Computer Vision Based Behavior Analysis

A DISSERTATION SUBMITTED TO
THE DEPARTMENT OF ELECTRICAL AND ELECTRONICS
ENGINEERING
AND THE INSTITUTE OF ENGINEERING AND SCIENCE
OF BILKENT UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

By

Zeynep Yücel

December 2009

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a dissertation for the degree of doctor of philosophy.

Prof. Dr. A. Bülent Özgüler(Supervisor)

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a dissertation for the degree of doctor of philosophy.

Asst. Prof. Dr. Pınar Duygulu Şahin(Co-supervisor)

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a dissertation for the degree of doctor of philosophy.

Asst. Prof. Dr. Selim Aksoy

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a dissertation for the degree of doctor of philosophy.

Prof. Dr. Billur Barshan

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a dissertation for the degree of doctor of philosophy.

Assoc. Prof. Dr. Uğur Güdükbay

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a dissertation for the degree of doctor of philosophy.

Prof. Dr. Eric Pauwels

Approved for the Institute of Engineering and Science:

Prof. Dr. Mehmet Baray
Director of Institute of Engineering and Science

ABSTRACT

Computer Vision Based Behavior Analysis

Zeynep Yücel
Ph.D. in Electrical and Electronics Engineering
Supervisor: Prof. Dr. A. Bülent Özgüler
December 2009

In this thesis, recognition and understanding of behavior based on visual inputs and automated decision schemes are investigated. Behavior analysis is carried out on a wide scope ranging from animal behavior to human behavior. Due to this extensive coverage, we present our work in two main parts. Part I of the thesis investigates locomotor behavior of lab animals with particular focus on drug screening experiments, and Part II investigates analysis of behavior in humans, with specific focus on visual attention.

The animal behavior analysis method presented in Part I, is composed of motion tracking based on background subtraction, determination of discriminative behavioral characteristics from the extracted path and speed information, summarization of these characteristics in terms of feature vectors and classification of feature vectors. The experiments presented in Part I indicate that the proposed animal behavior analysis system proves very useful in behavioral and neuropharmacological studies as well as in drug screening and toxicology studies. This is due to the superior capability of the proposed method in detecting discriminative behavioral alterations in response to pharmacological manipulations.

The human behavior analysis scheme presented in Part II proposes an efficient method to resolve attention fixation points in unconstrained settings adopting a developmental psychology point of view. The head of the experimenter is modeled as an elliptic cylinder. The head model is tracked using Lucas-Kanade optical flow method and the pose values are estimated accordingly. The resolved poses are then transformed into the gaze direction and the depth of the attended object through two Gaussian regressors. The regression outputs are superposed to find the initial estimates for object center locations. These estimates are pooled to mimic human saccades realistically and saliency is computed in the prospective region to determine the final estimates for attention fixation points. Verifying the extensive generalization capabilities of the human behavior analysis method given in Part II, we propose that rapid gaze estimation can be achieved for establishing joint attention in interaction-driven robot communication as well.

ÖZET

BİLGİSAYARLA GÖRÜ TABANLI DAVRANIŞ ÇÖZÜMLEMESİ

Zeynep Yücel
Elektrik ve Elektronik Mühendisliği Bölümü, Doktora
Tez Yöneticisi: Prof. Dr. A. Bülent Özgüler
Aralık 2009

Bu tezde görsel girdi ve otomatik karar tabanlı davranış tanıma ve anlaması araştırılmıştır. Davranış çözümlemesi hayvan davranışlarından insan davranışlarına kadar geniş bir kapsamda yürütülmüştür. Bu geniş kapsam dolayısıyla incelemelerimizi ana iki bölüm içinde sunuyoruz. Bölüm I'de ilaç görüntüleme deneyleri bakımından laboratuvar hayvanlarının lokomotor hareketleri ve Bölüm II'de ise görsel ilgi bakımından insan davranışlarının çözümlemesi incelenmektedir.

Bölüm I'de sunulan hayvan davranış çözümleme yöntemi arka plan çıkarımına dayalı hareket izleme, elde edilen yol ve hız bilgisinden ayıredici davranışsal nitelikleri belirleme, bu nitelikleri öznitelikler yoluyla özetleme ve bu öznitelikleri sınıflandırma işlemlerinden oluşmaktadır. Bölüm I'de sunulan deneyler göstermektedir ki önerilen hayvan davranış çözümleme sistemi davranışsal ve nörofarmakolojik çalışmalarda olduğu kadar ilaç görüntüleme ve toksikoloji çalışmalarında da fayda sağlayacaktır. Bu durum, önerilen yöntemin ayırdedici davranışsal değişiklikleri tespit etmedeki üstün yeteneğine bağlıdır.

Bölüm II'de sunulan insan davranış çözümleme şeması, sınırlandırılmamış ortamlarda ilgi sabitleme noktalarının belirlenmesi için etkili bir yöntem önermektedir. Deneyi yapan kişinin kafası eliptik bir silindir olarak modellenmektedir. Bu kafa modeli Lucas-Kanade optik akış yöntemi ile izlenmekte ve buna göre duruş değerleri kestirilmektedir. Ardından çözümlenen duruşlar iki Gauss bağlanımı ile bakış doğrultusuna ve bakılan nesnenin derinliğine dönüştürülmektedir. Bağlanım çıktıları nesne merkezi konumlarının birincil kestirimlerini bulmak amacıyla çakıştırılmaktadır. Bu kestirimler insan seğirmelerini gerçekçi bir şekilde taklit etmek bakımından biriktirilmekte ve nihai kestirimleri elde etmek için muhtemel bölge üzerinde belirginlik hesaplanmaktadır. Bölüm II'de sunulan insan davranış çözümleme yönteminin kapsamlı genelleme kabiliyetini kanıtlayarak, hızlı bakış doğrultusu kestiriminin etkileşim güdümlü robot iletişiminde birleşik ilgi kurulmasını sağlayabileceğini öngörülmektedir.

ACKNOWLEDGMENTS

I owe my sincere gratitude to my supervisor Dr. Arif Bülent Özgüler for his supervision, guidance, suggestions and support throughout my studies leading this thesis. I would like to express my deepest thanks to my co-supervisor Dr. Pınar Duygulu Şahin for her positive attitude, help and guidance in this study.

I would like to thank Dr. Eric Pauwels, Dr. Billur Barshan, Dr. Uğur Güdükbay, and Dr. Selim Aksoy for their revisions and suggestions on my thesis. I would like to express my sincere gratitude to Dr. Yıldım Sara and Dr. Rüştü Onur for their cooperation. I am grateful to Dr. Emre Esen for his help with the drug screening experiments. I would like to thank Çetin and Tekin Meriçli for their providing me the dataset on attention modeling. This study would never be complete, if it were not for them.

I am thankful to Dr. Enis Çetin for his encouraging guidance, which took my research to the next level. I owe my most sincere gratitude to Dr. Nicu Sebe and Dr. Theo Gevers, who gave me the opportunity to work with them in The Intelligent Sensory Information Systems Group of the University of Amsterdam. I warmly thank Roberto Valenti for his friendly help and the extensive discussions around our work. It is a pleasure to express my special thanks to Dr. Albert Ali Salah, who introduced me to Dr. Eric Pauwels giving me an opportunity to join Centrum voor Wiskunde en Informatica. I am grateful to Ali for supporting my research with his vast innovative ideas during my studies in CWI as well as during our later cooperation.

I would like to thank to Mehmet Kök for being my friend and helping me with my every problem. I warmly thank Selçuk Sandıkcı for his friendship and discussions. I also would like to thank all my friends in Department of Electrical and Electronics Engineering, my colleagues at University of Amsterdam and CWI. Special thanks are to Çağdaş Öğüt for being my friend, my guide, and my mentor. I would like to thank Zoja, Taylan, Alessia and Hüseyin, who gave me a family away from home. My deepest thanks are to Sinan Ayyorgun, who put up with all my whims and gave me a smile even in my darkest moments.

My special gratitude is to my parents, Asuman and Yılmaz Yücel, for their understanding and loving support. To them I dedicate this thesis.

The financial support of the Scientific and Technical Research Council of Turkey (TÜBİTAK grants BTT-105E065, MILDAR-107A011 and Career-104E065) is gratefully acknowledged.

Contents

List of Figures

List of Tables

Dedicated to my parents Asuman & Yılmaz Yücel. . .

Chapter 1

Introduction

In this thesis, we handle recognition and understanding of behavior using automated decision schemes based on visual inputs. The proposed methods share the general characteristics of automated behavior analysis methods such as objectivity and precision among others. In addition to those, they offer advantages due to the particular capabilities of the visual capturing method and the data gathered visually.

Automated schemes prevail over the manual evaluation methods in several respects. The primary quality of automatic analysis tools lies in their objectivity through experimenter-independent analysis and decision opportunities. They can also withstand cumbersome experimental conditions, such as long durations. Moreover, by excluding manual labor, they eliminate human error providing more precise results in comparison to the conventional evaluation schemes. In addition to its accuracy, the data is also suitable to enable drawing inferences. From a behavioral analysis point of view, this enables definition and identification of various types of actions, which are otherwise very hard to describe in accurate terms with reliable integrity. This permits precise quantification of behavioral characteristics and definition of analytically detectable and tractable features.

Moreover, automatic analysis offers the possibility of employing various kinds of sensors and capturing devices, which in turn allows detection and tracking of different sorts of actions and behavioral attributes.

We use only visual inputs obtained from simple web cameras. The proposed vision-based behavior analysis methods offer additional opportunities due to the input media. One of additional favorable features is due to the low-cost and easy implementation of the equipment. Such apparatus become smaller in size in recent years and offer higher quality. Moreover, they are widely used and already established in various settings ranging from home security to public surveillance applications. Their application scopes can be readily extended by courtesy of easy incorporation with computer processing. Video capturing provides perception of the environment close to human perception. This fact proposes advantages in particular regarding artificial intelligence applications and enables design of reasonable decision mechanisms, which mimic human reasoning.

In this thesis, we make use of these favorable features of automated visual behavior analysis and propose several methods that operate on a wide scope ranging from animal behavior to human behavior. We present our work in two main parts, where Part I is dedicated to animal behavior analysis for drug screening purposes and Part II is dedicated to human behavior analysis for attention resolution.

Part I of this thesis investigates locomotor behavior of lab animals with particular focus on drug screening experiments. For this purpose, an effective behavior analysis tool, which discriminates locomotor activity changes with respect to administered psychotropic drugs, is described.

The proposed method has two main components, consisting of

- representation of locomotor activity in terms of feature vectors, and

- classification of features with respect to drug types.

For behavior representation, locomotor activity is expressed as spatial measurements of cumulative distance traveled and mean instantaneous speed. Feature vectors, which summarize the spatial distributions of these parameters, are employed in classification stage in order to match the test subjects and the drugs.

Our proposed vision-based behavior discrimination method presents numerous advantages over the previously described vision-based drug identification tools. Some contributions that come forth are

- elimination of manual labor, and mistakes due to subjectivity of the experimenter,
- formulation of locomotor activity changes in response to pharmacological manipulation,
- summarization of these behavior alterations in terms of feature vectors,
- significantly high detection rate.

These observations indicate that due to its superior capability of detecting discriminative behavioral alterations in response to pharmacological manipulations, the proposed system proves very useful in behavioral and neuropharmacological studies as well as in drug screening and toxicology studies.

Part II of the thesis investigates analysis of behavior in humans, with specific focus on visual attention. In investigation of human behavior, attributes relating interests, intentions, goals, and desires, have been an area of growing interest to many researchers. The significance of these factors lies particularly in the fact that they constitute some very important components of natural communication and continuous interaction. Thus, resolution of attention emerges as a fundamental paradigm in that respect.

Therefore, we propose a method to resolve attention fixation points visually. In order achieve robustness and resilience as well as continuous progress, we first identify relevant cognitive skills in humans and then mimic them on a digital platform.

Among all cognitive skills relevant for attention, resolution of gaze direction emerges as a prominent one. Moreover, in estimation of gaze direction, eye locations and head pose are recognized as two principal factors. Therefore, we design a method for gaze direction estimation inspired by the natural composition of these principal factors. A 3D model based head pose estimation method is described along with an isophote based eye localization scheme. Then we incorporate these methods in a unified framework, improving the accuracy and extending the operating range of both modules.

It has been shown in literature that neither head pose nor eye locations do not describe the gaze direction completely [85]. Therefore, a Gaussian process regression is applied to interpolate gaze direction from the derived parameters. In addition to this, a bottom-up feature-based saliency model is employed to improve the gaze direction estimates. Subsequently, the attention fixation points are assessed by analyzing the restricted visual field indicated by the gaze direction.

The described attention resolution mechanism proves very useful in understanding of human behavior from an attention point of view. Hence a human robot interaction scenario is considered to be a suitable practical implementation. Establishment of natural communication and maintenance of continuous interaction between a human caregiver and a robotic agent is shown to be facilitated immensely by the proposed attention resolution method. In our application framework, a human caregiver selects and attends to objects among several alternatives and the embodied agent resolves attention fixation points identifying the attended objects with the proposed method.

We demonstrate our system on a number of recordings and conduct interclass evaluations as well as intraclass evaluations verifying the extensive generalization capabilities of our method. Our results suggest that rapid gaze estimation can be achieved for establishing joint attention in interaction-driven robot communication as well.

The organization of the thesis is as follows. In Part I, discrimination locomotor activity changes based several psychotropic agents is studied. Chapter 2 discusses some previous work undertaken in this field. Chapter 3 gives details about the experimental setup and the experiment protocol, whereas Chapter 4 explains the derivation of behavioral characteristics and summarization of those using feature vectors. Chapter 5 discusses the performance of the proposed method by presenting the classification results. In Part II, human behavior is studied from an attention perspective. Chapter 6 mentions some previous work relevant for our problem. Chapter 7 presents the attention resolution experiment scenario and the experimental setup. The details of the head pose estimation and eye localization methods are given in Chapter 8, where Chapter 9 explains the incorporation of these to get the final object center location and attention fixation point estimates. Chapter 10 presents the experimental results and the discussion section.

Index Terms: Behavior analysis, motion tracking, head pose estimation, eye localization, attention resolution, joint attention modeling, classification.

Part I

Vision-Based Animal Behavior Analysis for Drug Screening

The term *drug screening* in general refers to a vast collection of iterative experiments on drug discovery and development. The pharmacological profiles of newly developed drugs are determined through these biological assays at several stages including molecular, cellular, organ system and the whole organism levels. Whole animal studies, which we undertake in this thesis with particular focus on visual behavior analysis, constitute a part of whole organism level tests together with human clinical trials.

Although the present essential objective of whole animal tests is derivation of intermediate inferences for the effect of the drugs on organisms and disease models, the statutory obligation is introduced mainly for protection of the people in clinical trials and later stages. Compulsory animal tests are enforced basically after the Diethylene glycol tragedy in the United States. In 1937, a chemical similar to antifreeze was added to a sulfa drug labeled “Elixir of Sulfanilamide” to make the medication more palatable to children. In the absence of any animal testing, terminal effects had gone unnoticed, resulting in more than 100 casualties. In response to this tragedy, the U.S. congress required safety testing of drugs on animals prior to general market release. The scope of drug screening on animals has later been significantly extended as a consequence of subsequent incidents similar to the Diethylene glycol tragedy.

In spite of the restricted scope of this thesis to the visual experiments, in pharmacokinetics, whole animal studies refer to a broad set of ADME-Tox tests, which is an acronym for absorption, distribution, metabolism, excretion and toxicity. All these tests are subject to extensive and strict regulations that may vary in different countries across the world. In this regard, most authorities aim to restrict the number of times individual animals may be used, the overall numbers used, and the degree of pain that may be inflicted without anesthetic. Experiments on vertebrate animals considered in this thesis are subject to the regulations of the local ethical committee of Faculty of Medicine of Hacettepe University [25].

Providing that these ethical reservations are fulfilled, we focus on experiments investigating behavioral alterations induced by psychotropic agents. Visual observation process is one of the primary analysis methods in this kind of experiments. As a matter of fact, until today such behavioral effects have been observed and discriminated in general by skillful authorities without any automatized auxiliary equipment. However, this process is quite troublesome due to several factors. To begin with, it may be very time consuming and laborious depending on duration of the experiment and the monitored agent. Moreover, errors arising from the human factor are inevitable in most cases. Finally, a precise quantification of the observed features is very hard to obtain relying only on bare visual perception.

We thus aim to design and implement an automatic tool that will help the medical authorities in investigation of behavioral alterations in such experiments. The proposed analysis method is suggested to facilitate the process immensely in several respects. First of all, human error is eliminated. The flexibility to design different sorts of experiments, which otherwise may cost extensive labor of human experts, is propounded. Furthermore, accurate evaluation of locomotor activity measures is offered.

In that respect, we first study certain well-surveyed psychotropic agents. Thereby, we aim to point out to the discriminative behavioral features, which may be associated with these agents. By re-deriving inferences on activity changes, which have already been designated by medical authorities, we intend to validate the effectiveness of the proposed tool. Therefore, the suggested tool can be affirmed as an auxiliary instrument in investigation of unknown substances or newly developed drugs.

In our experiments, we considered four systematically well-known psychotropic drugs, namely amphetamine, cocaine, morphine, and diazepam. Certain doses of these agents are known to induce particular effects on laboratory mice. These effects are formulated in quantitative terms, which are derived using the path covered by the mice, and further utilized in automatic classification of locomotor activity.

