The navigation challenges, which arise in phases (1) to (3), are discussed. Particular attention is given to the sensor-technology readiness-level, accuracy, applicability range, mass and power budgets. In order to navigate in the vicinity of an asteroid, an information fusion algorithm is required that aggregates multi-sensor data as well as a-priori knowledge and solves the task known as simultaneous localization and mapping (SLAM). In order to deal with uncertain and inconsistent information and to explicitly represent different dimensions of uncertainty, a belief-function-based SLAM approach is used, which is a generalization of the popular FastSLAM algorithm.

The objective of the guidance task is the autonomous planning of optimal transfer trajectories according to mission driving criteria, e.g. transfer time and fuel consumption. Optimal control problems and the calculation of trajectory sensitivities for on-board stability analysis as well as real-time optimal control are explained.

Bringing cognitive autonomy to a spacecraft requires an on-board computational module as a central spacecraft component. This module is responsible for state evaluation, mission planning and decision-making regarding selection of potential targets, trajectory selection and FDIR. A knowledge-base serves as a database for decision making processes.

With the aim to validate and test our methods, we create a virtual environment in which humans can interact with the simulation of the mission. In order to achieve real-time performance, we propose a massively-parallel software system architecture, which enables very efficient and easily adaptable communication between concurrent software modules within KaNaRiA.

I. INTRODUCTION

Following the developments and the news on current space missions such as Rosetta or Dawn, one of the biggest challenges for small body rendezvous and landing missions is the large communication delay that leads to operational problems. Operations need to be planned thoroughly in advance. Nevertheless failures and anomalies often result in the complete loss of the spacecraft or lander. One approach to improve the reliability of complex operations is to enhance the autonomy, decision making and FDIR (fault, detection, isolation and recovery) capabilities of the spacecraft.

This is the approach that the project KaNaRiA takes up. The German acronym KaNaRiA stands for *Kognitionsbasierte, autonome Navigation am Beispiel des Ressourcenabbaus im All*, which translates into *Cognitive Autonomous Navigation for Deep Space Resource Mining.* As an interdisciplinary project, KaNaRiA focuses on autonomous mission planning, navigation and guidance in a-priori unknown environments dealing with the challenges of future space missions to minor planets. KaNaRiA strives to increase on-board spacecraft autonomy in the context of an asteroid mining scenario. The development of these concepts takes place in a virtual simulation environment, which serves as a test bed for a mission study. In this paper we give an overview of the KaNaRiA mission concept and the individual components of the system.

The paper is structured as follows. In section II and III, the engineering solutions applied to the particular mission scenario of KaNaRiA are presented, specifically the mission concept and reference scenario followed by the navigation system design and autonomous navigation concept.

Section IV covers the contribution of information fusion, which combines a-priori knowledge with sensor data to provide an information basis for autonomous decision-making.

In section V it is explained how the mathematical field of optimization and optimal control is used to calculate optimal interplanetary trajectories by solving infinite-dimensional optimal control problems.

In section VI the central component for on-board mission planning and autonomous decision-making is presented.

Section VII describes functionality of the simulation environment and its underlying software architecture.

II. MISSION: ASTEROID MINING

As an application for the proposed autonomous navigation, guidance and simulation solutions, an asteroid mining mission concept is defined.

The aim of asteroid mining opens up a huge space of scenarios and possibilities to implement a successful mission. The mission design changes depending on the desired resource, the purpose of usage or the location of the asteroid target. In order to specify a scenario, the JPL Rapid Mission Architecture [1] method has been applied.

II.I Mission Processes

The mission concept derivation is based on a separation and identification of processes that have to be fulfilled with the goal of mining a space body. First, the targets have to be mapped and characterized according to their natural resources and potential consideration for mining. These activities are done under the scope of Mapping, Characterization and Resource Determination (MCRD). Second, after having appointed a suitable target, the resource is mined by a separate miner (Resource Extraction and Exploitation, REaE). As an asteroid mining mission is by default a long-term mission, the transportation of the resources from the mining site to the refinery or designated user as well as the maintenance of the space elements involved have to be taken into account. Those activities are covered within the Maintenance and Logistics. For a more detailed description and definition of the mission, it is referred to Probst et al. [2]

As each of the processes involved in a successful mining mission imposes different requirements on the spacecraft architecture, separate spacecraft elements have been selected, each of them specialized for one specific process. The selection trade-off for each mission element architecture was done using a numerical method based on relative judgments with respect to suitable trade-criteria. The selection process is described in Probst et al. [2]

II.II Mission Elements

The following mission elements are involved in the mission scenario:

The *Potential Target Characterization Modules* (PTCMs) are in charge of exploring the considered targets in order to analyse their potential resource character.

The *KaNaRiA Miner Spacecraft* (KMS) lands on the designated target and excavates the resource.

The Refuel- and Repair- Elements (RF/RP) take care of the maintenance problems that occur.

An unmanned, autonomous *Operational Centre* (OC) serves as the main communication and delegation hub. It coordinates the mission elements and their tasks, sustains and collects the data and inherits the overall power of decision.

To complete the mining cycle, *resource transporters* are needed that carry the material from the mining site to the refinery or from there to the costumer.

II.III Mission Reference Scenario

As the mission scenario serves as a basis for the navigation, sensor fusion, guidance and autonomy algorithms and their implementation in a simulator, the mission scenario starts with a mission setup at a circular, Sun-bound parking orbit (2PO) with a semi-major axis of 2.8 AU. [2]

On 2PO, the OC, KMS and the maintenance spacecraft as well as the transporter are stationed whereas several PTCMs swarm out for their investigation and search for potential precious resources. Each PTCM consists of an orbiter and a re-docking lander so that it can visit more than one asteroid without coming back to 2PO. This way it is able to characterize each target thoroughly. The data obtained is relayed to the OC, which selects a definite target to which the KMS will head for mining.

In the simulator and further course of this project, the implementation and design will focus on the design of the PTCM as the developed technologies and algorithms can be transferred and applied to the other involved modules as well.

III. NAVIGATIONAL CONCEPT FOR DEEP SPACE MISSIONS

The KaNaRiA reference mission scenario envisages four main operational phases according to the mining processes described in section II.I: MRCD (Mapping, Characterization and Resource Determination), REaE (Resource Extraction and Exploitation), Maintenance and Logistics. Each of these phases imposes stringent performance requirements for the navigation subsystems of the various mission elements and their navigation autonomy capabilities. Within this section, the MRCD mission operations timeline is presented. The navigation requirements for PTCM spacecraft during the MRCD phase are discussed. The navigation system design of the PTCM is described and an autonomous navigation concept for interplanetary cruise is introduced.

PTCM Mission Operations Timeline

The operational concept for PTCM spacecraft is built upon the on-board autonomous capability for mission planning. Based on available system status information and collected knowledge about the target asteroid shape and dynamics, the spacecraft shall be able to select between 3 main concepts of operations while approaching an asteroid: an encounter mission and a lander mission with an additional option on re-docking the lander with the orbiter. The operational timeline for the scenarios is depicted in Fig. 1.

In an encounter mission scenario, the PTCM will perform remote sensing of the asteroid from a safe distance during a pre-planned time span. After finalization of the remote sensing campaign the PTCM will continue its course to a second target asteroid.