Our experiments are designed to grasp visual cues concerning locomotor activity changes. The mice are observed in an open-field arena and their activity is recorded for 100 minutes by a simple surveillance camera. For each animal the first 50 minutes of observation is carried out as the drug-free period. Each animal is then exposed to only one drug by intraperitoneal injection with either amphetamine or cocaine as the stimulant drugs or morphine or diazepam as the inhibitory agents. The arena is divided into virtual grids and the number of visits (sojourn counts) to the grids is calculated along with instantaneous speeds within these grids. The spatial distributions of sojourn counts and instantaneous speeds are utilized in construction of the feature vectors, which are fed to the classifier algorithms for the final step of matching the animals and the drugs.

We determine the animals that are drug-treated with a success rate of 96%. In sorting the data according to the increased or decreased activity, we achieve 92% accuracy. In the last stage, the method differentiates the type of psychostimulant or inhibitory drugs with a success rate of 70% and 80%, respectively.

The outline of the Part I of the thesis is as follows. Chapter 2 summarizes some relevant studies to our objectives. In Chapter 3, experimental setup and the formation of the database are explained together with the outline of the algorithm. The motion tracking algorithm, sample tracking results as well as some inferences about the drug effects and the steps of the proposed hierarchical classification algorithm are presented in Chapter 4. A discussion of the performance of the algorithm and some conclusions are presented in Chapter 5.

Chapter 2

Related Work

Behavioral studies in biological research are mostly based on the observation and evaluation of motor activity of animals in experimental models. In that respect, discrimination of variations in the locomotor activity is particularly important. A system, which is capable of detecting behavioral alterations in response to pharmacological manipulations, could prove very useful in behavioral and neuropharmacological studies, as well as in drug screening and toxicology applications. As yet, a wide variety of methods has been described for investigating motor activity.

One of the early works, which investigates rodent behavior focuses on mice, hamsters and rats, which suffer from chronic fatigue syndrome [15]. In order to monitor the activity of test subjects, force sensors located are at the bottom of the cage. In addition to these, an infrared photo beam, which detects and counts the turns of the running wheel, is employed.

A suitable arrangement of these infrared sensors is termed as *photo beam apparatus*. The photo beam apparatus can be a regarded as a conventional method in investigation of locomotor activity. It is based on the principle of generating a signal when an animal interrupts the infrared light. A coherent configuration of the infrared sensors can register movements in the desired directions, so that horizontal and vertical locomotor activity, area entries, and the occurrence of different activities, such as rearing, can be monitored. Due to its rich capabilities, the standard photo beam apparatus is used for recording motor activity for preclinical drug evaluation in numerous studies [14, 22, 74, 88]. The study of Drai et al. is worth mentioning, where they demonstrate the effects of amphetamine and phencyclidine in rats employing data measured by a standard photo beam tracking system [30].

Continuous-wave Doppler radar (CWDR) is used as an alternative to photo beam apparatus [71]. In [7] and [8], Austin et al. employ CWDR, to classify behavioral activation in rodents. Multilayer feed-forward neural networks, which are fed with the power spectrum estimation and root mean square values of these signals, classify them into exploring, grooming, and behavioral stillness classes.

Aside from force sensors, infrared photo beam and CWDR, video data has started to be used in tracking of rodent motion in recent years [69, 97]. The responses against therapeutic interventions and genetic mutations as well as behavioral responses to psychoactive drugs are observed on video data in visual terms. Computer systems utilizing suitable software are employed to analyze these video sequences in order to evaluate the animal behavior.

Automated visual observation and evaluation of locomotor activity presents significant advantages over the previous methods. Locomotor activity is incorporated within phenotype, which is very hard to quantify. However, the proposed method provides a reliable way of recording and a precise means of evaluation of locomotor activity. Besides, this evaluation is not prone to any operator bias. In contrast to bare visual observation, video tracking may also perform pattern analysis on a video sequence and derive quantitative measures for the behavior of interest [69].

In his work on motion tracking in medical imaging, Coatriex divides the vision based observation approaches into two classes, namely, boundary based and region based approaches [23]. Boundary based methods rely on active contour modeling and free-form curve-fitting. They give successful results in the absence of restrictions on shape and motion type, but fail in tracking the newly appearing objects and face instability problem. On the other hand, region based approaches make use of the information from an entire region, which makes them give robust results leading to a more stable system.

According to this classification of tracking approaches, the method described in this thesis can be characterized as region based. Another region based approach belongs to Zurn et al. [107]. The algorithms defined in [107] for light and dark cycle behavior analysis of rodents are simplified in [108] and a single algorithm is suggested for both cycles. Automated observation using video tracking is particularly suitable for recording locomotor activity. Activity is expressed as spatial measurements of distance traveled, speed, and acceleration [19, 29, 83, 84]. Andrews et al. investigate rodent behavior using path and speed data [4]. Smoothing operation on path data is followed by construction of a Gaussian Mixture model for the speed curve. Expectation maximization optimizes the parameters of a two-element mixture model, which is is sufficient for classifying motion into two categories of lingering and progression.

One of the late works belongs to Branson et al. where they handle a social scenario and describe a method for tracking the motion of three mice in a cage from side view video [17]. A simple foreground/background labeling is done, so as to overcome the occlusion problem. After thresholding the difference between the current frame and the background model, a two-element Gaussian Mixture model is applied onto the mouse images in order to detect any possible occlusions. The mice involved in occlusions are labeled to be foreground/background mice by using depth order heuristics. By optical flow estimation the location of the mouse in the background is estimated and therefore continuous tracking is achieved. In a recent study Shih and Young reported a combination of an accelerometer and video camera system to simultaneously measure vibration and locomotion activity and compare the effects of amphetamine and pentobarbital on mice [80].

This study aims developing an automated system for recording and analyzing the locomotor activity of mice in response to pharmacological manipulation. We present a video tracking method which utilizes an algorithm to detect and discriminate drug responses elicited by diverse pharmacological groups. In order to test the efficiency of the proposed method we employed typical pharmacological agents with well-described behavioral effects and carried out several experiments for discriminating the video recordings of the test animals. In the next chapter, we present details concerning the video dataset and give an overview of the proposed algorithm.

Chapter 3

Materials and Outline

This chapter elaborates on the details of the experimental procedure. In addition, formation of the dataset and the general characteristics of the psychotropic drugs are briefly discussed. We note that the methods and the procedures described in this chapter have been approved by the ethics committee of Hacettepe University with issue number 2008/71-4 [25]. The experiments are carried out together with Dr. Yıldırım Sara, Dr. Rüştü Onur, and Dr. Emre Esen of Department of Pharmacology of Hacettepe University and details of part of this study has been published in [103].

3.1 Experiment environment

3.1.1 Animals

The test subjects are chosen from among healthy adult male Swiss-albino mice weighing 30-35g . Mice are housed in groups of three per cage in a temperature-controlled room ($23 \pm 1^{\circ}$) with a relative humidity of $45-70\%$. They are kept in a 12h:12h light/dark cycle (illuminated between 18.00 and 06.00) with unrestricted access to food and water. Each test subject is exposed to a single drug. None of the mice has been to an arena or used in any kind of experiment before.

3.1.2 Drugs

The psychotropic drugs of amphetamine, cocaine, morphine, and diazepam are dissolved in saline and injected to the animals intraperitoneally. D-Amphetamine hydrochloride, and diazepam were obtained from Sigma Chemical Co. (USA), whilst cocaine hydrochloride and morphine hydrochloride were obtained from Etablissements Roques, France and Verenigde Pharmazeutische Fabriken, Holland, respectively.

All injections were given intraperitoneally in a volume of 10ml/kg dissolved in saline. The common effects of the drugs are listed below:

- **Amphetamine** is a prescription stimulant which is used in the treatment of attention-deficit hyperactivity disorder. The effects could include decreased appetite, increased stamina and physical energy, involuntary bodily movements, hyperhidrosis, hyperactivity, jitteriness, tachycardia, irregular heart rate, hypertension, and headaches.

- **Cocaine** is a potent central nervous system stimulant. The signs of stimulation are hyperactivity, restlessness, increased blood pressure, increased heart rate, and euphoria.

- **Morphine** acts directly on the central nervous system and relieves pain. In the management of severe pain, no other narcotic analgesic is more effective or superior to morphine.

- **Diazepam** is a benzodiazepine derivative drug. It is commonly used for treating anxiety, insomnia, alcohol withdrawal, and muscle spasms. It may

Table 3.1: Abbreviations, expansions and sample numbers.

Abbreviation	Expansion	# of Samples
N	Drug-Naive	24
T	Drug-Treated	24
I	Activity Increasing	12
R	Activity Reducing	12
A	Amphetamine	6
C	Cocaine	6
M	Morphine	6
D	Diazepam	6

also be used before certain medical procedures to reduce tension and anxiety, and in some surgical procedures to induce amnesia.

Henceforth, the drug sets are denoted with the initial letter of the drug name, i.e., with A, C, M, and D for amphetamine, cocaine, morphine, and diazepam, (see Table 3.1). For each drug, six test subjects are used. From the common drug effects, it is clear that certain doses of amphetamine and cocaine have stimulating effects, and thus the corresponding 12 samples are grouped into activity-increasing type of drugs, which is denoted by I in Table 3.1. Morphine and diazepam are in the set of activity-reducing type of drugs, designated with R. The combination of sets I and R constitute set of drug-treated samples, indicated by T, where the corresponding drug-naive videos of the same test subjects are contained in set N. The distribution of sample numbers and organization of the dataset are summarized in Table 3.1.

3.1.3 Open field experiments

The arena is an open field of $0.45m \times 0.45m$ with glass barriers (see Figure 3.1). The black sheet on the base helps to detect the location of the albino test subject by employing the color contrast. A CCD camera, which is an adjustable surveillance camera (Fly WC-OML300, China), is positioned at a height of $0.6m$ at the top of the cage and is connected to a personal computer. Illumination is obtained by means of an incandescent lamp of 40W, which is positioned next to the camera, and providing a homogeneous illumination on the arena.

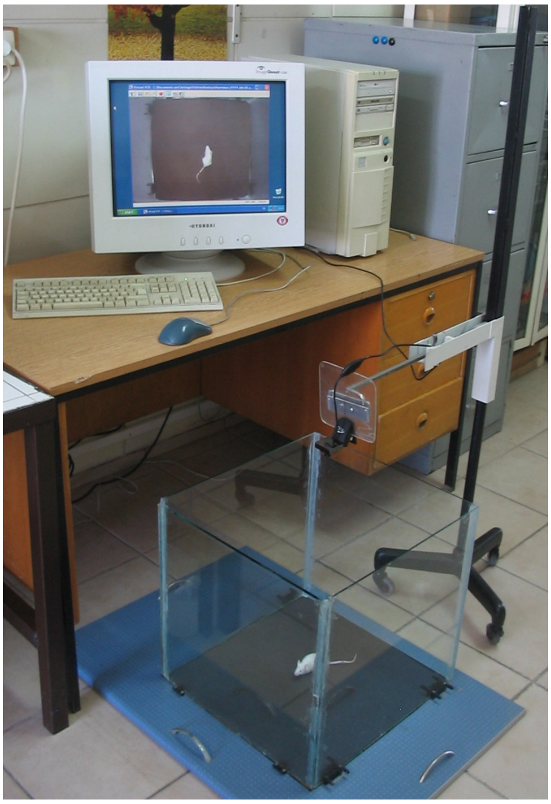


Figure 3.1: Experimental setup.

3.1.4 Experiment protocol

Experiments are performed according to a regular time schedule, namely between 09.00 and 15.00, in the weekdays except Monday, which is the weekly purification day of the vivarium.

A neat experiment environment is adapted throughout the series of experiments. The laboratory is purified from any odor or sound, which can lead to interfering effects. Moreover, the room is completely dark apart from the single illumination source positioned above the arena.

The test subjects are handled with great care. Before the experiment, mice are taken one at a time from their standard home cages, weighed and marked. Then animals are transferred to the open field apparatus and as they explored, video sequences are recorded at a frame rate of $10fps$.

Each animal is used only once and the concerning video sequences are recorded in two following sessions. In the first session, baseline activity of the mice is recorded for 50 minutes without drug administration. Immediately after this session, animals receive an intraperitoneal injection of amphetamine, cocaine, morphine, or diazepam and are placed back into the arena for another 50 minutes. Initial 10 minutes of each session were discarded. During this period animals resumed their baseline locomotor activity following manipulation.

3.2 Outline of the method

The proposed method is composed of two main building blocks, namely representation and classification of behavior (see Figure 3.2).

So as to achieve a comprehensive and yet simple representation, we need to draw only those features that indicate the distinctive behavior alterations induced by the psychotropic agents. For this purpose, we propose to use the traveled distance and instantaneous speeds of test animals. This requires tracking of motion and derivation of features from the tracked path. Therefore, motion tracking and feature extraction are recognized as two sub-blocks of behavior representation as indicated in Figure 3.2.

The inferences about the behavior of the subject are summarized in the feature vectors comprehending the behavioral distinctions due to the administered psychotropic agents. Subsequently the feature vectors are fed to the classifiers and an hierarchical scheme is applied resolving the drug properties in a gradual manner. In what follows, we elaborate on the details concerning these operations. The motion tracking algorithm is introduced in Chapter 4 together with sample tracking results and inferences about the drug effects. In addition to these, the two classification schemes are described in detail. The performance of the algorithm and some conclusions are presented in Chapter 5.

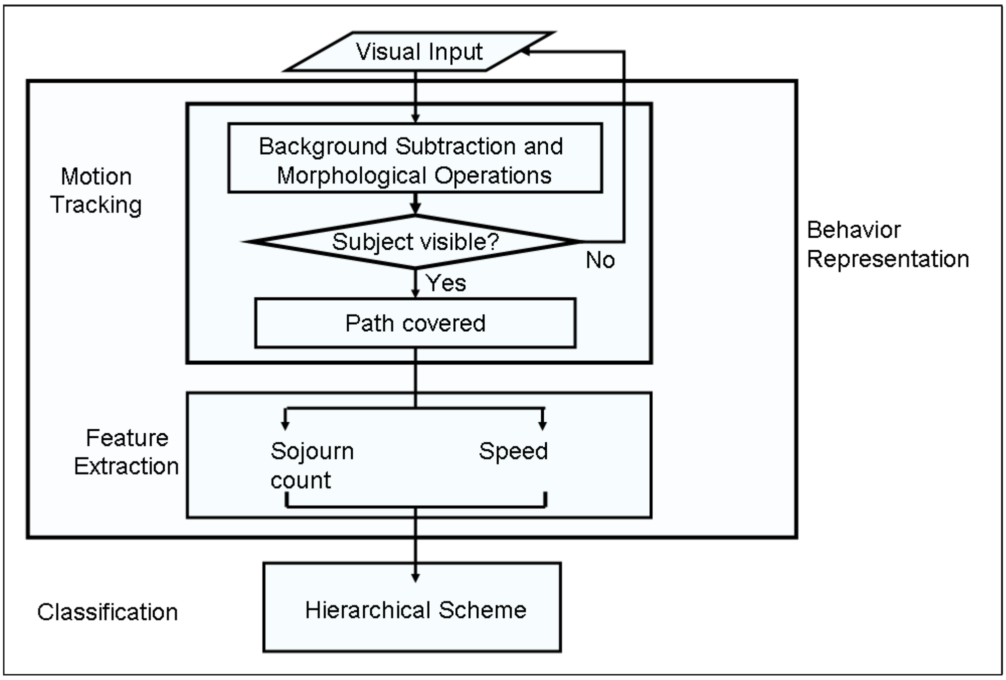


Figure 3.2: Outline of the method.

Chapter 4

Behavior Representation and Classification

The pertinent information reserved in locomotor behavior alterations is suggested to be summarized by a vector with certain number of quantified attributes. This requires, to begin with, interpretation of locomotor activity by medical authorities on bare visual observation. The benefit of this process is two-fold. First of all, the nature of discriminative changes in behavior are drawn forth. Secondly, we make sure that the well-surveyed drugs considered in the experiments induce effects that are in line with the expected results. Thereby, the possibility of overlooking any undetected medical disorders in mice is eliminated and it is ascertained that the videos qualify to be processed further.

Provided that this condition is satisfied, the embedded information is tractable to be formulated based on the visual feedback from the skilled observers. The first step in behavior representation is motion tracking. The derived path information is examined once more by the pharmacologists and it is affirmed that the initial feedback based on bare visual observation agrees with the outcome of the tracking phase. Subsequently, the feature vectors are derived with a special focus on locomotor activity changes indicated by the medical experts.

This chapter elaborates on the details of motion tracking and feature extraction processes. Details of background model and morphological operations are described in Section 4.1, whereas sample tracking results and related inferences are discussed in Section 4.2. The hierarchical classification scheme is explained in Section 4.3. The details concerning derivation of feature vectors are elaborated on in Section 4.4.

4.1 Motion tracking algorithm

The motion tracking algorithm relies on background subtraction and thresholding. The background image BG is recorded prior to the experiment just before the test subject is released into the arena. In order to get the region occupied by the test subject on the video image, a difference image is calculated for each frame of the video F_n, where $n = 1, \ldots, N$ and N is the total number of frames in that particular video [32]. In addition to this, a clipped difference image D_n is formed so as to avoid any reflections of the test subject on the glass barriers of the arena,

$$D_n(i - h_1 + 1, j - h_2 + 1) = F_n(i, j) - BG(i, j),$$
$$\forall i, j, n,\ h_1 \leq i \leq h_2,\ w_1 \leq j \leq w_2,\ 1 \leq n \leq N,$$

where h_1, h_2, w_1, and w_2, determine the borders of the arena. Since the camera and the arena are both stationary, h_1, h_2, w_1, and w_2 are constant for the whole video sequence.

To elicit the exact image of the test subject, the clipped difference image D_n is thresholded with a suitable bound. Subsequently, a series of morphological operations are carried out so as to eliminate any disturbances on the thresholded image due to possible bumps on the floor.