A lander mission scenario is selected if the asteroid target shows promising results after the remote sensing. The lander is released from the PTCM and uses its steering capabilities for safe landing on the designated landing site. The surface operations include a deep investigation of the asteroid's composition with a Low Frequency Radar as well as a Laser-Induced Breakdown Spectroscopy (LIBS) of the surface material. The data shall be relayed to the PTCM orbiter. The lander steering capabilities enable the performance of hopping or hovering manoeuvres between sample sites of interest. Additionally, the PTCM lander can ascent from the asteroid surface and re-dock to the PTCM orbiter in order to continue its course to a new asteroid. In case the PTCM delta-v capability is insufficient to perform a flight to a follow-up asteroid, the PTCM stays in the orbit around the asteroid and awaits - if profitable - the RF for refuelling.

The navigation system design of the PTCM spacecraft has been developed in order to ensure the spacecraft's capability to determine its location either absolutely in space or relatively to the target throughout all mission phases.

PTCM Navigation Requirements

The KaNaRiA mission concept proposes the deployment of 5-15 medium-size spacecraft, called PTCM, from a cargo control centre located in Sun-bound orbit about 2.8 AU distance from the Sun and 1.8 AU from Earth. At such distances, two-way ground-spacecraft communication delays exceed thirty minutes. Free-space transmission losses are as high as 290 dB in Ka-band, in which future deep-space communication

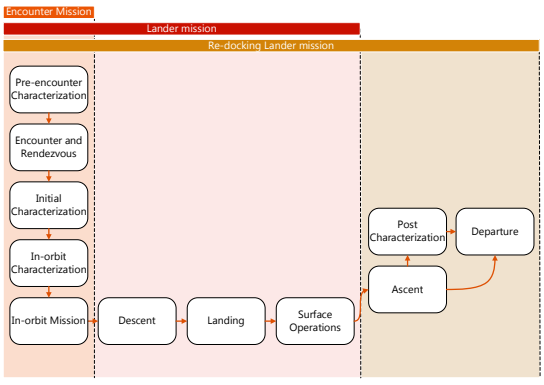


Fig. 1: Operational timeline (from left to right) for PTCM spacecraft,.

infrastructure will operate. The generation of sufficient power to frequently communicate with Earth for tracking purposes, while keeping all subsystems thermally conditioned and performing asteroid characterization operations is not a trivial problem given that the solar flux does not exceed 200 W/m^2. Furthermore, the simultaneous operation of 5-15 missions is a challenge for the already busy tracking and processing schedule of the deep-space ground infrastructure. It is therefore necessary to design the PTCM with a sensible balance between system complexity and self-contained autonomous navigation capabilities.

It has been determined that a PTCM shall be capable of performing on-board orbit determination (OD) at the 100 km precision during cruise in order to support guidance and control during orbit manoeuvring. OD shall be performed fully autonomously without ground support. The stability of the on-board solution shall be guaranteed for a transfer time as long as 4 years. OD updates from ground shall be expected regularly assuming a tracking campaign of maximum 1 week every 5 months.

The PTCM shall be targeted to a rendezvous-plane crossing point between 100 and 500 km from the asteroid surface depending of the volume sphere of influence of the particular object. The 3-σ error ellipsoid at rendezvous condition shall be constraint to 100 m – a requirement that has been fulfilled comfortably by previous asteroid fly-by missions.

During the asteroid in-orbit phase, a thorough characterization of the asteroid surface properties, internal structure as well as landing site selection and mapping will be carried out. During the observation campaigns a position accuracy in the order of meters relative to the asteroid surface shall be achieved.

The landing sequence will consist of a horizontal equalization phase and a subsequent vertical descent. The landing strategy has been designed to ensure soft landing (the survival of the PTCM lander structure), safe landing (safety of landing site avoiding obstacles bigger than 50 cm and slopes higher than 10 degrees) and hazard detection capability up to 10 minutes from touchdown.

Navigation System Design for a KaNaRiA PTCM

The PTCMs have been designed to perform inertial-aided optical navigation throughout all mission phases. In Table 1 a list of the navigation instruments has been provided including their type, mass and primary usage.

Cruise navigation

During cruise the angular observations of planet chords, star-planet and star-Sun angles are combined with the relative Doppler shift of the optical Sun spectra to derive spacecraft position and velocity. The self-contained navigation approach is based on the method proposed by Guo [3] and further investigated by Yim [4].

Instrument	Mass [kg]	Usage
Resonance Scatter Interferometer	42.2	Optical Sun Doppler observations
Coupled Star-Sun tracker	1.98	Stellar attitude and star-planet observations
Fine Sun Sensor	0.65	Coarse Sun attitude
Wide-Angle Camera	2	Asteroid detection and mapping
Narrow-Angle Camera	6	Surface mapping
Lidar Altimeter	3.52	Range finder
3D Lidar	6.5	Asteroid mapping
Space Inertial Reference Unit	7	Inertial position and attitude reconstruction

Table 1: PTCM navigation sensor suite

Spacecraft attitude is reconstructed from the stellar attitude provided by star tracking and from the rate-gyro integration during manoeuvring. Coarse Sun attitude sensors are mounted as back-up solution.

Fig. 2 shows the power flux available from planetary emission and chord lengths of solar system planets in the optical bandwidth as observed by a spacecraft flying a sun-bound circular orbit at 2.8 AU.

Planetary atmospheric and surface albedo has been taken into account. The main selected bodies to be observed for navigation are the Sun, Jupiter and Earth. However other planetary bodies, including the targeted asteroid, are observed when illumination and geometry

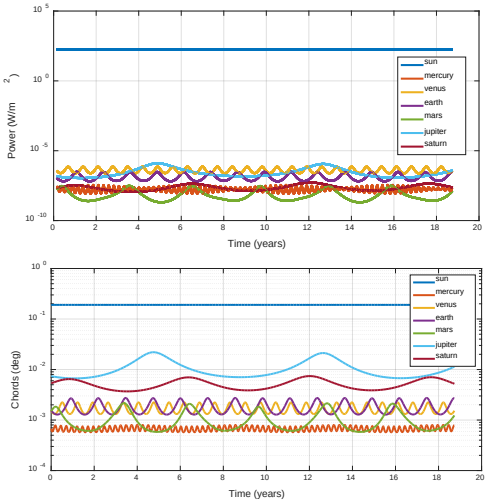


Fig.2: Power flux (top) and angular chord length (bottom) of Sun and solar system planets as observed from a Sun-bound circular orbit with semi-major axis of 2.8 AU.

conditions are favourable.

During cruise, Doppler frequency shift measurements from the Sun optical spectra are used to derive the spacecraft radial velocity. The derived radial velocity measurements are combined with planet chord length angles, planet-star and Sun-star angles. Angular measurements are processed according to standard celestial navigation procedures together with radial velocity measurements in an unscented Kalman filter. A particle filter is simultaneously executed in parallel with timely state updates from the Kalman filter. The particle filter (see section I.V) allows for a robust estimation in mismodelled dynamic environments, as for instance, the vicinity of an asteroid whose gravity field has not been probed. Fig. 3 illustrates the optical cruise navigation system of a PTCM spacecraft.

Optical navigation is aided by means of inertial measurements from the space inertial reference unit during orbit and attitude manoeuvring.

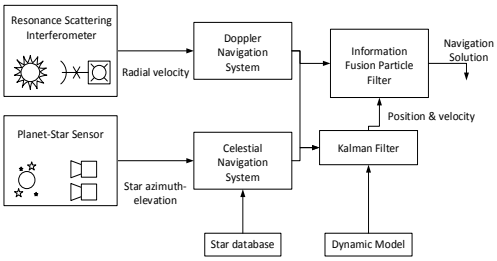


Fig. 3: Integrated celestial and optical Sun Doppler navigation system.

Asteroid relative navigation

In the vicinity of the target asteroid, optical navigation is implemented by means of feature tracking with two optical cameras and a 3D LIDAR. Visual SLAM (simultaneous localization and mapping) is used to reconstruct the asteroid shape and global map, and to locate the spacecraft relative to the generated surface map (see section I.V). A parallel estimation of both, spacecraft state and map, allows for increasing accuracy in the asteroid spin-state knowledge i.e., rotation axis orientation, rotation rate, tumbling modes, etc.

Star trackers are used for stellar attitude reconstruction as long as the asteroid covers between 60 and 80% of the instrument field of view. Rotation-rate measurements are collected from gyros to integrate attitude between stellar-blind phases and during the descent of the PTCM lander.

During descent, the PTCM lander uses a LIDAR altimeter to reconstruct height and vertical speed independently from the main SLAM navigation engine. The LIDAR altimeter solution is fed as input for the collision avoidance decision process handled by the on-board mission planning autonomy.

IV. MULTI-SENSOR FUSION FOR SPACE NAVIGATION

The information fusion subsystem aggregates multi-sensor data and a-priori knowledge to a unified representation, which serves as a basis for cognitive autonomous decision-making (Fig. 4). This bio-inspired model of decision-making relies on perceptions governed by top down as well as bottom up information flows. [5,6]

In particular, the aggregated information is comprised of *i)* top-down a-priori knowledge about the world and the spacecraft as well as *ii)* bottom-up perceived knowledge, which consists of fused data from multiple sensors. In conjunction, this information results in an estimate of the current spacecraft and environment state.

The multi-sensor fusion and state estimation solves the versatile challenges posed by the different mission phases (see section II) within one framework. Throughout all mission phases, a particle filter is used to approximate the desired probability distribution.

In the interplanetary cruise phase, the distribution $p(\boldsymbol{x}_t|\boldsymbol{z}_{0:t}, \boldsymbol{u}_{1:t})$[‡‡] over the current spacecraft state

$$\boldsymbol{x}_t = [\boldsymbol{r}_t^T, \boldsymbol{q}_t^T, \dot{\boldsymbol{r}}_t^T, \dot{\boldsymbol{q}}_t^T, \ddot{\boldsymbol{r}}_t^T, \ddot{\boldsymbol{q}}_t^T\ ]^T$$

given all measurements $\boldsymbol{z}_{0:t}$ and controls $\boldsymbol{u}_{0:t}$ is estimated in a heliocentric reference frame, where $\boldsymbol{r}_t$ is the position, $\boldsymbol{q}_t$ the attitude, $\dot{\boldsymbol{r}}_t$ the velocity, $\dot{\boldsymbol{q}}_t$ the angular velocity, $\ddot{\boldsymbol{r}}_t$ the acceleration, and $\ddot{\boldsymbol{q}}_t$ the angular acceleration of the spacecraft. $\boldsymbol{z}_t$ contains measurements from the interferometer, the coupled Sun-star tracker and the wide-angle camera (see section III).

In the MCRD phase, the camera suite and the mapping LIDAR are able to perceive the asteroid. This enables the multi-sensor fusion module to estimate a map Y of the approached asteroid. This provides a

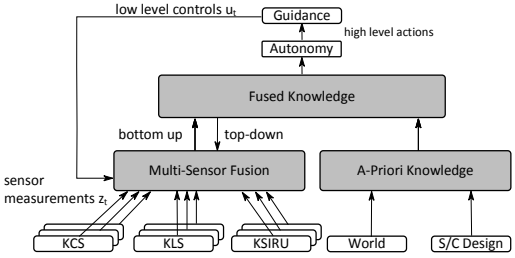


Fig. 4: Knowledge acquisition process for cognitive autonomous decision-making.

[‡‡] For convenience reasons we use $\boldsymbol{a}_{0:t}$ as a short notation for a time series of variables $\boldsymbol{a}_0, \boldsymbol{a}_1, \ldots, \boldsymbol{a}_t$.

physical description of the asteroid and, even more essential, can be used as a reference for relative spacecraft state estimation.

Although the two tasks of localization and mapping can be solved separately, they are not independent of each other. It is a joint estimation problem commonly known as Simultaneous Localization and Mapping (SLAM) [7] (Fig. 5). However, using a conditional independence assumption, the corresponding joint probability distribution can be factorized into one conditional distribution over the trajectory $\boldsymbol{x}_{0:t}$ and one over the map Y:

$$p(\boldsymbol{x}_{0:t}, Y|\boldsymbol{z}_{0:t}, \boldsymbol{u}_{1:t}) = \underbrace{p(\boldsymbol{x}_{0:t}|\boldsymbol{z}_{0:t}, \boldsymbol{u}_{1:t})}_{Trajectory}\underbrace{p(Y|\boldsymbol{x}_{0:t}, \boldsymbol{z}_{0:t})}_{Map}.$$

This allows us to use a technique called Rao-Blackwellization. [8] In the first step, the distribution over the trajectory is approximated by the particle filter [9] using controls, measurements and map estimate. In the second step, the current state is assumed to be known and the distribution over the map is computed analytically.

Initially, a landmark-based map is estimated in order to establish robust relative navigation in an asteroid-centric reference frame. The landmarks will be extracted by performing bio-inspired feature detection and description using Intrinsic 2 Dimensional (I2D) features [6,10] on the images obtained by the on-board cameras and with the distance information provided by the LIDAR instruments.

When the landmark map has full coverage and allows for a robust localization, it is extended by a belief-function-based grid-map of the asteroid in the proximity operations phase. It divides the volume into discrete grid cells where each grid cell represents an estimate of a corresponding piece of the physical environment. While the uncertainty regarding the true state is usually represented by a Bayesian probability, we are using belief functions [11,12] here, which allow to assign probability mass not only to the singletons $a \in \Theta$ of a hypothesis space Θ but also to all subsets of the power set $A \subseteq \wp(\Theta)$ including the superset Θ and the empty set $\emptyset$. This approach makes different dimensions of uncertainty explicit. E.g. a full lack of evidence is expressed by assigning all mass to Θ while conflicting evidence is expressed by mass assigned to $\emptyset$. In the Bayesian probability framework, both cases would result in an equal distribution and would be therefore undistinguishable. There are several works on mapping using belief functions [13,14,15] while a belief-function-based SLAM approach as a generalization of the successful grid-map based FastSLAM [16] algorithm was presented by Reineking and Clemens. [17] This approach was already applied in the context of extra-terrestrial exploration. [18,19]

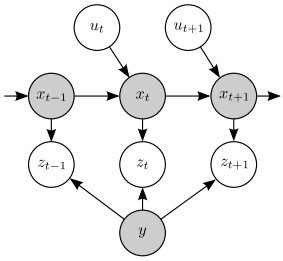


Fig. 5: Bayesian Network depicting the SLAM-problem.

The combination of belief functions and a grid map allow for *i)* a finer representation of the physical environment and *ii)* a better representation of the cognitive uncertainties. [20] This in turn enables the autonomy to pursue advanced exploration strategies to actively investigate possible landing sites, with respect to commodities, hazardous areas and fuel consumption. Based on the uncertainty information in the maps (grid-map as well as landmark based) the autonomy can be provided with desired actions with respect to every navigation instrument. Thus, particular actions can be assessed for their expected information gain.