Let γ be the threshold and let the thresholded image be in the form a matrix X_n, each entry of which is formed according to,

$$X_n(i,j) = \begin{cases} 1 \text{ if } D_n(i,j) > \gamma \\ 0 \text{ if } D_n(i,j) < \gamma \end{cases}$$

$$\forall i, j, n,\ 1 \leq i < h_2 - h_1,\ 1 \leq j < w_2 - w_1,\ 1 \leq n \leq N.$$

The thresholded image X_n is enhanced through a series of morphological operations. First of all, the regions with an insignificant extent, which do not lie close to a large blob, are squeezed out. Then the unconnected blobs, which are in close proximity, are combined. Following these morphological operations, images such as the one presented in Figure 4.1 are obtained. Here the location of the test subject is designated as the center of mass of the white region. In order to grasp, the temporal evolution of behavioral changes in a piecewise fashion, a number of consecutive frames are grouped and processed in the described manner. By convention, for this kind of drug-screening experiments, the video is processed in non-overlapping bins of 40 seconds. For our case, this follows,

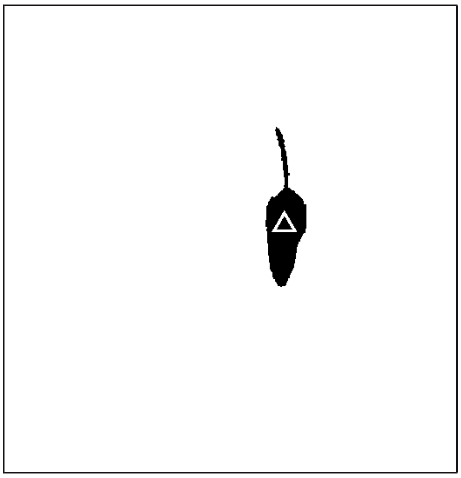


Figure 4.1: Center of mass for an example frame.

$$B_k = \{X_{(k-1)L+i};\ 1 \leq i \leq L\},\quad 1 \leq k \leq \left\lfloor \frac{N}{L} \right\rfloor,$$

where the bin B_k is a set of enhanced video frames, $L = 40 \times fps$ as fps stands for the frame rate and $\lfloor . \rfloor$ is the floor operator. Denoting the center of mass of the test subject in the i^{th} thresholded image of the k^{th} bin as p_k^i, a sequence of center of mass locations P_k concerning bin B_k is formed as,

$$P_k = \{p_k^i;\ 1 \leq i \leq L\}, \ , \ 1 \leq k \leq \left\lfloor \frac{N}{L} \right\rfloor ,$$

where $p_k^i = (x_k^i, y_k^i)$.

Connecting the points p_k^i to p_k^{i+1}, we obtain the path graphs concerning five different bins as in Figure 4.2, which illustrates sample paths for drug-naive and amphetamine-, cocaine-, morphine-, and diazepam-treated test subjects. The

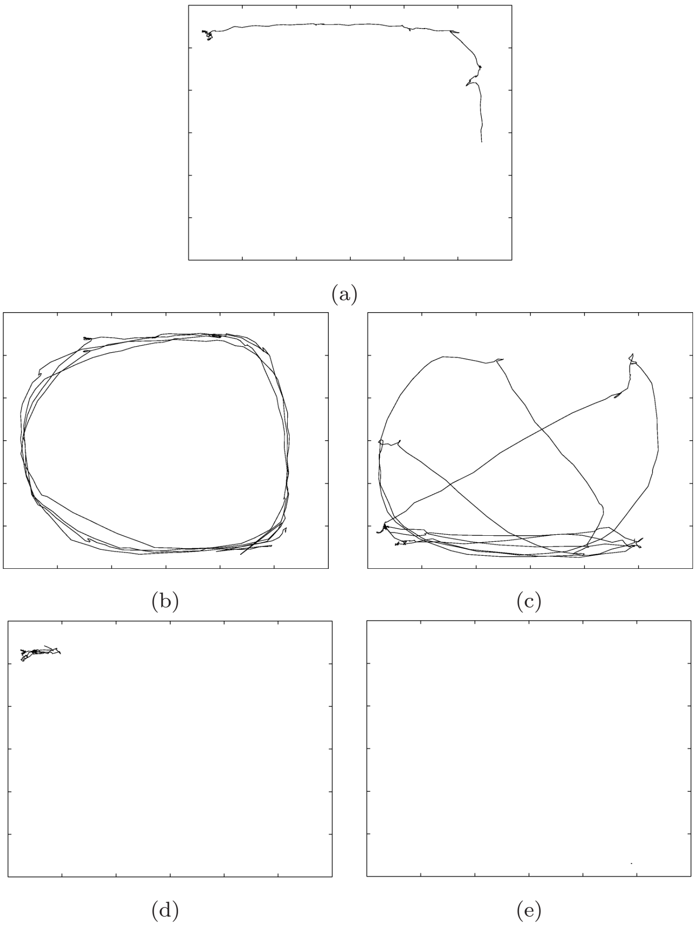


Figure 4.2: Example paths of (a) drug-naive, (b) amphetamine-, (c) cocaine-, (d) morphine-, (e) diazepam-treated test subjects along 40 seconds.

set of all path graphs in the dataset is examined by medical authorities and the validity of the dataset is confirmed once more.

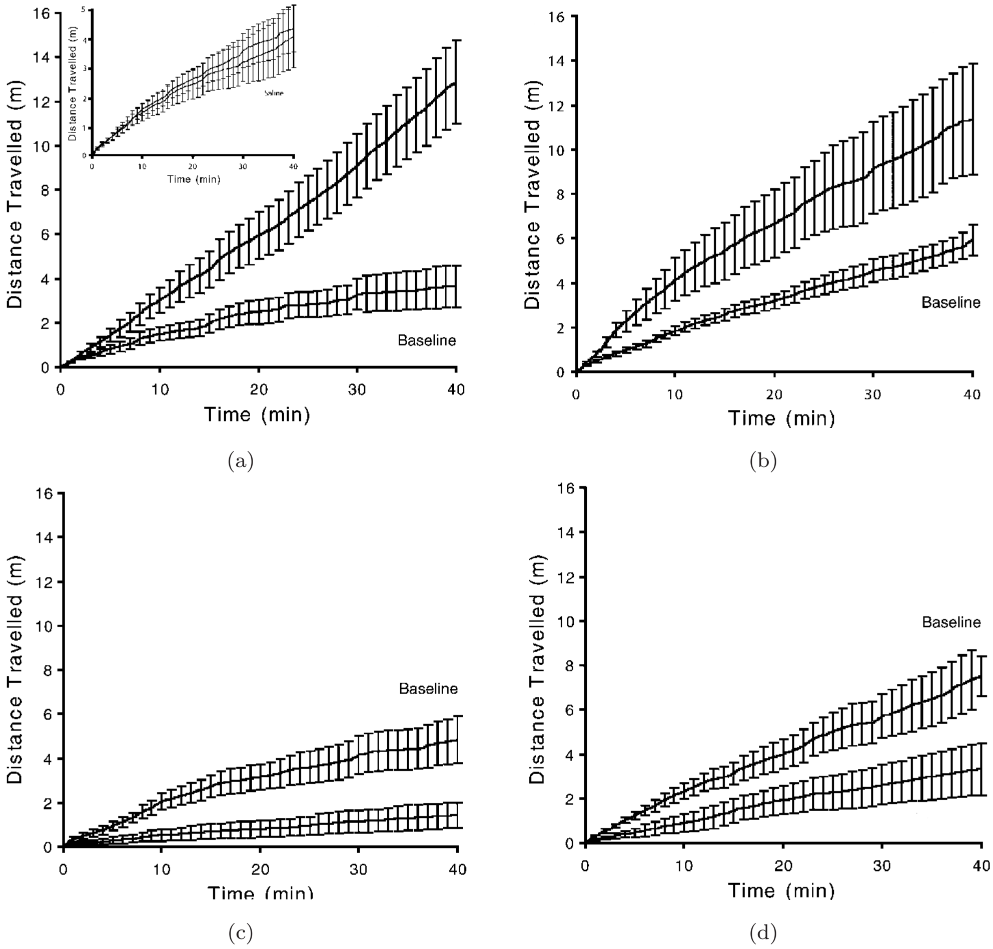
4.2 Inferences from the motion tracking results

Prior to the identification of the psychotropic drugs injected to the test subjects, we need to clarify several points. First of all, the drug-naive videos need to be shown to qualify to be compared to the drug-treated cases. Thereby, any bias due to stress and injection pain or the possibility of placebo effects are eliminated. The activity following saline injection is studied. Six mice are injected with saline and the covered path is calculated for that purpose. The cumulative traveled distance curves are observed to overlap before (i.e., baseline) and after the injections (see Figure 4.3-(a) inset). So discomfort of injection and placebo effect are inferred not to induce behavioral alterations in terms of the investigated attributes. Therefore, untreated baseline activity is shown to qualify to be compared to the stimulated locomotion.

Subsequently, we proceed by examining drug induced behavior. The cumulative traveled distances in Figure 4.3 reveal that amphetamine and cocaine increase locomotor activity compared to the pre-drug control period, while morphine and diazepam inhibit locomotion. However, further distinctive properties are inherent within both increased and decreased locomotor activities.

Stimulant drugs segregate mainly in terms of spatial distribution of activity on the arena. Amphetamine administered animals prefer to move along the edges of the arena, while cocaine-treated animals move throughout the arena including the central reagents, displaying a motion of more distributed nature (see Figures 4.6 and 4.7).

Morphine and diazepam inhibited locomotion display different characteristics as well (see Figures 4.3, 4.6 and 4.7). Under the influence of morphine, the animals remain sedated in one restricted area, generally located near the corners of the arena. Diazepam-treated animals remain sedated to a lesser extent, appearing slightly more active around the edges of the arena, with respect to the morphine group.



(a)

(b)

(c)

(d)

Figure 4.3: Cumulative traveled distances before and after psychotropic drug treatment for (a-inset) saline injection and, (a) amphetamine-, (b) cocaine-, (c) morphine- and (d) diazepam-treated test subjects. (Data is expressed as the mean cumulative distance traveled ± standard error of the mean (SEM) over the 40min of test period.)

In what follows, we give an account of summarization of the behavioral characteristics discussed in this section. Thereby, the amount of handled information is aimed to be reduced significantly.

4.3 Classification scheme

The observations listed in Section 4.2 indicate that several groups of drugs induce similar effects with inherent distinctions within the group. Based on this remark, a gradual inference method is suggested to attain a more efficient decision mechanism.

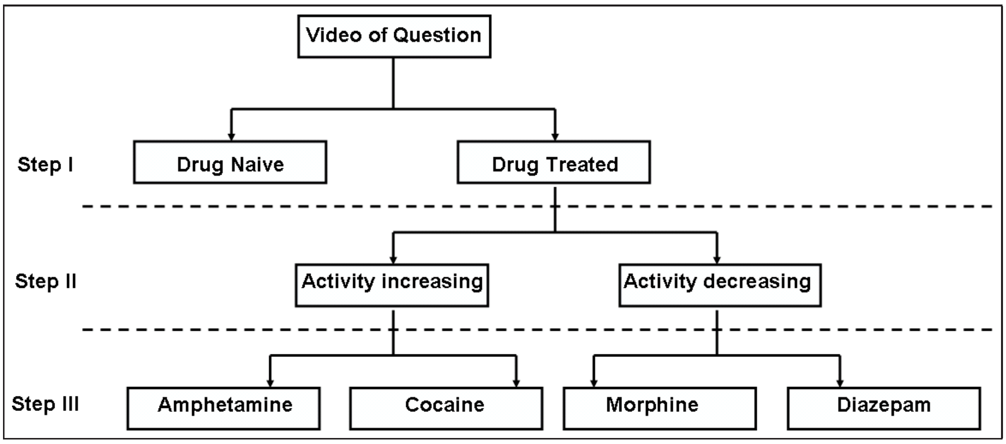


Figure 4.4: Hierarchical classification.

A hierarchical scheme is thus described to differentiate the administered drugs in a progressive manner (see Figure 4.4). In Step I, it is investigated whether the test subject is exposed to any kind of drug or is exhibiting a drug-naive behavior. If the video of question is detected to be drug-naive, no further investigation is performed. If it is detected to be drug-treated, the drug effect is ascertained as activity increasing or activity decreasing in Step II. Finally at the last step, taking the previously resolved drug characteristics into account the drug is detected.

Since these three steps utilize similar attributes from various view points, different feature vectors are defined for each step. The following section describes the formation of the feature vectors employed in each of the three steps of this hierarchical classification scheme.

4.4 Formation of feature vectors

Section 4.2 reveals that there are several variations in behavior on certain regions of the arena. For quantifying these distinctions, we accommodate the floor of the arena into a $\Lambda \times \Lambda$ grid-like structure. Let X be a particular video frame. A square grid, g_{uv}, covers the following region on X:

$$\begin{gathered} g_{uv}(i,j) = X((u-1)\lambda + i, (v-1)\lambda + j), \\ \forall i,j,u,\ 1 \leq i,j \leq \lambda,\ 1 \leq u,v \leq \Lambda, \end{gathered} \tag{4.1}$$

where λ is the number of pixels along one edge of g_{uv}. The grids g_{uv} are grouped based on the similarity of behavior on certain regions of the arena as corner, edge, and central regions. The boundaries are as indicated in Figure 4.5. Corners are denoted by $\mathbf{C_1}$, $\mathbf{C_2}$, $\mathbf{C_3}$, $\mathbf{C_4}$, and each of them covers $\varepsilon \times \varepsilon$ many grids. Edges are denoted by $\mathbf{E_1}$, $\mathbf{E_2}$, $\mathbf{E_3}$, $\mathbf{E_4}$ covering $\varepsilon \times (\Lambda - 2\varepsilon)$ many grids each. The central region with number of grids $(\Lambda - 2\varepsilon) \times (\Lambda - 2\varepsilon)$ is denoted by $\mathbf{M}$.

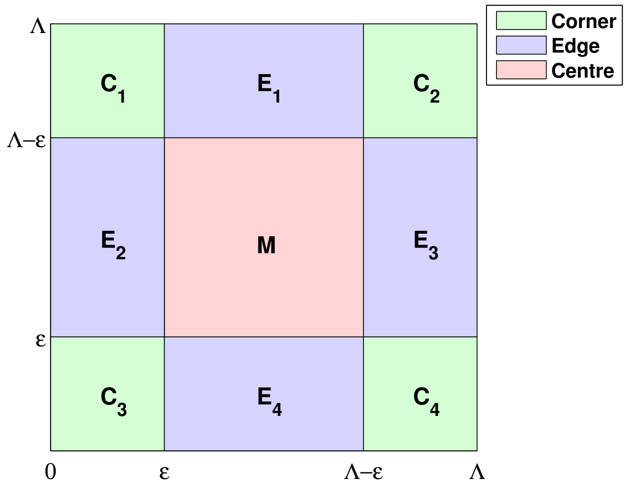


Figure 4.5: Locations and borders of three main regions.

The behavior of the test subject in each of these regions is investigated in terms of two measures, sojourn count and mean instantaneous speed. For a particular grid, sojourn count is defined as the number of visits to that grid, and the mean instantaneous speed is the mean value of the displacement vectors originating from that grid. The set of center of mass locations falling into grid g_{uv} is,

$$\pi_{uv} = \{p_k^i : \; p_k^i \in g_{uv}, \; \forall i, k, \; 1 \leq i \leq L, \; 1 \leq k \leq \lfloor \tfrac{N}{L} \rfloor\},$$
$$\forall u, v \; 1 \leq u, v \leq \Lambda.$$

It follows that the sojourn count ψ_{uv} of grid g_{uv} equals the number of elements of the set π_{uv}

$$\psi_{uv} = \mathcal{C}(\pi_{uv}), \forall u, v \; 1 \leq u, v \leq \Lambda.$$

where $\mathcal{C}(.)$ denotes the cardinality of a set. For evaluating mean instantaneous speed, we first need to obtain the displacement values. The displacement of the test subject from i^{th} frame to the $(i+1)^{st}$ frame of the k^{th} bin is denoted by ζ_k^i. A displacement vector ζ_k^i is derived from the coordinate vector P_k as follows

$$\zeta_k^i = \left| p_k^{i+1} - p_k^i \right|, \; \forall i, k, \; 1 \leq i \leq L-1, \; 1 \leq k \leq \left\lfloor \frac{N}{L} \right\rfloor.$$

This can also be regarded as a scaled version of the instantaneous speed of the test subject. Since the coordinates are determined at equal time intervals, $\zeta_k^i \times fps$ is the mean instantaneous speed. The set of displacements originating from the grid g_{uv} is denoted by ρ_{uv} and is given by

$$\rho_{uv} = \{\zeta_k^i : \; p_k^i \in g_{uv}, \; \forall i, k, \; 1 \leq i \leq L-1, \; 1 \leq k \leq \lfloor \tfrac{N}{L} \rfloor\},$$
$$\forall u, v \; 1 \leq u, v \leq \Lambda.$$

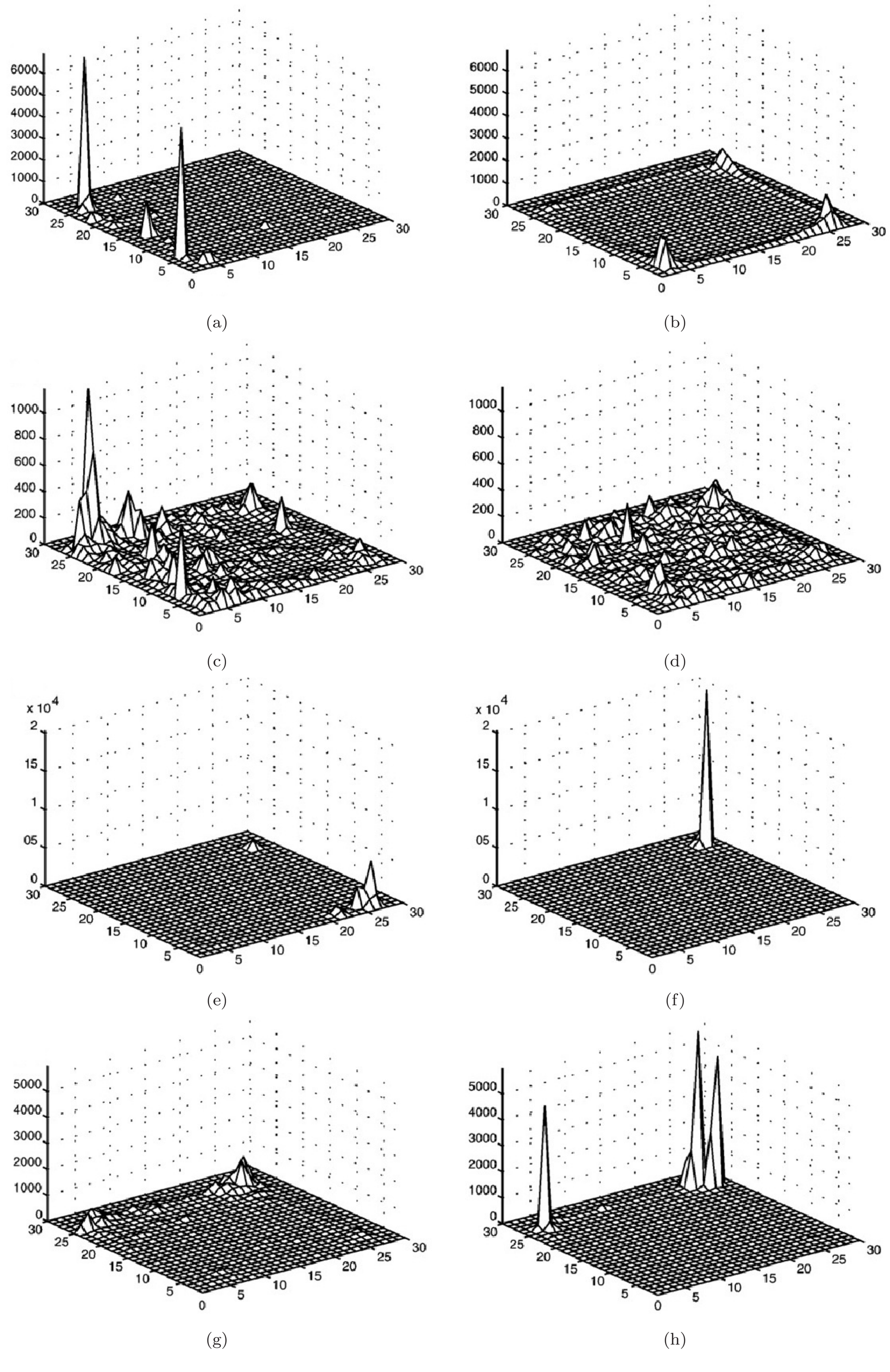
The mean instantaneous speed for grid g_{uv} is denoted by δ_{uv} and is obtained by calculating the mean value of the set ρ_{uv} and scaling by *fps* according to

$$\delta_{uv} = \mu(\rho_{uv}) fps, 1 \leq u, v \leq \Lambda,$$

where $\mu(.)$ denotes the mean value of a set.

The evaluation of sojourn count and mean instantaneous speed values are presented in Figures 4.6 and 4.7, respectively. In Figure 4.6, the value on the **z**-axis indicates the number of visits to the corresponding grids, i.e., sojourn count for (a) amphetamine-, (b) cocaine-, (c) morphine- and (d) diazepam-treated test subjects. In Figure 4.7, **z**-axis represents the average of instantaneous speeds that a particular test subject had when it left a grid.

One could infer similar results to those of Figure 4.2 by examining Figures 4.6 and 4.7. Both sojourn count and mean instantaneous speed of cocaine-treated test subjects present a more distributed pattern compared to those of amphetamine-treated test subjects. The figures indicate a clear difference in sojourn count and mean instantaneous speed for morphine- and diazepam-treated subjects as well. A morphine-treated test subject displays more activity in comparison to a diazepam-treated test subject.



(a)

(b)

(c)

(d)

(e)

(f)

(g)

(h)

Figure 4.6: Representative samples of the distribution of sojourn counts prior to (left column) and after (right column) administration of psychotropic drugs (a-b) amphetamine, (c-d) cocaine, (e-f) morphine and (g-h) diazepam.

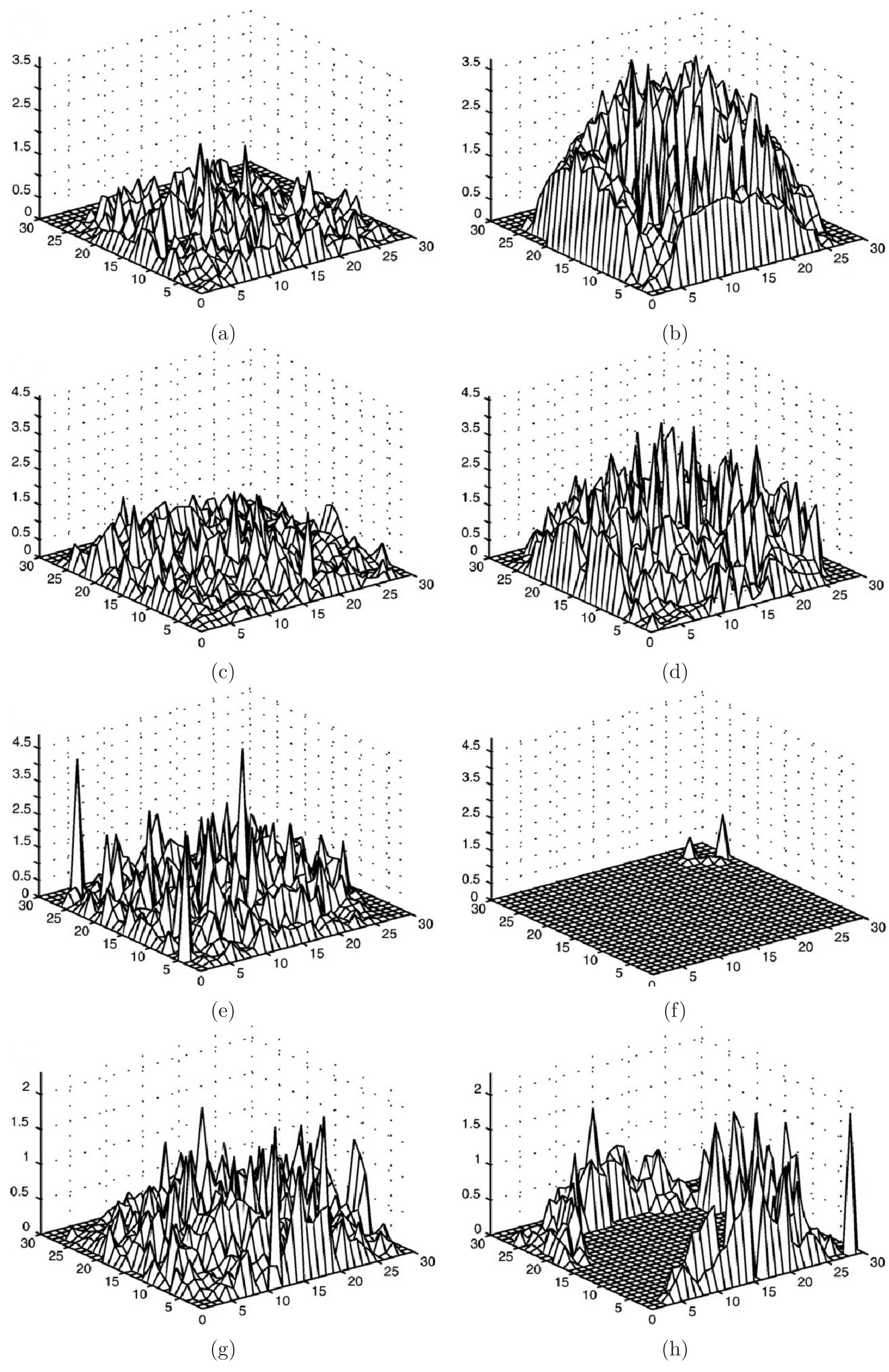


Figure 4.7: Mean instantaneous speeds of mice in each grid they visited prior to (left column) and after (right column) administration of psychotropic drugs (a-b) amphetamine, (c-d) cocaine, (e-f) morphine and (g-h) diazepam.

A feature vector contains information on sojourn count and mean instantaneous speed of a test subject on corner, center, and edge regions of the arena. The aggregate information for the corners and edges are obtained by aligning and adding the corresponding portion of the sojourn count and mean instantaneous speed matrices

$$\begin{aligned}
\psi_C &= \psi(\mathbf{C_1}) + \psi(\mathbf{C_2}) + \psi(\mathbf{C_3}) + \psi(\mathbf{C_4}),\\
\delta_C &= \delta(\mathbf{C_1}) + \delta(\mathbf{C_2}) + \delta(\mathbf{C_3}) + \delta(\mathbf{C_4}),\\
\psi_E &= \psi(\mathbf{E_1}) + \psi(\mathbf{E_2})^T + \psi(\mathbf{E_3})^T + \psi(\mathbf{E_4}),\\
\delta_E &= \delta(\mathbf{E_1}) + \delta(\mathbf{E_2})^T + \delta(\mathbf{E_3})^T + \delta(\mathbf{E_4}),\\
\psi_M &= \psi(\mathbf{M}),\\
\delta_M &= \delta(\mathbf{M}),
\end{aligned}$$

where for each $\mathbf{C}_l$ the matrices $\psi(\mathbf{C}_l)$ and $\delta(\mathbf{C}_l)$ are defined by

$$\psi(\mathbf{C}_l) = [\psi_{ij}] \text{ and } \delta(\mathbf{C}_l) = [\delta_{ij}],\ (i,j) \in \mathbf{C}_l.$$

$\psi(\mathbf{E}_l)$ and $\delta(\mathbf{E}_l)$ are defined similarly for $l = 1, \ldots, 4$.

4.4.1 Feature vectors for Step I of HC

The activity in regions **C**, **E**, and **M** is expressed in terms of the mean and standard deviation of the sojourn count and mean instantaneous speed. Thus the feature vectors for each region are

$$\begin{aligned}
\phi^I(C) &= [\mu(\delta_C)\ \sigma(\delta_C)\ \mu(\psi_C)\ \sigma(\psi_C)],\\
\phi^I(E) &= [\mu(\delta_E)\ \sigma(\delta_E)\ \mu(\psi_E)\ \sigma(\psi_E)],\\
\phi^I(M) &= [\mu(\delta_M)\ \sigma(\delta_M)\ \mu(\psi_M)\ \sigma(\psi_M)],
\end{aligned}$$

where functions $\mu(.)$ and $\sigma(.)$ give the mean and standard deviation of the indicated values. This follows that the feature vector Φ^I of a particular video for classification Step I is,

$$\Phi^I = [\phi^I(C)\ \phi^I(E)\ \phi^I(M)]^T,$$

which is the concatenation of the feature vectors for three regions.

4.4.2 Feature vectors for Step II of HC

In this step, the change induced by the drug in the sojourn count and mean instantaneous speed is investigated. Let the classifier at Step I label the input feature vectors Φ^I as Φ^I_N and Φ^I_T for the drug-naive and drug-treated recordings, respectively.

If the test subject is detected to be drug-treated at Step I, then the difference of the feature vectors of Step I is calculated and fed to the second step of HC as input

$$\Phi^{II} = \Phi^I_T - \Phi^I_N.$$

The classifier processes Φ^{II} and labels it either as Φ^{II}_i or as Φ^{II}_d, depending on whether the detection is "activity increasing" or "activity decreasing".

4.4.3 Feature vectors for Step III of HC

This step of the classification scheme aims at differentiating between amphetamine-, cocaine-, morphine-, and diazepam-treated test subjects.

Amphetamine-Cocaine classification

If Step II classifier labels Φ^{II} as Φ^{II}_i, then Step III classifier should decide whether the test subject is amphetamine- or cocaine-treated. This is done by training and testing the classifier with the same feature vectors of Step II but with different labels.

Morphine-Diazepam classification

If Step II classifier labels Φ^{II} as Φ^{II}_d, then Step III classifier should decide whether the test subject is morphine- or diazepam-treated.

Since morphine and diazepam both inhibit locomotor activity, unlike to the previous classification steps, only a small part of the arena provides behavioral information. We focus on the grids on which sojourn count of test subject are higher. If the maximum of the sojourn count appears at the grid $g_{u^*v^*}$, we focus on an $\eta \times \eta$ sub-matrix around grid $g_{u^*v^*}$. The sub-arena, denoted by r^*, is the set of the following grids

$$r^* = \{g_{uv} : \ u^* - \frac{\eta}{2} \leq u \leq u^* + \frac{\eta}{2} - 1, v^* - \frac{\eta}{2} \leq v \leq v^* + \frac{\eta}{2} - 1\}.$$

A procedure similar to the one illustrated in Figure 4.5 is applied to this sub-arena. The sub-arena r^* is divided into 9 sub-regions, r^*_{ij}, of equal size, similarly to Figure 4.5, where $1 \leq i, j \leq 3$. Thus, r^*_{ij} is given by

$$\begin{aligned} r^*_{ij} = \{g_{uv} : \ & u^* - \tfrac{\eta}{2} + (i-1)\tfrac{\eta}{3} \leq u \leq u^* - \tfrac{\eta}{2} + i\tfrac{\eta}{3} - 1 \\ & v^* - \tfrac{\eta}{2} + (j-1)\tfrac{\eta}{3} \leq v \leq v^* - \tfrac{\eta}{2} + j\tfrac{\eta}{3} - 1 \\ & 1 \leq i, j \leq 3, \ 1 \leq u, v \leq \Lambda\}. \end{aligned}$$

The sojourn count and mean instantaneous speed matrices for these sub-regions are formed according to,

$$\psi(r^*_{uv}) = [\psi_{ij}], \, \delta(r^*_{uv}) = [\delta_{ij}]. \ (i, j) \in r^*_{uv}, \, 1 \leq u, v \leq 3.$$

Similarly, corner, edge, and center regions, r^*_C, r^*_E, and r^*_M , are formed by grouping the sub-regions. The sojourn counts and mean instantaneous speeds are calculated by adding the corresponding portions of sojourn count and mean instantaneous speed matrices as,

$$\begin{aligned} \psi_{r^*_C} &= \psi(r^*_{11}) + \psi(r^*_{13}) + \psi(r^*_{31}) + \psi(r^*_{33}), \\ \delta_{r^*_C} &= \delta(r^*_{11}) + \delta(r^*_{13}) + \delta(r^*_{31}) + \delta(r^*_{33}), \\ \psi_{r^*_E} &= \psi(r^*_{12}) + \psi(r^*_{21}) + \psi(r^*_{23}) + \psi(r^*_{32}), \\ \delta_{r^*_E} &= \delta(r^*_{12}) + \delta(r^*_{21}) + \delta(r^*_{23}) + \delta(r^*_{32}), \\ \psi_{r^*_M} &= \psi(r^*_{22}), \\ \delta_{r^*_M} &= \delta(r^*_{22}). \end{aligned}$$

The formation of the feature vectors for corner, edge, and center parts are carried out in exactly the same manner as in Section 4.4. For instance, the feature vector for the corner part is given by,

$$\phi_{r_C^*} = [\mu(\psi_{r_C^*})\ \sigma(\delta_{r_C^*})\ \mu(\psi_{r_C^*})\ \sigma(\psi_{r_C^*})]^T.$$

Similarly for the edge and the middle parts. The feature vector for the third classification step Φ^{III} is the concatenation of the feature vectors for all parts, i.e., r_C^*, r_E^*, and r_M^*,

$$\Phi^{III} = [\phi_{r_C^*}\ \phi_{r_E^*}\ \phi_{r_M^*}]^T.$$

Finally the classifier processes Φ^{III} to label it as morphine- or diazepam-treated.

4.5 Classifiers

In classification step, Linear Discriminant Classifier (LDC) and Support Vector Classifier (SVC) are used [31]. LDC employs linear discriminant functions in classification, where SVC is based on support vector machines [2]. The details about these classification algorithms are given in Sections 4.5.1 and 4.5.2.

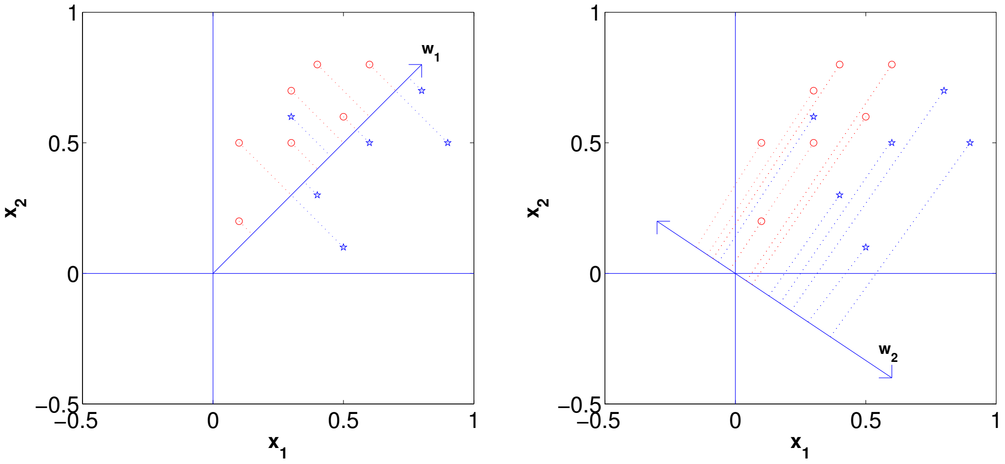
4.5.1 Linear discriminant classification

Linear discriminant analysis looks for a function that gives the most efficient direction for discrimination, namely linear discriminant function [12]. Let $\mathbf{x}$ be a feature vector. A linear discriminant function is a linear combination of components of a vector $\mathbf{x}$ and can be written in the form

$$g(x) = \mathbf{w}^T\mathbf{x} + b,$$

where $\mathbf{w}$ is the weight vector and b is threshold weight.