V. OPTIMAL TRAJECTORY PLANNING

Trajectory planning for deep space missions is a topic of great interest. Mathematical fields like optimization and optimal control can be used to realize autonomous missions while protecting resources and making them safer. A perturbed *optimal control problem* (OCP(p)) has the form

$$\begin{aligned} \min_{x,u} F(x,u,p,t) &:= g(x(t_f),t_f) + \int_0^{t_f} f_0(x(t),u(t),t,p)dt \\ s.t. \quad \dot{x}(t) &= f(x(t),u(t),t,p) \\ x(0) &= x_0 \\ \Psi(x_0,x(t_f),p) &= 0 \\ C(x(t),u(t),t,p) &\leq 0 \end{aligned}$$

with F being the objective function depending on the state $x(t)$ at time $t \in [0,t_f]$, the vector p describing model perturbations and the control function $u(t)$ by which the system's dynamic f can be influenced via differential equations. The control u has to be chosen in such a way that the constraints C as well as the initial and terminal conditions Ψ are fulfilled while minimizing the objective function F.

In principle, there exist two ways to solve an OCP(p), the so called indirect and direct methods. The indirect methods are being studied since several decades and need advanced skills regarding optimal control theory. Some algorithms are described in Bürlisch [21], Deuflhard [22], Ho and Bryson [23] as well as Miele [24]. The direct approach transcribes the infinite-dimensional OCP(p) into a finite-dimensional

non-linear optimization problem (NLP(p)) via discretization of states and controls. [25,26] An NLP(p) consists of an objective function *F* and constraints *G*:

$$\begin{aligned} \min_{z} \quad & F(z,p) \\ s.t. \quad & G_i(z,p) = 0,\ i = 1,\dots,M_e \\ & G_i(z,p) \le 0,\ i = M_e+1,\dots,M \end{aligned}$$

The objective function F depends on the optimization vector $z := (x_1^T, \dots, x_{N_t}^T, u_1^T, \dots, u_{N_t}^T)$ with $x_i, u_i, i = 1, \dots, N_t$ representing the former x and u at discrete time points $0 = t_1 < t_2 < \cdots < t_{N_t} := t_f, x_i \approx x(t_i), u_i \approx u(t_i)$ and the perturbation vector p. For a fixed parameter $p = p_0$ an optimal solution is called the *nominal* or *undisturbed solution* indicated by $z(p_0)$.

The OCP(p) formulation's dynamic model describes the movement of the spacecraft due to main gravitational influences of the sun and other planets as well as the thrust commands through ordinary differential equations (ODEs):

$$\dot{x} := \begin{pmatrix} \dot{p}_{sc} \\ \ddot{p}_{sc} \\ \dot{m}_{sc} \end{pmatrix} = \begin{pmatrix} \dot{p}_{sc} \\ \sum_{i \in I} \mu_i \cdot \frac{r_i}{\|r_i\|_2^3} + \frac{T}{m_{sc}} \\ -\frac{\|T\|}{g_0 I_{sp}} \end{pmatrix}$$

Herein p_{sc} is the position vector of the spacecraft, $\mu_i, i \in [Sun, Mars, Jupiter, Saturn]$ is the gravitational constant of the according celestial body and r_i the direction vector between spacecraft and body, $T = [u_x, u_y, u_z]$ is the thrust vector, m_{sc} the spacecraft's recent mass, I_{sp} its specific impulse and g_0 the gravitational constant of Earth.

Within the optimization there exist several methods to solve such ODE systems. One is the so-called full discretization, where all states and controls are calculated for a chosen number of discrete time points. An alternative is to use *multiple shooting methods*. Here the solution space is divided into several sections by so-called *multi-nodes* and for each section a *single shooting method* is applied. [27] It is sufficient to combine the sections by additional constraints in order to gain the correct solution in the end. In the KaNaRiA implementation the position of the multi-nodes is let free for optimization.

These methods will be investigated to achieve a robust and efficient optimization for each of the systemically different navigation phases of a space mission. The resulting non-linear high-dimensional optimization problems are solved using the software package WORHP [28] ('We Optimize Really Huge Problems'). This is especially efficient for solving high-dimensional problems like those resulting from the discretization of optimal control problems as it uses for example the sparsity information of the derivative matrices.

Additionally, an on-board-capable *parametric sensitivity* and *stability analysis* of optimal nominal solutions towards perturbations will be performed in KaNaRiA. Perturbations are for example deviations in the assumed amount of left over fuel, the magnitude of the solar pressure or the asteroid's gravitational influence, which may have a great impact on the practicability of a planned trajectory. Changes in the optimal solution of the undisturbed problem in case of deviating values p from nominal values p_0 can be estimated by calculating the solution vector

$$z(p) \approx z(p_0) + \frac{dz}{dp}(p_0)(p - p_0)$$

while only the nominal solution $z(p_0)$ and its sensitivities $\frac{dz}{dp}(p_0)$ need to be computed.

Whereas offline calculations of optimal trajectories allow for their investigation, a practical online-realization can only be achieved through special *real-time capable methods*. Based on the parametric sensitivity analysis and dependent on the different phases of a space mission and their special claims different trajectory optimization and real-time tracking strategies will be developed for differing time scales. When approaching the asteroid further and especially when entering the landing phase the challenges of efficient real-time capable control interventions increase due to the weak, inhomogeneous gravity field resulting from the relative small mass, irregular form and unknown rotation of the asteroid.

Implementation:

A simple way to achieve an orbit transfer is the Hohmann transfer orbit, but it is only applicable under strong constraints. That is why in KaNaRiA another approach was chosen. For the cruise phase a maximum of three thrust commands may be applied, one at the beginning of a trajectory, one at the end and one at an optimized time point in between. These commands are sufficient regarding the long time frame of the flight without serious perturbation forces. To model impulsive thrusting more accurately an application-adapted model is developed. By using the objective function

$$F = w_{tf} t_f - m_f (1 - w_{tf})$$

with t_f being the total flight time, m_f the spacecraft's final mass and $w_{tf} \in [0,1]$ a weighting factor where any fit between time- and energy-optimization can be chosen. The start mass of the spacecraft is 4000 kg, the fuel mass 1500 kg, the I_{sp} 318 seconds and the thrust is limited to 340 to 440 Newton. The optimization was

performed considering the influences of the planets Mars, Saturn and Jupiter. The boundary condition was meeting the position and velocity of the asteroid within a certain range sufficient for the cruise phase. The solutions for full time and full energy optimization can be seen in Fig. 6 and Table 2. With 2157.56 kg of fuel consumption and a total flight time of 796.747 days the flight of the energy-optimal trajectory needs 125.55 kg of fuel less but 30.493 days longer than flying the time-optimal trajectory (Table 2). The energy-optimal trajectory contains two thrust commands whereas the time-optimal trajectory consists of three thrust commands in order to meet the objective. This way in order to meet the energy-optimal objective, the spacecraft might orbit on the original trajectory before thrusting for the first time. The changes in the z-position differs the most since changing the inclination of a trajectory is highly energy consuming. In comparison to the x-/y-positions, the thrusts lead to only a small adjustment in the z-position. For both trajectories the last thrust is applied at the end of the trajectory, whereas only the time-optimal trajectory has a thrust at the beginning of the manoeuvre (Fig. 6).

The solution trajectories show strong differences according to the chosen objective priorities which means being able to save a lot of mission time or fuel consumption according to the mission's needs and allowing for various and considerably different autonomous decisions.

VI. COGNITIVE SPACECRAFT AUTONOMY

The autonomy module is the central component for autonomous reasoning and decision-making regarding normal mission operation as well as emergency

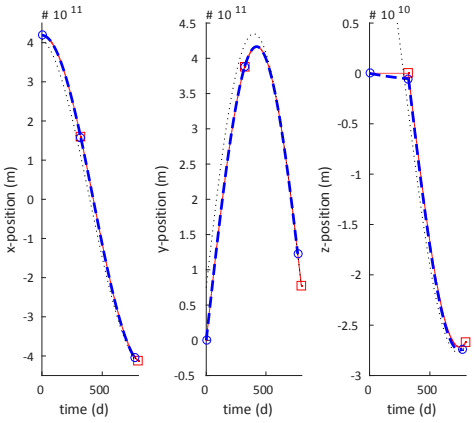


Fig. 6: X-, y- and z-position in meters of time-optimal (dashed blue line) and energy-optimal (solid red line) trajectory over time in days. The dotted black line shows the asteroid's position. Circles (time-optimal) and squares (energy-optimal) show the time points of the thrust commands.

	$w_{tf} = 0$	$w_{tf} = 1$	diff
Opt. criterion	energy	time	
Flight time (d)	796.747	766.254	30.493
Fuel (kg)	2157.56	2283.11	125.55
Line color (Fig. 6)	red	blue	

Table 2: Optimization criterion, flight time in days and fuel consumption in kg for two different mission trajectories.

situations. It controls all sub-modules of the spacecraft and all processes related to reasoning, plan generation, plan evaluation, plan execution and FDIR during all phases of the mission.

During normal mission operation, the autonomy module monitors both phase-specific mission objectives and the current state of the spacecraft. Based on these it generates plans to either achieve primary (e.g. locate the asteroid using optical sensors, maintain a stable orbit around the asteroid, perform the docking/landing operation) or secondary objectives (e.g. calculate alternative trajectory to further increase information on possible landing site). As the scenario is of a highly dynamic nature, the system periodically requests re-evaluation of plans to check whether they are still applicable. The appropriate strategy for re-evaluation is based on current system resources and time constraints. The autonomy module has to decide on and ensure commitment to one plan, yet retain the option to reconsider the commitment at a later point – when new information becomes available.

Uncertain knowledge resulting from incomplete or incorrect data poses a central challenge to reasoning and decision making, therefore the system has to consider these kind of uncertainties in the decision making process. Based on the biologically inspired principle of information maximization, the autonomy module seeks to minimize and resolve these uncertainties by employing information gain strategies and active perception to extend and improve the amount and quality of the available knowledge.

As autonomous handling of emergency situations is vital, the module utilizes FDIR algorithms to react to anomalies as they are detected, by reprioritizing primary and secondary mission objectives as well as planning and executing appropriate fault-detection, fault-isolation and fault-recovery plans.

Situation Analysis and Evaluation

To create a basis for decision-making and plan generation, the current state of the spacecraft and all information available to it has to be analysed and evaluated. This includes a-priori knowledge (spacecraft configuration, mission phase specific objectives), internal data (navigation variables, fuel, mass, health status) and external data (sensor measurements, asteroid properties, potential targets). Sensor information from

optical cameras, an imaging LIDAR and a LIDAR altimeter are provided by the sensor fusion [5] to the autonomy module and combined to create maps that assign potentially hazardous areas, points of interest and potential landing sites to regions on the asteroid. In addition, boundary conditions for trajectory requests regarding different mission phases and actions are added.

Plan Generation, Assessment and Execution

During plan generation, the system decomposes high-level objectives into a sequence of actions. These are selected from a dynamic set of currently available actions and based on the current beliefs of the spacecraft. At the atomic level, actions can be executed by the spacecraft actuators, which include spacecraft propulsion, reaction wheel control, and communication with other entities, sensor control and deployment of other vehicles (PTCM orbiter and lander). As the environment is dynamic, objectives can become unachievable and thus plans can become obsolete. The autonomy must be able to assess whether a given plan is still feasible and react accordingly.

Attitude and Sensor Control

To fulfil phase specific mission objectives that require distinct sensor and actuator alignments, the autonomy module has to provide an attitude control sequence based on both proposed priority rankings of measurable information and communication requirements.

This attitude control sequence is based on a previously calculated trajectory, where a trajectory is represented as a sequence of positions and time points. This sequence is split into segments at the control points of the trajectory. For each of these segments a spacecraft orientation is calculated for which all available sensors potentially provide the best measurements with respect to the maximisation of gained information.

From these orientations along the trajectory, the required attitude controls can be determined. Taking into account the potential information gain and hazards along this path, a sensor control plan for the trajectory is calculated, which specifies the sensor activation and deactivation at all time points.

Autonomous FDIR

To enable the system to autonomously perform fault detection, isolation and recovery (FDIR), current knowledge about the spacecraft and the world is used to infer about possible erroneous states. Algorithms for anomaly detection are utilized to determine unusual world- or spacecraft state configurations (e.g. conflicting datasets, unusual high uncertainty) that indicate a hard- or software problem. These are analysed regarding fault-identification and fault-recovery. If available, information on error-models of sensors and probabilities for different error scenarios will be incorporated in this analysis. If one or more recovery strategies exist, the necessary actions to be performed and possible constraints on the further action selection and plan generation (e.g. an actuator ceased to function) will be evaluated. In addition findings of this analysis are provided to the sensor fusion to enable this module to adapt the corresponding sensor models accordingly.

VII. MASSIVELY PARALLEL AND PHYSICALLY-BASED SIMULATION

In this section, we highlight two key aspects of KaNaRiA's simulation software. First, we give an overview of our simulation software with a focus on its novel approach to concurrency control management. Second, we will present the challenges for our novel concept of gravity field simulation for irregularly shaped celestial bodies.

Realistic spacecraft simulations have to cover all aspects of a mission scenario in real-world detail. Internal spacecraft components, the space environment with its physical forces and disturbances, the sensor data acquisition chain, and the spacecraft actuator and propulsion systems have to be modelled and simulated.

One key aspect of such simulations is the validation and testing of specific performance aspects (e.g. navigation algorithms), enabling sophisticated analyses for engineers that would otherwise be impossible. These analyses (e.g. spacecraft landing procedure performance) require comprehensive simulation and the monitoring of vast amounts of generated data.

In recent years, simulation has emerged as a key technology for improving and streamlining the conceptualisation and design of vehicles by simulation in "virtual testbeds". [29,30] Virtual testbeds are constituted by a sophisticated physically-based simulation of both the vehicle and its designated environment, as well as real-time, immersive rendering and 3D interaction techniques. These testbeds give engineers the opportunity to interact with the simulated vehicle in order to gain comprehensive understanding of possible design flaws as early as possible during the design process. [29,31]

Consequently, the main challenge of such virtual end-to-end simulations for space missions is real-time simulation with highly responsive interactivity while maintaining realistic physical models. In this context, an enormous amount of software components is working in order to simulate both, spacecraft behaviour and required input data. Additionally, spacecraft engineers would, ideally, have the ability to easily manipulate parameters of the spacecraft(s), change aspects of space

environment such as disturbances, add or remove sensors or other spacecraft components, and interactively test the spacecraft(s) under a variety of conditions.