Let c_i be a class, where $i = 1, \ldots, C$. Denoting the linear discriminant function associated with class c_i as g_i, a feature vector $\mathbf{x}$ is assigned to c_i if the



(a)

(b)

Figure 4.8: Two examples of linear discriminant function for the same binary classification problem.

following condition is satisfied

$$g_i(\mathbf{x}) > g_j(\mathbf{x}) \ \forall j \neq i$$

In a binary class classification case, if the decision surface is a hyperplane, $g(\mathbf{x})$ is linear. The orientation of the decision surface is determined by the vector $\mathbf{w}$ and location is determined by b. Figure 4.8 illustrates a two class classification problem. Two possible linear discriminant functions are shown. The first one is not able to provide an efficient discrimination while the second achieves a more satisfactory separation.

4.5.2 Support vector classification

Among all hyperplanes that separate the given classes, there exist a unique hyperplane, which gives the maximum margin of separation. The highest margin of separation implies that the distances from the hyperplane to the nearest data points in the separated classes are maximized [79]. The margin, measured perpendicularly to the hyperplane, equals $\frac{2}{\|\mathbf{w}\|}$. Support vectors are employed in finding this particular hyperplane, making margin of separation maximum [37].

Given a set of training examples and class labels, $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n) \in \mathbb{R}^d \times \{\pm 1\}$, the goal is to find a classifier function $f : \mathbb{R} \to \{\pm 1\}$ such that $f(\mathbf{x}) = y$ will correctly classify new patterns. There exist a class of hyperplanes that separate the two classes [79]

$$(\mathbf{w} \cdot \mathbf{x}) + b = 0, \ w \in \mathbb{R}^d \ , b \in \mathbb{R},$$

corresponding to decision functions

$$f(x) = \text{sign}((\mathbf{w} \cdot \mathbf{x}) + \mathbf{b}).$$

To construct the optimal hyperplane, the following optimization problem is defined:

$$\begin{aligned} &\text{minimize } \tfrac{1}{2}\|\mathbf{w}\|^2 \\ &\text{subject to } y_i((\mathbf{w} \cdot \mathbf{x}_i) + b) \geq 1, \ i = 1, \ldots, n. \end{aligned}$$

The solution to this minimization problem can be obtained using quadratic programming techniques. The solution vector, then, becomes,

$$\mathbf{w} = \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i.$$

for some α_i $i = 1, \ldots, n$. In this case, it is obvious that the solution is the summation of a subset of training examples. The training patterns, which have at least one nonzero α_i, are called the support vectors. Support vectors lie on the margin and carry all the relevant information. The value of b is calculated using

$$\alpha_i(y_i((\mathbf{w} \cdot \mathbf{x}_i) + b) - 1) = 0,$$

Any one of the support vectors satisfy the equation above. Replacing the previously found $\mathbf{w}$ and one of the support vectors, one can solve for b.

As an example, consider the binary classification problem depicted in Figure 4.9. In this case, the stars are separated from the balls using a hyperplane. The optimal separating hyperplane is orthogonal to the shortest line connecting the convex hulls of the two classes and it intersects it half way between the classes.

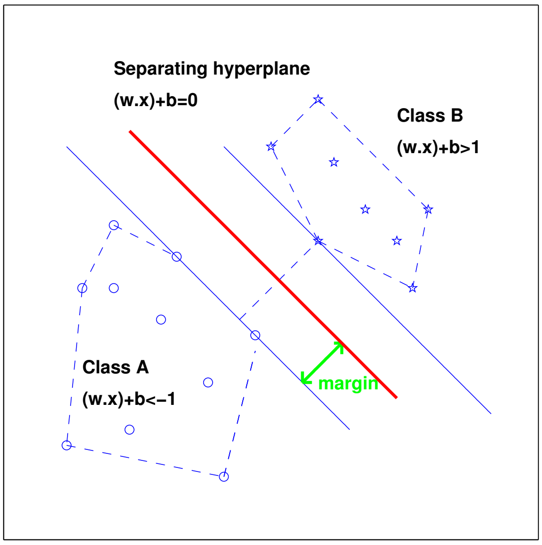


Figure 4.9: Implementation of SVM for a binary classification problem.

Chapter 5

Performance

This chapter discusses the performance of the proposed method at each stage of the hierarchical classification scheme described in Chapter 4. A series of experiments are carried to investigate the performance in training and test stages separately.

Training performance is described by how well the classifier learns the characteristics of the classes. While exploring training performance, the classifier is trained with a number of training examples and then tested with exactly the same set of patterns. The evolution of classification performance against training set size is observed by increasing the size of the training set gradually. It is then investigated whether a classifier is able to apprehend the class properties or not. If the training performance turns out to be satisfactory, then the classifier is said to be convenient for solving the problem of interest.

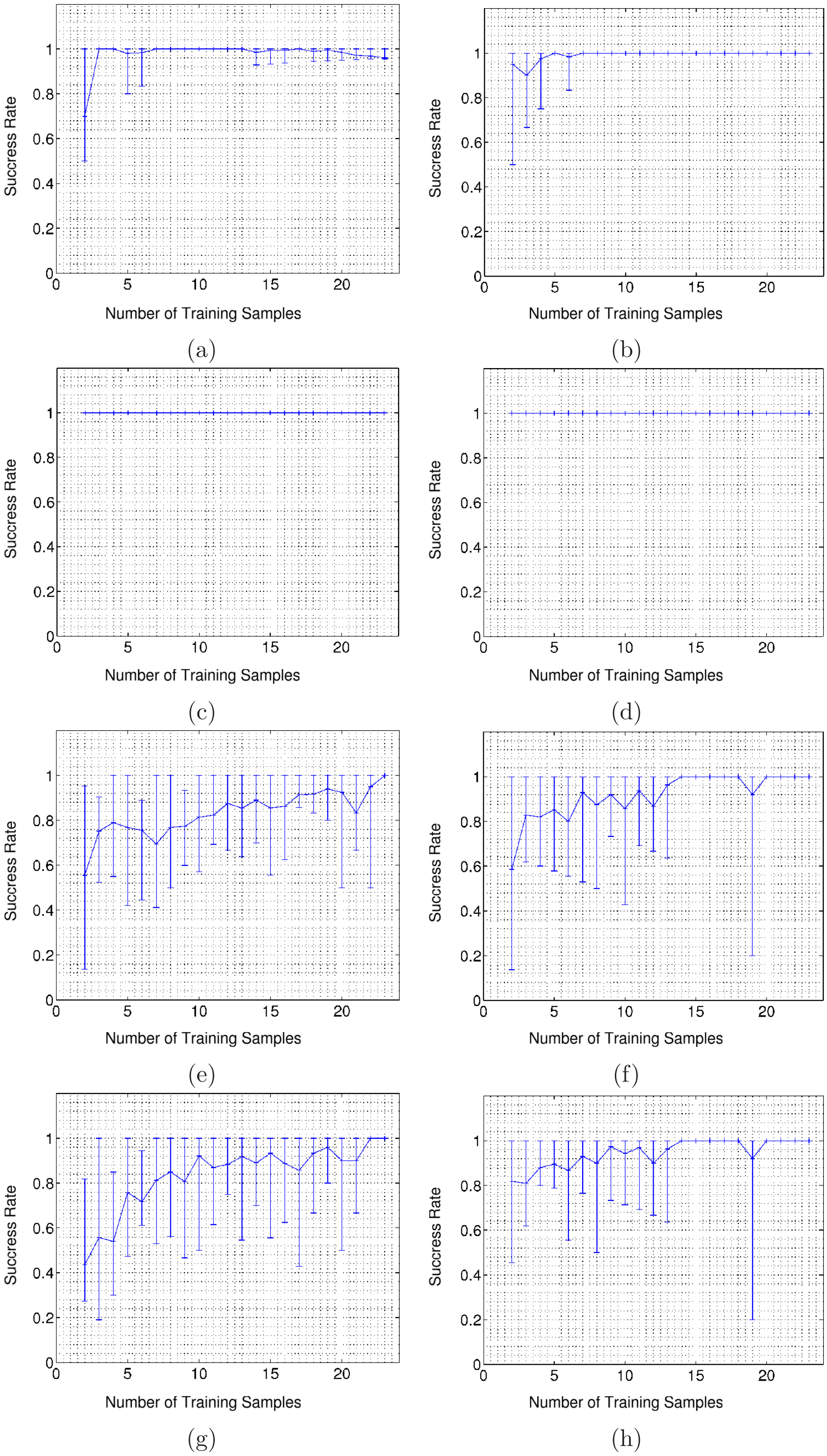
Test performance indicates how well the classifier performs when new patterns are investigated for class membership. While measuring test performance, the classifier is trained with a number of training patterns and then tested by new patterns. The number of training patterns is increased step by step and the classifier is tested by the rest of the dataset at each step. As we increase the number of training examples, the classification performance is expected to increase provided that there exists a certain pattern in the data and settle down around a steady state value. In this manner, we see how large a data set suffices to describe the classes thoroughly. We can check whether our dataset is large enough to comprehend the properties of all different kinds of class members.

In addition to this gradual scheme, a leave-one-out (LOO) cross-validation method is applied as well. As a matter of fact LOO is a special case of the gradual exploration pattern, where the test set size is one, and the rest of the dataset is used as training patterns. That is to say, LOO uses a single observation as the validation data and the remaining observations as the training data. This procedure is repeated such that each observation in the dataset is used once as the validation data. For LOO classification, we check whether the classifiers mislabel the inputted feature vectors (items) consistently. In some cases, not all of the items in a class present alike characteristics. This causes mislabeling in a consistent manner among the classifiers. Such misclassification is considered not to be attributed to the classifiers. In the particular experiments carried out in this work, the reason for such kind of errors may be originating from the age, weight, or metabolism of the mice.

We present in Sections 5.1, 5.2, and 5.3, the evolution of training and test performance with respect to varying number of training samples. The graphs presented in these sections indicate the average success rate calculated by taking the mean of all possible combination of training and test set pairs. In addition to this, the minimum and maximum success rates are indicated by the vertical lines in Figures 5.1, 5.2, 5.3, and 5.4.

5.1 Evolution of training and test performance in Step I of HC

Figures 5.1-(a) and 5.1-(b) show training performance for classes N and E with LDC classification scheme, while Figures 5.1-(c) and 5.1-(d) present training performance for the same classes with SVC classification. These figures indicate that both classifiers are able to grasp the class characteristics with notable success rates. Thus, the classifiers do qualify to be used in the test phase. Figures 5.1 (e-h) illustrate evolution of test performance with respect to increasing number of training samples with a similar organization to Figures 5.1 (a-d). As expected, test performance increases as the number of training samples increase. Moreover, the figures indicate that it settles down in general to values around 80%-90% for number of taining samples more than 10.



(a)

(b)

(c)

(d)

(e)

(f)

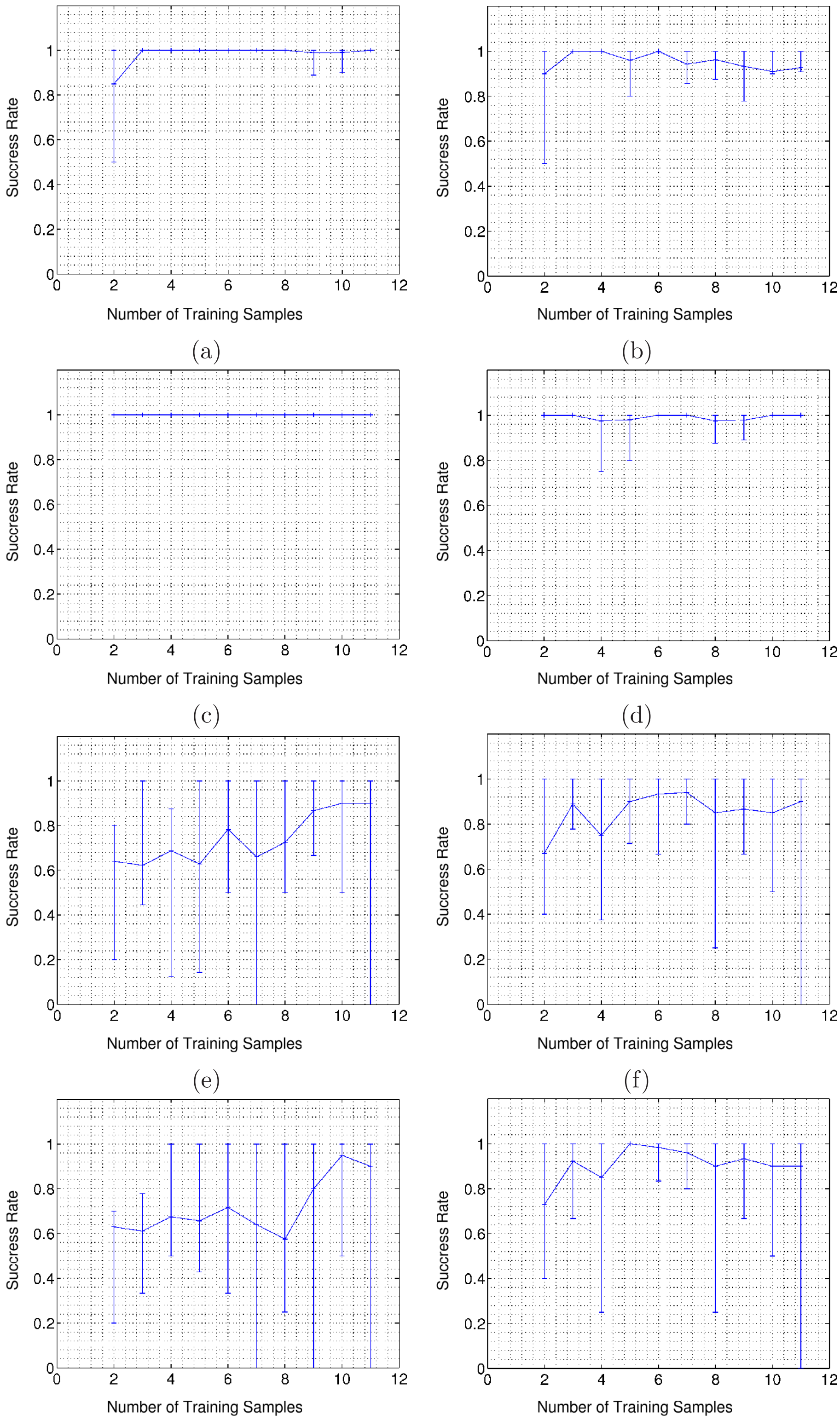
(g)

(h)

Figure 5.1: Evolution of (a-d) training and (e-h) test performance for Step I of HC. Figures (a) and (b) indicate the detection rates in training phase with LDC for drug-naive and drug-treated test subjects, where Figures (c) and (d) are organized similarly for SVC. Figures (e-h) are have a similar arrangement for test performance.

5.2 Evolution of training and test performance in Step II of HC

The evolution of performance is investigated in a similar manner to Section 5.1 and the results are presented with the same organization as in Figure 5.1. The training performance results presented in Figures 5.2 (a-d) prove that the LDC and SVC classifiers are able to apprehend the class properties with a success rate of more than 90%. Due to the specific focus of Step II of HC, the sizes of the training and test sets reach half as much of Step I. Nonetheless, for increasing number of trainig samples, the success rate in general reaches over 80%.



(a)

(b)

(c)

(d)

(e)

(f)

(g)

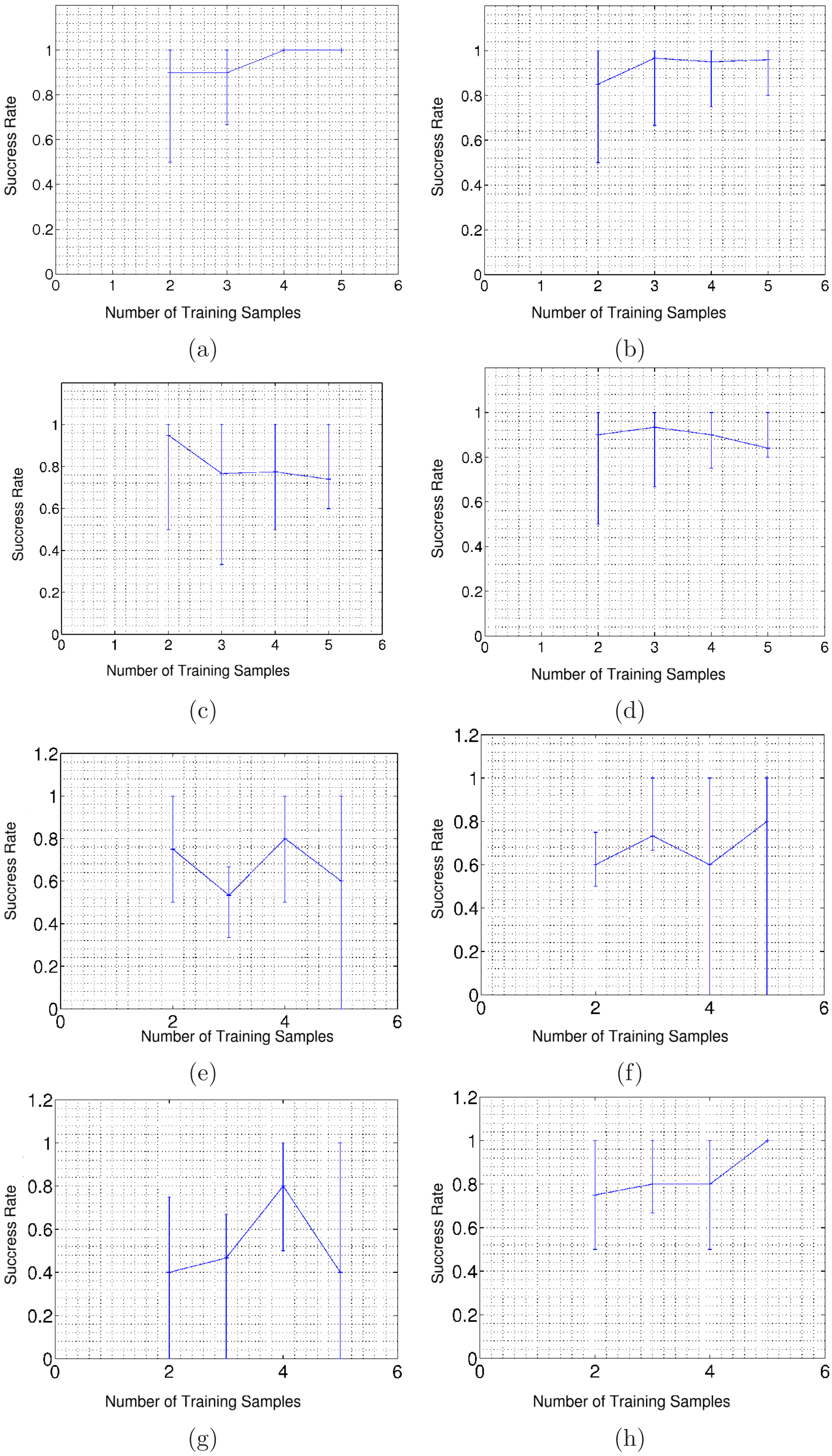
(h)

Figure 5.2: Evolution of (a-d) training and (e-h) test performance for Step II of HC. Figures (a) and (b) indicate the detection rates in training phase with LDC for activity increasing and activity decreasing type of drugs, where Figures (c) and (d) are organized similarly for SVC. Figures (e-h) are have a similar arrangement for test performance.

5.3 Evolution of training and test performance in Step III of HC

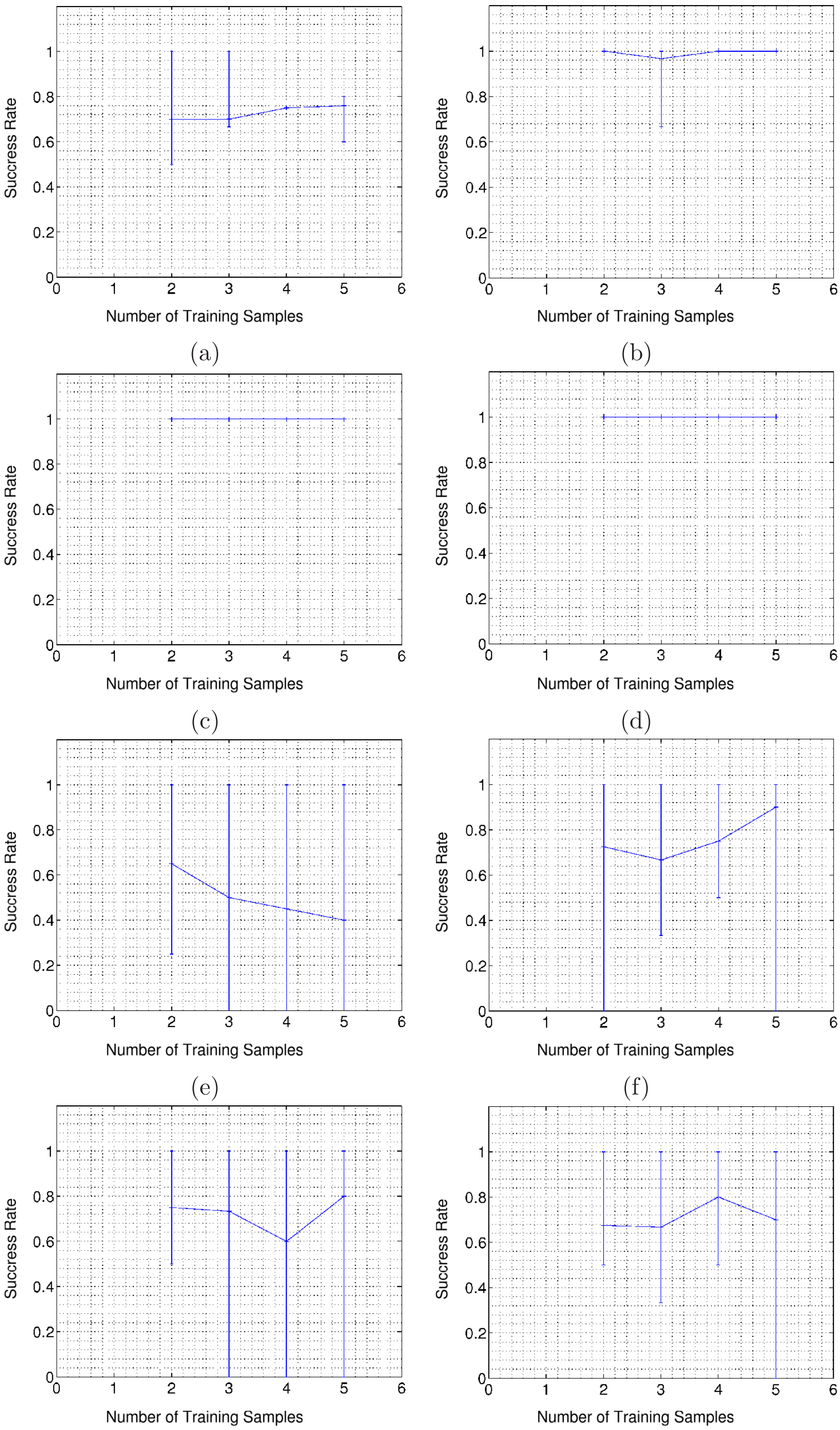
This section presents training and test performance for Amphetamine-Cocaine and morphine-diazepam classification with similar organization to Sections 5.1 and 5.2.

The evolution of training performance in Figures 5.3 (a-d) and 5.4 (a-d) indicates the limited size of the number of training patterns is not enough to comprehend the class properties extensively. Therefore, we expect that increasing number of training samples will not have an obvious effect in the improving test performance. Nevertheless, we investigate test performance and find results that are in line with the inferences from an examination of training performances. Figures 5.3 (e-h) and 5.4 (e-h) indicate deficiencies due to the limited number of feature vectors.



(a) (b) (c) (d) (e) (f) (g) (h)

Figure 5.3: Evolution of (a-d) training and (e-h) test performance for amphetamine-cocaine classification. Figures (a) and (b) indicate the detection rates in training phase with LDC for amphetamine- and cocaine-administered test subjects, where Figures (c) and (d) are organized similarly for SVC. Figures (e-h) are have a similar arrangement for test performance.



(a)

(b)

(c)

(d)

(e)

(f)

(g)

(h)

Figure 5.4: Evolution of (a-d) training and (e-h) test performance for morphine-diazepam classification. Figures (a) and (b) indicate the detection rates in training phase with LDC for morphine- and diazepam-administered test subjects, where Figures (c) and (d) are organized similarly for SVC. Figures (e-h) are have a similar arrangement for test performance.

5.4 LOO classification performance

The outcomes of LOO classification are presented in a confusion table, which indicates success and failure rates in terms of the predicted (assigned) and actual (true) classes. Each row of the table represents the instances in a predicted class, while each column represents the instances in an actual class. Table 5.4 gives a visualization of the confusion table with the designation of each cell. From Tables 5.2-(a) and (b), taking into account the number of test patterns, we conclude that there are 2 false-positives for LDC and 1 false-positives for SVC. Going back to the videos, it turns out that one video appeared as false-positive for both of the classifiers. From Tables 5.3-(a) and (b), number of false-positives for LDC and SVC are 2 and 1, respectively. There are no common mistakes

Table 5.1: Designations of cells of confusion table.

		Predicted Class	
		Positive	Negative
Actual Class	Positive	True Positive	False Negative
	Negative	False Positive	True Negative

Table 5.2: LOO with (a) SVC and (b) LDC classifiers for Step I of HC.

		Assigned Class	
		N	T
True Class	N	0.96	0.04
	T	0	1

(a)

		Assigned Class	
		N	T
True Class	N	0.92	0.08
	T	0	1

(b)

Table 5.3: LOO with (a) SVC and (b) LDC classifiers for Step II of HC.

		Assigned Class	
		I	R
True Class	I	0.92	0.08
	R	0.08	0.92

(a)

		Assigned Class	
		I	R
True Class	I	0.83	0.17
	R	0.08	0.92

(b)

among these false-positives. The number of false-negatives for both LDC and SVC is 1 and the misclassified video sequence turns our to be the same for both classification methods. As seen in Tables 5.4-(a) and (b), the number of false-negatives is 3 and 2 for LDC and SVC, respectively. Both false-negatives of SVC are among the false-positives of LDC. There is 1 false-positive for both of the classifiers and it is common. There are 2 false-negatives for LDC and 1 for SVC from Tables 5.5-(a) and (b). The one false-negative of SVC appeared also among the ones of LDC. The number of false-positives are 2 and 1 for LDC and SVC, respectively. The video that turned out to be false-positive in SVC classification is assigned to the wrong class in LDC classification as well.

Table 5.4: LOO with (a) SVC and (b) LDC classifiers for Amphetamine-Cocaine classification in Step III of HC.

		Assigned Class	
		A	C
True Class	A	0.67	0.33
	C	0.17	0.83

(a)

		Assigned Class	
		A	C
True Class	A	0.5	0.5
	C	0.17	0.83

(b)

Table 5.5: LOO with (a) SVC and (b) LDC classifiers for Morphine-Diazepam classification in Step III of HC.

		Assigned Class	
		M	D
True Class	M	0.83	0.17
	D	0.17	0.83

(a)

		Assigned Class	
		M	D
True Class	M	0.67	0.33
	D	0.33	0.67

(b)

5.5 Discussion

This section provides a detailed interpretation of the experimental results presented in Sections 5.1- 5.4. We first walk through the performance rates at all stages of the HC and comment on the results for each experiment set. Based on these, we discuss the competence of the dataset and on probable causes of misclassification. Finally, we provide a comparative evaluation of performances of SVC and LDC classifiers.

The experimental results presented in Sections 5.1- 5.4 provide a comprehensive outlook for the efficiency of the proposed method. The discrimination of drug-naive and drug-treated animals is obtained with considerably high success rates in Step I of HC. Although the baseline activity of mice displayed variations among the groups, it is observed that over 96% of drug-naive animals are labeled correctly. The proposed method also displayed 92% accuracy in Step II of HC, where the drug-treated animals were sorted according to their increased or decreased locomotor activities. The proposed scheme matches the animals and the drugs correctly with reasonably high success rates, in cases, where both of the drugs yielded quite similar cumulative distance and mean instantaneous speed curves. As an example, for Amphetamine-Cocaine classification in Step III of HC, the cumulative distances traveled and mean instantaneous speed values are influenced in the same direction by the administered drugs. Nonetheless the algorithm achieved 70% success rate in drug-animal matching. Similarly, the correct classification rate of Morphine-Diazepam classification is still around 80% despite the comparable effects of the drugs. Finally, the conclusion of the hierarchical classification scheme has an accuracy of $70 - 80\%$. The efficacy of the proposed feature vectors and the proposed classification scheme is thus ascertained with the findings from these performance experiments.

The competence of the dataset for such kind of drug identification experiments is another question that we need to answer. In the training phase of Steps I and II of HC, the classifiers are observed to learn the characteristics of all classes reliably and quickly. This indicates that the class characteristics are represented by the feature vectors in an extensive and precise manner. Therefore, 24 mice per group is concluded to yield sufficient discrimination capabilities at Steps I and II of HC. However, in Step III neither the training performance nor the test performance results present a clear evolution with varying number of training samples. This is due the fact that the six feature vectors for each drug type is not able to represent the class characteristics extensively. However, the performance of proposed method can be improved by introducing additional data.

Since the dataset is shown to represent the class properties for Steps I and II of HC, we can also comment on the convenience of the SVC and LDC classification schemes for these stages of HC. When the performance evaluations presented in Sections 5.1,- 5.4 are examined, it is observed that SVC and LDC are capable to resolve the distinctions between the classes considered in Steps I and II. Nonetheless, several cases of misclassification occur. When we have a closer look at these cases, it is observed that feature vectors, which are labeled with a wrong label by one classifier, are usually mislabeled by the other classifier as well. This consistent mislabeling indicates that the feature vectors and the locomotor activities of the test animals have a dissimilar character compared to the rest of the group. This finding is also confirmed by the medical authorities by examining the locomotor activities, distance and speed curves of these animals. Both classifiers are inferred to be suitable for the solution of this problem. It should be noted that SVC and LDC do not perform equally well in classification. In most cases SVC is observed to perform better than LDC.

In conclusion, the feature vectors and classifiers used in our study proved to be effective and sensitive enough to represent the behavioral characteristics of the animals under the influence of psychotropic drugs. Future work includes investigation of the parameters like bin size, number of grids on performance. Moreover the results presented in Sections 5.3 and 5.4 indicate that a wider dataset is needed to determine the efficiency of the proposed method at Step III of HC.

Part II

Human Behavior Analysis for Attention Resolution

Automatic analysis of human behavior is a very broad research field with a multidisciplinary nature associating, among others, cognitive science, artificial intelligence, and natural language processing. Here, we restrict this broad scope and discuss comprehension of human behavior only from an attention point of view with an application in a human-robot interaction framework.

This narrows the scope of human behavior analysis to a large extent, yet still leaves a considerably broad pursuit. Within this broad field, the identification of behavioral attributes concerning interests, intentions, goals, and desires emerge as the key elements of attention resolution. An efficient resolution of human attention is enabled on a fundamental basis, provided that a profound integration of these factors is achieved.

For the realization of such a system, a recent popular design approach is to imitate the evolution of relevant elements of human comprehension. The advantages of this design approach is multifaceted. It leads not only to an automatic interpretation of human activity and behavior but also to a spontaneous generation of complement response. This introduces the capability of establishing joint attention with humans as well as engaging into natural communication. Moreover, perception of the world is achieved in a natural manner leading to a flexible training scheme. The system learns from the interactions with environment and humans. This results in robust systems resilient in unconstrained settings. Taking these into account, we thus propose a system, that will comprehend and interpret human attention very efficiently. In order to reflect the efficacy and the benefits of the evolutionary perspective to the fullest extent, we consider naturally interacting robotic agents to be a suitable implementation bed. On that account, developmental robotics, which adopts a mutual standpoint of developmental psychology and robotics [5, 106], arises as a convenient design paradigm.

Developmental psychology suggests that, in order to establish social contact and fulfill the desire for knowledge, infants get engaged in communication and hence obtain joint attention with the caregivers. These social skills are observed to improve gradually at primary stages of infancy. It appears that young infants first follow the head movements of others, and only in time develop the ability to follow the gaze direction [27]. Subsequently, infants learn to relate this information with attention [20]. With these lessons learned from developmental psychology, the learning process of infants, as well as the evolution of auxiliary skills that contribute to learning are utilized in training of intelligent agents. Among these skills are those that relate to the construction and restructuring of different types of memory, but also skills to actively explore the sensors and effectors of the agent, and consequently, the environment. These cognitively inspired developmental systems allow the experimenter to obtain naturally interacting embodied agents with inherent evolution of communicative skills. These skills are particularly important for developing language and communication, as well as for *imitation-based learning*, which allows the experimenter to demonstrate a behavior rather than explicitly design algorithms to produce the behavior in the agent.

Recent models of imitation-based learning rely on Meltzoff and Moore's active intermodal mapping framework for action imitation learning [55]. Important work in this area includes [38] and [82], which use Bayesian principles to explore action spaces statistically, followed by gradual learning of action groups and communicative preferences. In [39] a goal-based action model is used to classify intentional actions in a controlled environment. The embodied agent extracts a large number of visual features from the scene and by tracking the trajectory of the experimenter's hand, determines which of the predefined actions is being performed. As a common factor of most research in this area, visual cues are extensively used for implementing working models on embodied agents, and the visual distinctions that can be perceived by the embodied agent serve as affordances [56].

In experiments concerning human robot interaction, the learned structure of a visual scene provides additional cues to the embodied agent in guessing the focus of attention of the communicating party. Hence, most approaches incorporate *saliency* as a part of the joint-attention system, and select appropriate saliency measures that will indicate what is inherently interesting in the scene depending on the application domain. When the saliency of a scene is determined, a visual feedback controller provides motor control commands to direct the gaze of the robot to a salient location, both for attending to the face of the experimenter and to other objects in the environment.

The saliency can be a function of natural image statistics. For instance in [67], a robotic system is described where the bottom-up saliency of a visual scene is computed by color, edge, and motion cues. Top-down influences can also be incorporated by modulating bottom-up channels, or by explicitly adding dedicated saliency components. Faces are particularly important for the natural interaction settings, consequently they are separately detected and made salient.