In order to achieve the above stated software requirements, we have proposed and implemented the KaNaRiA virtual simulation (KVS) [32], which proposes an easily adaptable and customizable massively-parallel virtual reality system architecture with a centralized software infrastructure to attain real-time performance of the overall simulation.

Consequently, KVS enables the analysis and testing of autonomous spacecraft operation, spacecraft navigation algorithms, and spacecraft subsystems in an enriched, virtual world. It leverages physics-based spacecraft models in conjunction with high-quality, multimedia visualization and immersive interaction techniques to form an intuitive, accurate engineering tool.

KVS has been designed to take advantage of an open source game engine targeted at the video game industry. Thus, KVS is able to bridge the gap between traditional, high-fidelity analysis tools [33] and graphically realistic, immersive, and interactive simulations.

Some of the highlights of KaNaRiA's virtual simulation include:

- Real-time 3D rendering of complex space environments & spacecraft models
- Real-time simulation of spacecraft subsystems, sensors as well as actuators
- The ability to observe internal spacecraft data intuitively
- Controlled, repeatable testing for advanced simulations
- Intuitive and consistent user interface.

Rendering, internal multi-component spacecraft simulation, and interaction with the overall system happens completely in parallel in KVS. To avoid any latency between those parallel software components, KVS uses our novel concurrency control management (CCM) for wait-free data exchange, with its core being a global hash map, called key-value pool (*KVPool*, Fig. 7). [34,35] The KVPool is a centralized data storage that maintains the complete shared world state of the simulation without being a traditional, heavy-weight database.

Every simulation aspect, such as spacecraft subsystems, sensors, actuators, and any physical models are implemented as *entities*, which can access the KVPool. Other software components can access the data by simply passing the key to the KVPool. The wait-free behaviour of KVS's KVPool results in a dramatic speed-up of several orders of magnitude compared to traditional lock-based approaches (see Fig. 8), while

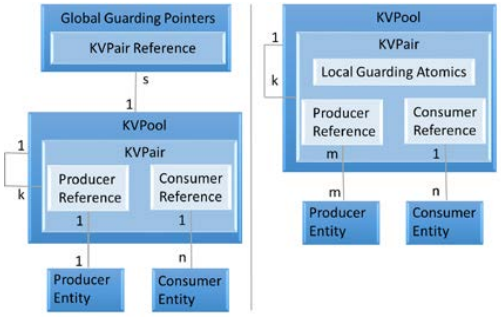


Fig. 7: Entity-relationship diagram of wait-free access to the shared simulation state by global guards (left) [34] and local guards (right). [35]

avoiding all their problems like deadlocks or thread starvation. Moreover, it overcomes the well-known many-to-many interface problem of the data-flow-based approach found in many traditional VR system architectures.

Furthermore, KVS's software infrastructure facilitates automatic code generation for virtual testbeds via domain specific modelling. [29] In addition, it can also be used for other data-driven simulation domains such as multi-agent-systems. [29]

Testing navigation and autonomous guidance algorithms for landing and orbiting an asteroid, under micro-g or milli-g gravity fields is crucial for developing fail-safe landing procedures. Therefore, KVS has to simulate a realistic gravity field around an asteroid for a given shape model (polygonal mesh) and density distribution at every point in space. We aim for fast and accurate computation of gravitational fields for any given asteroid. Currently spherical or ellipsoid harmonics approaches are the computationally least inexpensive compared with other approaches.

However, spherical harmonics series diverge within the Brillouin sphere [36] (see Fig. 9); hence, the gravitational field computed close to the surface of an asteroid is inaccurate. [37] This results in incorrect simulated gravitational forces acting on the spacecraft during landing phase.

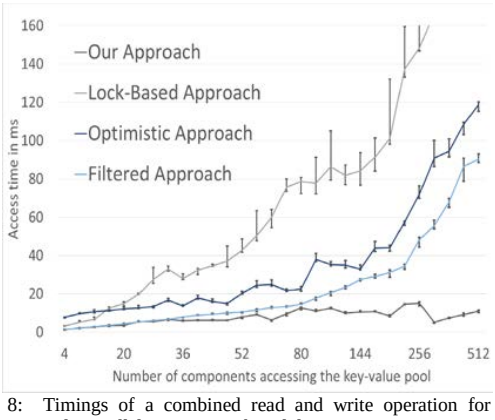


Fig. 8: Timings of a combined read and write operation for massively parallel access to a shared data structure.

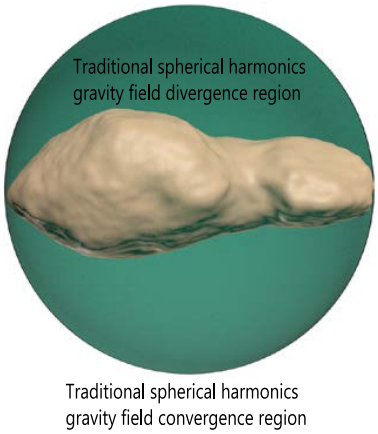


Fig. 9 : Brillouin sphere of asteroid Toutatis.

Takahashi et al. [38] overcame this issue with interior gravity field approach and alternatively with the interior spherical Bessel gravity field model. [39]

However, the former approach is computationally expensive as different sets of interior spherical harmonic coefficients have to be computed separately for each and every point on the asteroid surface, these different sets of coefficients are only applicable for gravity field computation within their respective interior sphere touching the respective point. [38] On the other hand, the pre-processing in the latter approach is computationally very expensive, which is not suitable for our purposes, since we need to be able to compute the field for any asteroid during runtime of the simulator. In our case, we generate the asteroids' shape models and density distributions procedurally in order to test guidance algorithms for landing on different types of asteroids (with respect to shape and density distributions) as well as sensor fusion algorithms for navigation. Therefore, we are currently working on an approach that computes the gravitational field of an asteroid in real-time while maintaining fast pre-processing. In our approach, we basically compute a sphere packing of a shape model of given asteroid using the modified protosphere algorithm from Weller et al. [39,40] with constraint on the radiuses of spheres based on the known prior asteroid density distribution. This method produces uniform density spheres that can be considered as point masses, then computing gravitational potential/acceleration at any given point is a trivial scalar/vector summation of potentials/accelerations applied by each sphere at that point (see Fig. 10). The sphere packing generation and summation computation for gravitational potential/acceleration are parallelizable. Hence, the pre-processing and gravity field computation are fast, which are suitable for our computational demands.

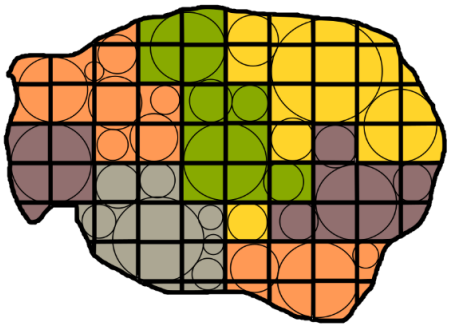


Fig. 10: Sphere packing of an asteroid with density distribution constraint on radius. The colours indicate different densities inside the asteroid.

VIII. CONCLUSION AND OUTLOOK

The KaNaRiA project focuses on the development of new autonomous decision-making, navigation, sensor fusion and guidance methods implemented on a virtual spacecraft. Within the spacecraft design, an autonomy module serves as the central controlling unit managing the data obtained and created by the navigation, fusion and guidance, generating and executing plans as well as controlling the attitude.

As an application for the development of this approach, an asteroid mining mission concept was developed. It considered asteroid mining as a long-term space activity. As the initial mission analysis led to the decision of a parking orbit located in the asteroid main belt, all involved mission elements need advanced autonomy strategies for navigation and guidance as well as mission planning and operations. During the KaNaRiA project, the design and application of the autonomous strategies focuses on the PTCMs consisting of an orbiter and a re-docking lander designed for a multi-rendezvous asteroid mission.

The navigation concept was designed for the PTCM operations timeline. It enables a thorough characterization of the asteroid surface properties as well as mapping including landing site selection with an envisaged position accuracy of several meters. The instrument suite to perform inertial-aided optical navigation under the imposed constraints by the mission concept was presented and the methods to conduct the observations were introduced. For cruise navigation, the navigation system design uses angular observation of planet chords combined with star-Sun relative Doppler shift to obtain the spacecraft position and velocity. The main selected bodies are Jupiter, the Sun and Earth with the option to consider other planets depending on their illumination conditions. The spacecraft radial velocity is calculated using the optical Doppler frequency shift measurement from the Sun and combined with planet chord length angles, planet star and Sun-star angles to

determine position. The latter measurements are processed in an unscented Kalman filter whose results are used for timely state updates for the particle filter running in parallel. The Kalman filter ensures a robust estimation for mismodelled dynamic environments (e.g. vicinity of asteroids) such as an unprobed gravity field. For the purpose of relative navigation, two optical cameras and 3D LIDAR are used for feature tracking, optical navigation and independent height and vertical speed reconstruction during descent.

Based on the data obtained by the navigation sensor suite, the multi-sensor fusion subsystem provides all necessary information for cognitive autonomous decision-making. The data is obtained by a particle filter-based SLAM approach with a combination of a landmark-based map with a belief-function-based grid-map. The spacecraft dynamic state and the corresponding maps of the asteroid are estimated with a level of detail corresponding to the respective mission phase. This approach is applicable in every exploration scenario where an autonomous agent has to estimate its own position in an unknown environment and map it at the same time. Furthermore, the uncertainties encoded in the map enable an autonomous system to take cognitive decisions.

The challenge of finding the right interplanetary trajectory is solved using optimal control methods from the mathematical field. In KaNaRiA, the implemented approach allows a maximum of three thrust commands, one at the beginning, one at the end and one at an optimized time point in between. A weighting factor allows a customized fit between time- and energy-optimization. Using the optimal nominal solution as baseline, a parametric sensitivity analysis towards perturbations will be performed. Based on the parametric sensitivity analysis and according to the need for optimality, robustness and calculation time at hand, three real-time capable optimal control methods will be implemented: a method for model-predictive control (MPC), a method for repeated adjustment and an optimal feedback controller. Additionally, the approach of modelling the spacecraft motion will be applied to the task of navigation on the asteroid's surface to investigate an adaptive autonomous consideration of state-space constraints.

Analysing and evaluating the data obtained by navigation, fusion and guidance as well as other information available, the autonomy module assesses the current state of the spacecraft. The module acts as central component for autonomous reasoning and decision-making. The situation assessment is used as input for the decision on the feasibility of applicable mission objectives. Mission objectives are broken down into a sequence of actions, which are used to generate a plan. Due to a dynamic environment, the objectives could become unachievable depending on the spacecraft or environment state. With a changing environment, periodic requests of plan re-evaluations are necessary to either ensure commitment or reconsideration. The autonomy module also takes into account the uncertainty of the obtained knowledge using biologically inspired principles such as information maximisation and active perception. Finally, the execution of the plan is based on the trajectory optimization of the guidance subsystem. Using given time and control points, the actuators can be commanded. Based on the known attitude, a sensor control plan can be generated to specify their de-/activation schedule. Last but not least, erroneous states are inferred from the current knowledge of the spacecraft and world state utilizing anomaly detection algorithms for FDIR. All in all, the methods and algorithms developed in this project can be used to enhance the level of autonomy of future space missions with regards to navigation, plan generation, action selection and FDIR. The system provides the ability to represent uncertainty and incorporate this knowledge into the plan generation step. It can modify existing plans to include utility objectives aiming on reducing uncertainty and therefore enhances the robustness of the system with respect to unexpected situations.

The developed autonomy and navigation methods and algorithms are tested and verified in the KaNaRiA virtual simulator (KVS) using the mission scenario of asteroid mining as application. The KVS uses our novel concurrency control management approach with wait-free data exchange between various software components. A centralized data storage called KVPool is used, which resolves the many-to-many interface problem typically encountered in traditional VR architectures. This wait-free approach outperformed standard approaches in terms of access time as shown in the Fig. 8. The above software infrastructure can also be applied in other data-driven simulation domains. Currently, we are experimenting on a new approach for generating gravity field of asteroid shape models, which is based on sphere packing method [32]. This approach considers variable densities and overcomes the gravity field divergence problem in the Brillouin sphere region (see Fig. 9). However, at the same time our method also focuses on a fast computation of gravity potential and acceleration, and on fast generation of pre-processed data used for computing gravity fields.

The KaNaRiA project had its project kick-off in October 2013 and is designated for a period of four years.

ACKNOWLEDGMENTS

This work has been carried out in the frame of the project KaNaRiA. KaNaRiA is a project financed by the German Ministry of Economy and Energy through the

German Aerospace Centre, Space Administration (DLR, Deutsches Zentrum für Luft- und Raumfahrt, FKZ 50NA1318 and 50NA1319). It is a collaboration of the University of Bremen and the Bundeswehr University Munich. The involved institutes as well as their contribution in the paper and the respective point of contacts are listed below:

Section II. Alena Probst, ISTA, Bundeswehr University Munich

Section III. Graciela González Peytaví, ISTA, Bundeswehr University Munich

Section IV. David Nakath, Cognitive Neuroinformatics, University of Bremen

Section V. Anne Schattel, Optimization and Optimal Control, University of Bremen

Section VI. Carsten Rachuy, Cognitive Neuroinformatics, University of Bremen

Section VII. Patrick Lange, Computer Graphics and Virtual Reality, University of Bremen

REFERENCES

[1] R.C. Moeller et al., "Space Mission Trade Space Generation and Assessment using the JPL Rapid Mission Architecture (RMA) Team Approach", in *IEEE Aerospace Conf.*, 2011.

[2] A. Probst et al., "Reference Mission Scenario Selection for Main Belt Asteroid Mining Missions", in *Planetary and Terrestrial Mining Science Symp. (PTMSS)*, CIM 2015 Conv., May 10-13, Montréal, Quebec, 2015.

[3] Y. Guo, "Self-contained autonomous navigation system for deep space missions", in *Spaceflight Mechanics*, 1999, pp. 1099-1113.

[4] J. Yim et al., "Autonomous orbit navigation of interplanetary spacecraft", in *AIAA Astrodynamics Specialist Conf.*, Reston, Virginia, 2000.

[5][IV.4]D. Nakath et al., "Active Sensorimotor Object Recognition in Three-Dimensional Space", in *Spatial Cognition IX*, Springer International Publishing, 2014, pp. 312-324.

[6] K. Schill et al., "Scene analysis with saccadic eye movements: top-down and bottom-up modelling", in *Journal of electronic imaging*, Vol. 10(1), 2001, pp. 152-160.