Considering the natural evolution pattern of interaction of infants, estimation of the head pose and gaze direction of the communicating party together with resolution of saliency are suggested to constitute primary visual skills necessary for joint attention modeling. Like most other works in this field, [67] employs a module that learns to associate facial appearance of the experimenter with angles that specify its pose. However, our study underlines the distinction of head pose and the gaze direction in particular. Inspired by the natural gradual development pattern, an initial cylindrical head model based pose estimator assesses the orientation of the head. The regression module transforms the estimated pose into gaze direction. Subsequently saliency is computed in the prospective region, ascertaining the point of attention fixation and segmenting the governing object.

In Chapter 6, we give a detailed overview of the related studies conducted in the last decade. Chapter 7 presents an outline of the proposed method in a modular framework. These modules are handled in Chapter 8 and Chapter 9. Finally, performance of the proposed method is evaluated quantitatively in Chapter 10.

Chapter 6

Related Work

In order to implement human-like complex social skills, one needs to develop a "theory of mind" model (ToMM) from an intelligent agent design perspective [78]. This requires the decomposition of the communication process into simpler cognitive skills, which can be implemented on an intelligent agent.

In this respect, the task based decomposition proposed by Scassellati presents practical advantages with its functional modularity [77]. According to his hypotheses, the primary tasks are recognition, maintenance of eye contact, and gaze following, which enable attention resolution and engagement into joint attention. Subsequently, imperative and declarative pointing are considered to permit feedback between the infant and the caregiver. Figure 6.1 presents the analogy between the developmental stages of infant learning and ToMM, classifying previous works according to their user, motion and tracking domains.

When we examine the gaze following methods, we see that there are two principal components, which determine the gaze direction to a certain extent. These are head pose and eye locations. Thereby, we would first like to give an overview of the recent models in head pose estimation and eye localization.

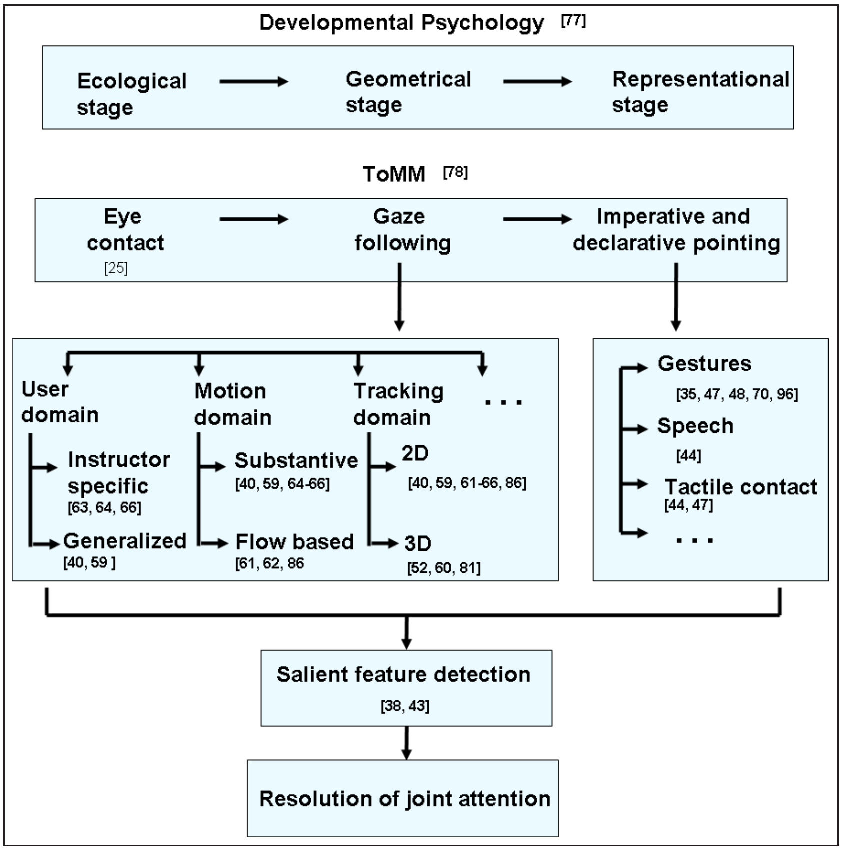


Figure 6.1: Analogy between developmental robotics and ToMM, overview of related work.

Subsequently, we present some integrated models for attention modeling based on naturally inspired development patterns along with pose and eye location estimation schemes.

6.1 Literature on head pose estimation

Throughout the years, different methods for head pose estimation have been developed. The 3D model based approaches achieve robust performance and can deal with large rotations. However, most of the good results are gained in restricted domains, e.g. some systems only work when there is stereo-data available [57, 76], when there is no (self-) occlusion, or when the head is rotating not more than a certain degree [51]. Systems that solve most of these problems usually do not work in real-time due to the complex face models they use [104]. However, if the face model complexity is reduced to simpler ellipsoidal or cylindrical shape, this creates a prospect for a real-time system.

The head pose estimation techniques reported in the literature are based mainly on two different approaches. These are active appearance models (AAM), and cylindrical head models (CHM). AAMs employ a set of reference points to track the head and estimate the pose, whereas CHM fits the face image onto a 3D model of the human head. AAMs have advantages over CHMs in terms of computational complexity. However they are sensitive to the slight changes in initialization and they are not robust against extreme poses. From this point of view, there is a trade off between complexity and robustness.

The individual approaches for head tracking and pose estimation, have been handled extensively in numerous studies. Here, we would like to address several recent works, which propose joint approaches for head tracking and pose estimation [60]. Matsumoto et al. [54] and Newman et al. [68] employ a stereo camera system to obtain 2D feature tracking and 3D model adaptation for tracking and pose estimation. Ba et al. [10] improve precision of pose estimate and accuracy of head tracking by considering these as two coupled problems in a probabilistic setting within a mixed state particle filter framework. They refine this method by fusion of four camera views in [9]. Huang et al. propose to integrate a skin-tone edge-based detector into a Kalman filter based robust head tracker and hidden Markov model based pose estimator in [42]. Hu et al. [41] describe a coarse-to-fine pose estimation method by combining facial appearance asymmetry and 3D head model. A generic 3D face model and an ellipsoidal head model are utilized in [91] and [3], respectively. In [58] an online tracking algorithm employing adaptive view based appearance models is proposed. The method provides drift-free tracking by maintaining a dynamic set of keyframes with views of the head under various poses and registering the current frame to the previous frames and the keyframes. Another AAM based scheme is described by Sung et al. in [87]. They combine AAM with CHM to overcome its drawbacks like sensitivity against large pose variations and initial pose parameters and problems of reinitialization. Similar to [87], we would like to make use of the competent attributes of the CHM. Instead of AAM, we propose using an eye locator in order to broaden the capabilities of the system and improve the precision of individual tracking schemes building a unified framework of modules working in tandem.

6.2 Literature on eye localization

Due to the swift advances in digital technology, eye location estimation has gained considerable importance in recent years. Numerous works in the literature studied the development of systems which can estimate eye location for various scenarios. Our aim is to design a method that is capable of doing accurate eye center localization and tracking in low resolution videos. In order to achieve higher robustness, the method should be able to cope with difficult conditions introduced by extreme head poses.

There are several methods proposed in the literature for eye center location but their common problem is the use of intrusive and expensive sensors [13] and the sensitivity to head pose variations. While commercially available eye trackers require the user to be either equipped with a head mounted device, or to use a high resolution camera combined with a chin rest to limit the allowed head movement, the methods using image processing techniques are considered to be less invasive and so more desirable in a large range of applications. Furthermore, daylight applications are precluded due to the common use of active infrared (IR) illumination used to obtain accurate eye location through corneal reflection. Non-infrared appearance based eye locators [28, 50, 92, 93] can successfully locate eye regions, yet have difficulties in dealing with non-frontal face conditions.

The method used by Asteriadis et al. [6] assigns a vector to every pixel in the edge map of the eye area, which points to the closest edge pixel. The length and the slope information of these vectors is consequently used to detect and localize the eyes by matching them with a training set. Cristinacce et al. [28] use a multistage approach to detect facial features (among them the eye centers) using a face detector, pairwise reinforcement of feature responses, and a final refinement by using active appearance model [26]. Türkan et al. [92] use edge projection [105] and support vector machines (SVM) to classify estimates of eye centers. Bai et al. [11] use an enhanced version of Reisfeld's generalized symmetry transform [73] for the task of eye location. Hamouz et al. [34] search for ten features using Gabor filters, use features triplets to generate face hypothesis, register them for affine transformations and verify the remaining configurations using two SVM classifiers. Finally, Campadelli et al. use an eye detector to validate the presence of a face and to initialize an eye locator, which in turn refines the position of the eye using SVM on optimally selected Haar wavelet coefficients [21].

Very promising to our goals is the method proposed in [93]. This method uses isophote (i.e., curves connecting points of equal intensity) properties to infer the center of (semi-)circular patterns which represent the eyes. However, the accuracy of the eye center location drops significantly in the presence of large head poses. This is due to the fact that the eye structure is no longer symmetric and thus the algorithm delivers increasingly poor performance. This observation suggests that it is desirable to correct the distortion given by the pose so that the eye structure under analysis preserves its symmetry properties. The results will be improved considerably, assuming that there is a way to compensate for the head pose so that we obtain a normalized image patch on which the eye center locator is deployed.

6.3 Literature on integrated head pose and eye location estimation

Several approaches have been reported in the literature for estimating the head pose, eye location and gaze. The authors of [54, 68] consider a tracking scenario equipped with stereo cameras and employ 2D feature tracking and 3D model fitting. The work proposed by Ji et al. [46] describe a real-time eye, gaze, and head pose tracker for monitoring driver vigilance. The authors use IR illumination to detect the pupils and derive the head pose by building a feature space from them. Although their compound tracking property promote them against separate methods, the practical limitations and the need for improved accuracy make them less practical in comparison to monocular low resolution implementations.

The approach proposed in [87] is very relevant to our work. The authors combine a cylindrical head model with an active appearance model approach to overcome the sensitivity to large pose variations, initial pose parameters, and problems of reinitialization. Similar to [87], we would like to make use of the competent attributes of the cylindrical head model together with the eye locator proposed in [93] in order to broaden the capabilities of both systems and to improve the precision of each individual component.

6.4 Literature on attention modeling

A developmental learning model for joint attention is proposed in [63] from a biological point of view. A neural network module is employed in modeling the visual system of the robot, where the layers represent the input, retina, visual cortex, and the output. However, in order to implement a learning scheme, which truly mimics the cognitive development pattern of infants, one should rather go beyond the biological properties. Nagai et al. [64, 66] propose a developmental learning scheme, which improves learning by passing through the ecological, geometric, and representational stages of joint attention [20]. They further improve their system by imposing non-supervised learning condition in an uncontrolled environment, which they name as *bootstrap learning* [65].

These methods, however, are not particularly designed for interaction with multiple people. Although a long training phase with an enormous dataset may lead to user-independent joint attention mechanisms, this is not practically feasible. Thus, a codebook of face images is generated by a self-organizing map and sensorimotor mapping is obtained accordingly in [59]. To perform joint attention with strangers, the robot generalizes its experiences with the caregiver by calculating the similarity between the input and the codebook vectors. Moreover, the self-organizing map makes the learning time shorter for cases with a single caregiver as well and lets the agent communicate with a human asynchronously [40].

The methods that we have mentioned so far treat the video frames as substantive images and omit the temporal connection. Humans, on the other hand, utilize motion information besides static information such as posture and face direction to infer about desires and intentions. For this reason, the robotic agent described in [86] alternates its gaze between a human caregiver and the object he attends by triggering motion using the cues introduced by the motion of the human's face. Besides triggering motion, [62] states that a realistic human-like learning scheme should utilize motion information to estimate gaze shift. The temporal relationship between the frames is expressed in terms of optical flow vectors and thereby a coarse estimate for gaze shift providing initial motor output to follow the gaze is obtained [61].

These approaches formulate visual attention based on the video frames employing the 2D information available. However, the common morphological characteristics can be employed in the derivation of 3D information form the 2D visual input. Since the perception of gaze direction depends to a large extent on head pose [52], one can model the head of the caregiver as a 3D object and resolve for the pose [60]. Hoffman et al. [38] employ an ellipsoidal model for human head and the inferred head angles are used in the estimation of the gaze vector. Saliency computation is performed around the estimated gaze incorporating the instructor-specific priors [81].

These methods mimic early stages of cognitive development of infants, i.e., mainly 6 to 12 months. Reciprocal communication, which is a part of the later stages, is achieved employing auxiliary modules such as person identification, speech recognition and synthesis along with natural aligned gestures [43], mutually entrained body movements and complex eye movements in [70, 44, 47, 48].

In this study, we model attention mimicking the early stages of infancy in robot learning and making use of the principals of developmental robotics. The modules of the theory of mind model described by Scassellati [78] is implemented from a robotic design perspective by integrating the eye localization method of Valenti et al. [94] and Xiao et al. [98]. In this way, we solve for the gaze direction and target depth. Subsequently, saliency is computed in the prospective region and the attention fixation points are estimated.

Chapter 7

Outline and Experiments

Estimating focus of attention receives a lot of interest for obvious reasons. Gaze direction estimation is regarded as the primary determining constituent for that purpose and has been used as a common supplementary component in numerous studies. Therefore, we handle the problem from a gaze point of view and derive attention fixation points accordingly.

In estimation of gaze direction, a popular approach is to employ the eye locations and in particular iris information (see [35] for a recent overview of eye and gaze models). However, this requires a decent view of the eye region and in particular the 3D model based methods need a considerably high resolution of the eye area [99]. However, the embodied agent is not sufficiently stable to extract an accurate estimate of the gaze direction only by analyzing the eye and iris area of the experimenter. On the other hand, the resolution of the input video does not provide a sufficiently good image of the eye region. It is clear from Figure 7.1, which illustrates the eye regions cropped from different experimenters' frontal views, that one cannot make a reasonable estimate for gaze direction using these patches. Furthermore, the experimenter does not always present a frontal view with a distinct image of the eye region and the eye ball. In particular, gazing sideways changes the image of the eye drastically, making it very hard to distinguish the direction. Therefore, instead of completely depending upon eye and iris information, we make use of the eye region as a supplementary constituent whenever it is available. Alternatively, we seek to determine the gaze direction basically from head pose estimates, by making the assumption that head pose is indicative of but not equal to the gaze direction. Consequently, gaze direction is derived mainly from the head pose estimates through proper regression and is enhanced with eye locations, when there is reliable information.

(a)

(b)

(c)

(d)

Figure 7.1: Eye regions for the four experimenters. (Approximately 15 × 25 pixels).

This chapter outlines the basic components of the described attention resolution scheme. The experimental setup is detailed in Section 7.1, whereas the organization of the main building blocks and the coupling in between is explained in Section 7.2.

7.1 Experiment environment

We performed a set of experiments to resolve human attention fixation points and to model joint attention between a human caregiver and an embodied agent.

Considering all the design details given in Chapter 6, the embodied agent of the Artificial Intelligence Laboratory of Boğaziçi University by Çetin Meriçli and Tekin Meriçli [102] is used. This agent is designed to be used for service and guiding purposes, is regraded as a suitable application platform (see Figure 7.2).

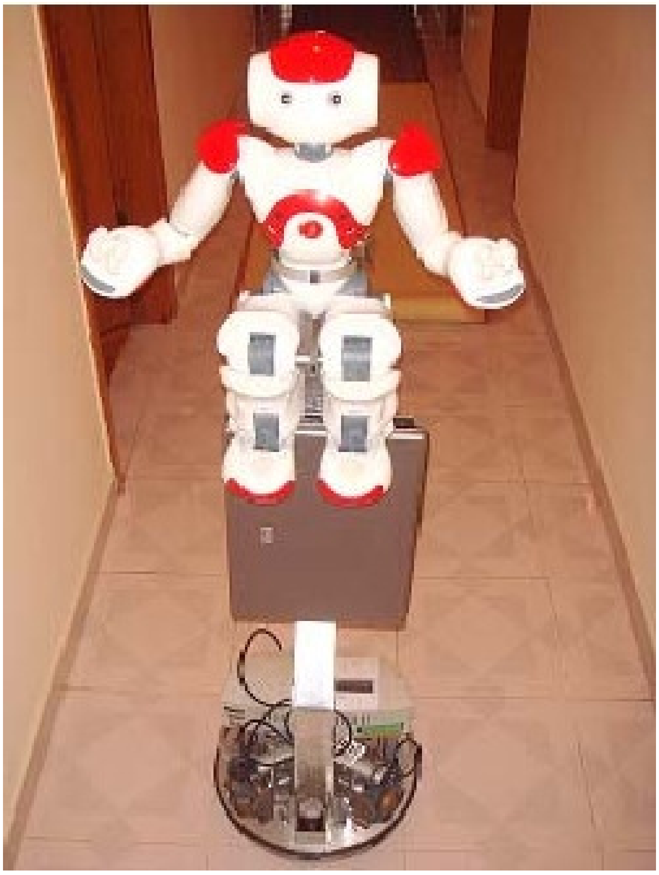


Figure 7.2: The robot platform used in the experiments.

The system is composed of three main components:

- The Aldebaran Nao humanoid robot is the main interaction and animation unit [1]. Aldebaran Nao is a 23″ tall humanoid robot with 25-DOF in total, two vertically aligned color cameras with 640 × 480 resolution, and a 500Mhz Geode processor. A Linux based operating system is running on the robot and pre-installed text-to-speech packages allow the robot generate speech. In our design, we utilize the upper torso of the Nao robot and use the Robotino robot to make the whole body wander around.
- The FESTO Robotino robot is used as the navigation unit [75]. Robotino is a wheeled robot capable of moving omnidirectionally. It is surrounded by 9 IR sensors and a bump sensor, and it has a 300Mhz processor. Also a 5 meters range Hokuyo laser range finder is installed on the body of the Robotino robot to have more accurate range data from the robot's environment.
- A laptop computer provides additional processing and and serves as a monitoring unit.