[7] H. Durrant-Whyte and T. Bailey, "Simultaneous localization and mapping: part I", in *IEEE Robotics & Automation Magazine*, 13(2), 2006, 99-110.

[8] A. Doucet et al. "Rao-Blackwellised particle filtering for dynamic Bayesian networks", in *Proc. of the 16th Conf. on Uncertainty in Artificial Intelligence (UAI)*, Morgan Kaufmann Publishers Inc., 2000.

[9] M. Montemerlo and S. Thrun, "FastSLAM 2.0. FastSLAM: A Scalable Method for the Simultaneous Localization and Mapping Problem", in *Robotics*, 2007, pp. 63-90.

[10] T. Reineking et al., "Adaptive Information Selection in Images: Efficient Naive Bayes Nearest Neighbor Classification", in *16th Int. Conf. on Computer Analysis of Images and Patterns (CAIP)*, 2015 (in press)

[11] G. Shafer, "A mathematical theory of evidence", in *Princeton: Princeton university press*, Vol. 1, 1976.

[12] P. Smets and R. Kennes, "The transferable belief model", in *Artificial Intelligence*, Vol. 66(2), 1994, pp. 191-234.

[13] D. Pagac et al., "An evidential approach to map-building for autonomous vehicles", in *IEEE Robotics and Automation*, Transactions on 14.4, 1998, pp. 623-629.

[14] M. Ribo and A. Pinz, "A comparison of three uncertainty calculi for building sonar-based occupancy grids". in *Robotics and Autonomous Systems*, Vol. 35.3 , 2001, pp. 201-209.

[15] J. Mullane et al., "Evidential versus Bayesian estimation for radar map building", in *9th IEEE Int. Conf. on Control, Automation, Robotics and Vision (ICARCV'06)*, 2006.

[16] D. Hahnel et al., "An efficient FastSLAM algorithm for generating maps of large-scale cyclic environments from raw laser range measurements", in *Proc. IEEE/RSJ International Conf. on Intelligent Robots and Systems*, Vol. 1., 2003.

[17] T. Reineking, T. and J. Clemens, "Evidential FastSLAM for grid mapping", in *IEEE 16th Int. Conf. on Information Fusion (FUSION)*, 2013, pp. 789-796.

[18] J. Clemens and T. Reineking, "Multi-Sensor Fusion Using Evidential SLAM for Navigating a Probe through Deep Ice", in *Belief Functions: Theory and Applications*, Springer International Publishing, 2014, pp. 339-347.

[19] H. Niedermeier et al., "Navigation system for a research ice probe for antarctic glaciers", in *IEEE / ION Position, Location and Navigation Symp. (PLANS)*, May 2014, pp. 959-975.

[20] T. Reineking and J. Clemens, "Dimensions of Uncertainty in Evidential Grid Maps", in *Spatial Cognition IX*, Springer International Publishing, 2014, pp. 283-298.

[21] R. Bulirsch, "Die Mehrzielmethode zur numerischen Lösung von nichtlinearen Randwertproblemen, Technical Report of Carl-Cranz-Gesellschaft e.V., Oberpfaffenhofen, 1971", reprint: *Department of Mathematics, Munich University of Technology*, Germany, 1993.

[22] P. Deuflhard, "A modified Newton method for the solution of ill-conditioned systems of nonlinear equations with applications to multiple shooting", in *Numerische Mathematik*, Vol. 22, 1974, pp. 289-315.

[23] Y. Ho and A.E. Bryson, “Applied Optimal Control Optimization, Estimation and Control”, in *Holsted Press Book*, 1975.

[24] A. Miele, “Gradient algorithms for the optimization of dynamic systems”, in *C.T. Leondes 17*, 1980, pp. 1-52.

[25] C. Büskens, “Direkte Optimierungsmethoden zur numerischen Berechnung optimaler Steuerprozesse”, Diploma Thesis, University of Münster, 1993.

[26] C. Büskens and H. Maurer, “Real-Time Control of an Industrial Robot Using Nonlinear Proramming Methods”, in *Proc. of the 4th IFAC Workshop on Algorithms and Architectures*, Vilamoura (Portugal), 1997.

[27] J. Stoer and R. Bulirsch, Introduction to Numerical Analysis”, in *Springer-Verlag*, New York, 1980.

[28] C. Büskens and D. Wassel, “The ESA NLP Solver WORHP”, in *Modeling and Optimization in Space Engineering*, J. D. Pintér (Hrsg.), Springer Optimization and Its Applications, Springer Verlag, Vol. 73, 2013.

[29] P. Lange et al., “Multi Agent System Optimization in Virtual Vehicle Testbeds”, in *8th EAI Int. Conf. on Simulation Tools and Techniques (SIMUtools)*, Athens, Greece, 2015.

[30] M. Cohrs et al. “A Methodology for Interactive Spatial Visualization of Automotive Function Architectures for Development and Maintenance”, in *Int. Symp. on Visual Computing (ISVC)*, Crete, Greece, 2013.

[31] A. Gomes de Sa and G. Zachmann, “Virtual Reality as a tool for Verification of Assembly and Maintenance Processes”, in *Journal of Computer graphics*, 1999, pp. 389-403.

[32] P. Lange et al., “Virtual Reality for Simulating Autonomous Deep-Space Navigation and Mining”, in *24th Int. Conf. on Artificial Reality and Telexistence (ICAT-EGVE)*, Bremen, Germany, 2014.

[33] J. Balaram et al.: “DSENDS – A high-fidelity dynamics and spacecraft simulator for entry, descent and surface landing”, in *IEEE Aerospace Conf.*, Montana, USA, 2002, pp. 7-3342 – 7-3359.

[34] P. Lange et al., “A Framework for Wait-Free Data Exchange in Massively Threaded VR Systems”, in *Int. Conf. in Central Europe on Computer Graphics, Visualization and Computer Vision (WSCG)*, Plzen, Czech Republic, 2014.

[35] P. Lange et al., “Scalable Concurrency Control for Massively Collaborative Virtual Environments”, in *ACM Multimedia Systems: Massively Multiuser Virtual Environments (MMVE)*, Portland, USA, 2015.

[36] M. Brillouin, “Equations aux Dériveées partielles du 2e ordre. Domaines à connexion multiple. Fonctions sphériques non antipodes”, in *Annales De L ’Institut H. Poincaré*, Vol. 2, 1933, pp. 173–206.

[37] R. A. Werner, “Evaluating Descent and Ascent Trajectories Near Non-Spherical Bodies”, *Tech. Report*, Jet Propulsion Laboratory (JPL), 2010. Available: http://www.techbriefs.com/component/content/article/8726.

[38] Y. Takahashi et al., “Surface Gravity Fields for Asteroids and Comets”, in *Journal of Guidance, Control, and Dynamics*, Vol. 36 (2), 2013, pp. 362-374.

[39] R. Weller and G. Zachmann, “ProtoSphere: A GPU-Assisted Prototype-Guided Sphere Packing Algorithm for Arbitrary Objects”, in *ACM SIGGRAPH Asia 2010 Sketches*, New York, USA, 2011, pp. 8:1 – 8:2.

[40] R. Weller et al., “Massively Parallel Batch Neural Gas for Bounding Volume Hierarchy Construction”, in *Virtual Reality Interactions and Physical Simulations (VRIPhys)*, Bremen, Germany, 2014.