The experiment scenario is the following. The experimenter stands in front of a table, on which six objects are placed in a non-occluding fashion as in Figure 7.3. The robotic platform, which is located at the opposite side in approximately 2 meters distance to the experimenter, has a view of this complete scene. As the experimenter fixates his/her attention to one of the six objects by looking at them in random order for a certain duration of time, it records the scene at a frame rate of 15*fps*. Subsequently, the attention resolution algorithm, which is detailed in Section 7.2, is employed by the robotic agent.

Eight experiments are carried out with the described scenario, where four different experimenters provide two sequences each. In each of these eight experiments, all of the six objects are attended for several seconds in a random order at least for once. The video sequences are composed of 1804 frames in total and they are recorded at 15*fps* frame rate. The ground truth for the attended

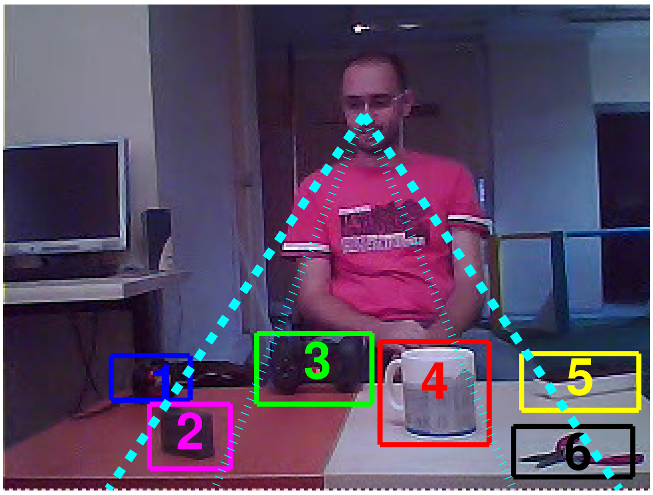


Figure 7.3: The experimental setup, object indices and clustering schemes.

objects is obtained by manual annotation, whereas the ground truth for the gaze direction follows as the slope of the line connecting the center of the annotated object and the head center resolved by the CHM.

7.2 Outline of the method

Our proposed approach mimics the natural strategy of resolution of focus of attention observed in infants. The basic steps of the proposed algorithm are summarized in Figure 7.4.

The experimenter initially provides a frontal view of his/her face and initializes a session of joint attention. This condition approximates the initialization of natural communication between humans. In addition to that, it facilitates the initialization of head pose and localization of eye centers. The first step is detecting the face of the experimenter with the Viola-Jones algorithm [95]. Subsequently, the eye locations are solved using a isophote based eye localization method. The head pose of the experimenter is initialized in line with these resolved eye locations and resolved by adapting a 3D elliptic cylindrical model to

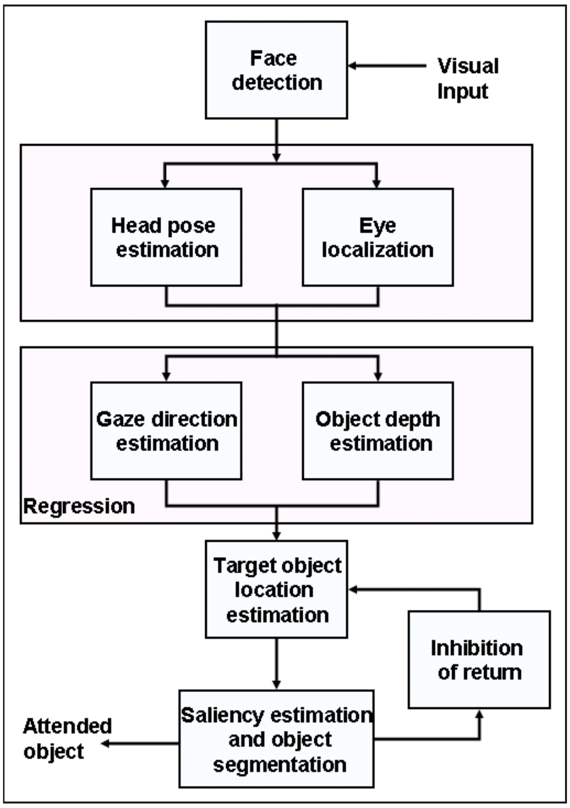


Figure 7.4: Basic steps of the algorithm.

the face region. By applying pose update, which is derived using Lucas-Kanade optical flow method, continuous tracking of head pose is maintained.

Two Gaussian Process (GP) regressors are employed in estimation of gaze direction and the distance of the target object along the gaze vector from these pose values. We stress the distinction between following the head pose and the gaze direction itself. Most of the joint attention approaches in the literature do not explicitly correct for the discrepancy between the head pose and gaze direction, which is reported to be normally distributed with a mean of five degrees in natural settings in [36, 90].

The two estimates concerning gaze direction and object depth are then probabilistically combined to yield a coarse estimate for the center of the target object. By pooling a number of estimates concerning corresponding frames, a more robust decision on the target is generated. The rough localization of the attended object is refined by a bottom-up saliency scheme, which also segments out the target object. If the experimenter continues to maintain a certain head pose, alternative target locations are eventually explored as a result of an inhibition-of-return mechanism.

Chapter 8

Head Pose Estimation and Eye Localization

Head pose and eye location estimation are two principal components of human attention resolution. In recent years, these problems have been studied individually in numerous works and several competent methods have been proposed for each.

In this chapter, we employ two of these competent methods and exploit physiologic characteristics of humans to build a basis for their nested architecture. This incorporation improves the performance of the individual methods in a profound integration framework. In this manner, we eliminate several shortcomings of the independent methods, which come into picture under certain circumstances like extreme poses. In addition to this, we introduce several enhancements, like extension of the operating range and improvement of precision and reinitialization capabilities.

As discussed in Chapter 6 previous research indicates that cylindrical head models and isophote based eye localization schemes are two prominent methods in head pose and eye location estimation, respectively. For head tracking and pose estimation, we employ the method of Xiao et al. [98] and combine their elliptic cylindrical model based pose estimator with the eye locator of Valenti et al. [93]. A feedback mechanism is established between the two components by evaluating the tracking quality constantly at each stage. As soon as it is detected that the head pose estimation and eye localization results yield conflicting outcomes, the tracking system is re-initialized and two modules are adjusted to get in line with each other. In [94], the tracking quality of each component method is shown to improve significantly within this unified framework.

The outline of the chapter is as follows. In Section 8.1, the details of head tracking and pose estimation algorithm is is presented, Section 8.2 gives an overview of the eye localization method. The integrated framework is described in Section 8.3.

8.1 Head pose estimation

The cylindrical head model approach has been used in a number of studies [18, 51, 98]. Among those, the implementation of Xiao et al. stands out with its superior performance [98].

Since the operational real-time requirement is imperative for the application framework, a number of simplifying assumptions are made in the formulation of the problem. The basic assumption concerns the morphological structure. Namely, human head is modeled as an elliptic cylinder, with the actual width of the head and the radii in line with the anthropomorphic measures [33, 98].

The Viola-Jones algorithm, which uses the Adaboost classifier with Haar wavelet features searches for a face on video frames until one is detected [95]. Subsequently, the cylindrical head model is accommodated on the detected face area. The pose of this 3D model on frame F_i is represented by a vector $\vec{p_i}$, which is a collection of rotation and translation parameters

$$\vec{p_i} = [r_x^i \ r_y^i \ r_z^i \ t_x^i \ t_y^i \ t_z^i]^T.$$

The initial values for these parameters are determined employing the initiation condition of a session of joint attention. In our scenario, we describe this condition as establishment of eye contact between the agent and the experimenter with a fully frontal view of the experimenter's face. Thus the pitch r_x^0, roll r_y^0, and yaw r_z^0 angles are all set to 0. The translations along $\mathbf{x}-$ and $\mathbf{y}-$ axes, t_x^i, t_y^i, are initialized in relation with the face region obtained by the Haar classifier after a normalization with respect to the center point of the image. Moreover, the depth of the head, t_z^0, which describes the distance of the head from the camera, is set to an approximate fixed value.

As soon as the eye contact is established between the experimenter and the embodied agent, the head pose is initialized as described and head tracking starts. Meanwhile, the agent estimates the head pose and applies a regression scheme on pose values to derive gaze direction. The estimation scheme is explained in detail in Chapter 9. Thereby, real time tracking is maintained in addition to pose and gaze direction estimation. Since the initial pose $\vec{p_0}$, is already determined employing the initiation condition, any pose value $\vec{p_{i+1}}$, $i \geq 0$, can be resolved by simply updating the previous value $\vec{p_i}$ [98]. An operator $\mathbf{M}$ is used to apply the pose update, $\vec{\Delta\mu_i} = [\omega_x^i, \omega_y^i, \omega_z^i, \tau_x^i, \tau_y^i, \tau_z^i]$, on $\vec{p_i}$ to derive $\vec{p_{i+1}}$

$$\vec{p_{i+1}} = \mathbf{M}(\vec{p_i}, \vec{\Delta\mu_i}),$$

which will be specified below. However one should note that a 2D image retrieved from a video sequence is employed in derivation of pose values representing the direction and orientation in 3D space. In order to cope with the ambiguity ensuing from the dimensionality disparity, a suitable mapping needs to be defined. For that purpose, we suggest using perspective projection and ray tracing through a pin hole camera. By this means, the relation between the 3D locations of the points on the cylinder and their corresponding projections on the 2D image

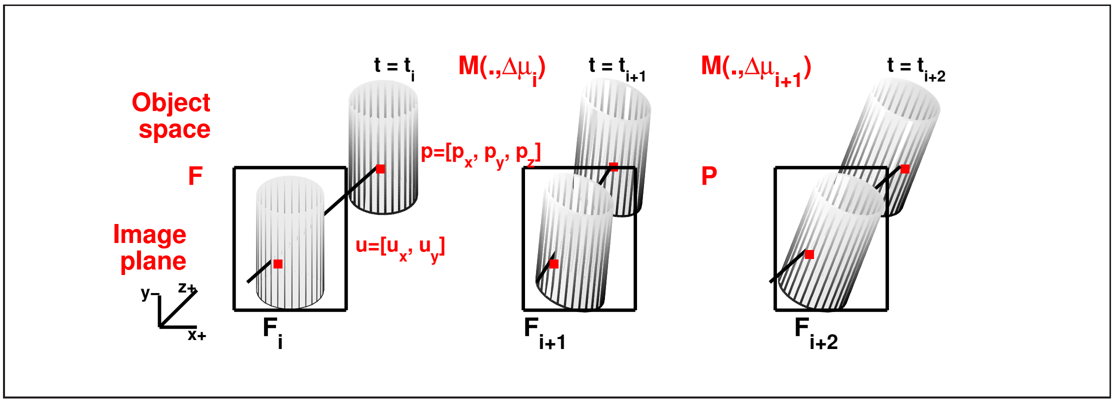


Figure 8.1: Perspective projection, illustrated for three consecutive frames in which the head moves and tilts.

plane is established (see Figure 8.1). The 3D coordinates of a set of points on the frontal part of the cylindrical head model are ascertained with respect to the reference frame. Any point $\mathbf{p} = [p_x\ p_y\ p_z]^T$ on the elliptic cylinder is known to satisfy the following equation,

$$\left(\frac{p_x}{\rho_x}\right)^2 + \left(\frac{p_z}{\rho_z}\right)^2 = 1, \tag{8.1}$$

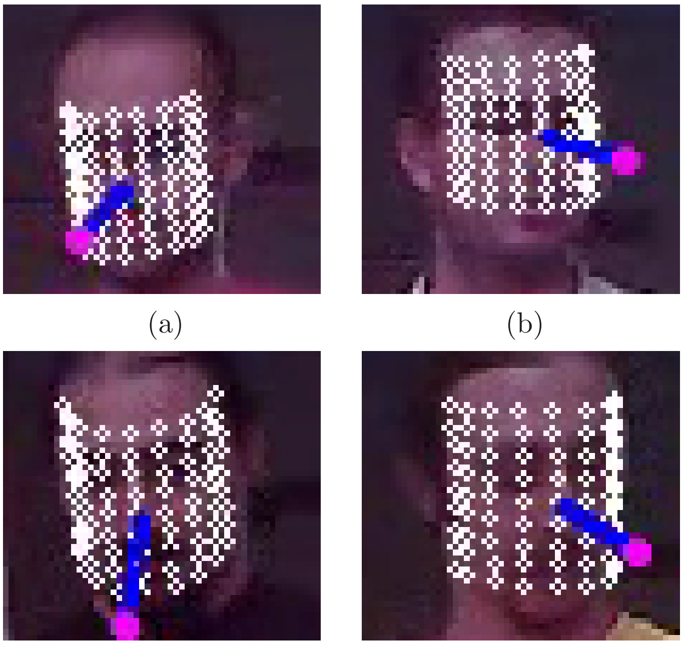
where $\mathbf{r_x}$ and $\mathbf{r_z}$ stand for the radii of the ellipse along $\mathbf{x}-$ and $\mathbf{z}-$axes, respectively. In order to get the coordinates of the points on the visible part of the cylinder, the front region is sampled in an $N \times N$ grid-like structure on $\mathbf{x}-\mathbf{y}$ plane (see Figure 8.2). The corresponding depth values are obtained using ray tracing technique, which traces the path of light through pixels in an image plane.

Let the starting point and the direction of the ray be $\mathbf{q} = [q_x\ q_y\ q_z]^T$ and $\mathbf{d} = [d_x, d_y, d_z]^T$, respectively. Let the point that the ray hits on the cylinder be a point sampled at $\mathbf{p} = [p_x\ p_y\ p_z]^T$. If the ray is considered to travel for a duration of t, then

$$\mathbf{p} = \mathbf{q} + \mathbf{d}t.$$

From the elliptic cylindrical model assumption, it follows that p_x and p_z satisfy Equation 8.1, so that,

$$\left(\frac{q_x + d_x t}{\rho_x}\right)^2 + \left(\frac{q_z + d_z t}{\rho_z}\right)^2 = 1. \tag{8.2}$$



(a) (b)

(c) (d)

Figure 8.2: Cylindrical head model for several frames from four video sequences with four different experimenters.

While passing through, the ray intersects the cylinder at entry and exit. Hence, the quadratic Equation 8.2 has two roots. We consider the solution with the smaller absolute value, as it indicates a point closer to the camera asserting a visible point.

Let $\mathbf{u} = [u_x \; u_y]^T$ be a point on the image plane and $\mathbf{p} = [p_x \; p_y \; p_z]^T$ be the corresponding point sampled from the cylinder as in Figure 8.1. Figure 8.3 illustrates the side view of this setting by making a pin hole camera assumption for the sake of simplification.

Using similarity of triangles in Figure 8.3, the following equations apply for the relation between $\mathbf{p}$ and $\mathbf{u}$,

$$p_x = \tfrac{p_z u_x}{fl},$$

$$p_y = \tfrac{p_z u_y}{fl},$$

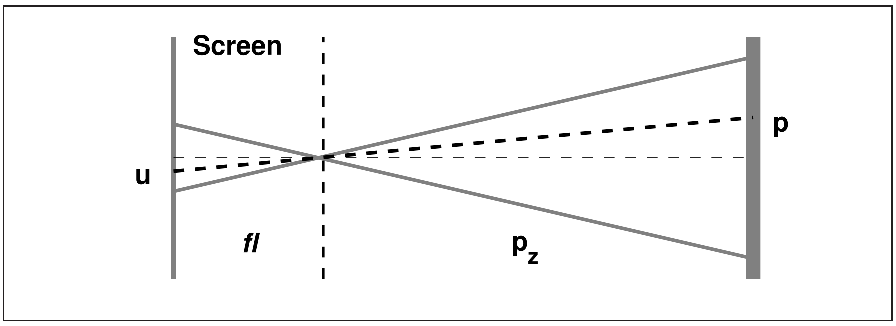


Figure 8.3: Pin hole camera model.

where fl stands for the focal length of the camera. This relation is summarized by the perspective projection function $\mathbf{P}$

$$\mathbf{P}\left(\mathbf{p}\right) = \mathbf{u},$$

by which we formulate the relationship between the 3D locations of the points sampled on the cylinder and their corresponding projections on 2D image plane.

As seen in Figure 8.1, the cylinder is observed at different locations and with different orientations at two consecutive frames F_i and F_{i+1}. This is expressed as an update on pose vector $\vec{p_i}$ by the rigid motion vector $\Delta\vec{\mu}_i$.

Let $\pi_{\mathbf{i}}$ denote the 3D location of a point sampled on the cylinder on frame F_i. The new location of the point at F_{i+1} is found by applying the transformation model, $\mathbf{M}$, which is defined by a rotation matrix R corresponding to $[\omega_x^i\ \omega_y^i\ \omega_z^i]^T$ and a translation vector $T = [\tau_x^i\ \tau_y^i\ \tau_z^i]^T$ as

$$\pi_{i+1} = \mathbf{M}(\pi_{\mathbf{i}}, \Delta\vec{\mu}_i) = R\pi_{\mathbf{i}} + T.$$

The location of the projected point $\mathbf{u}_{i+1}$ on F_{i+1} is found by using the 2D parametric function $\mathbf{F}$ and applying the rigid motion vector $\Delta\vec{\mu}_i$, which summarizes the motion between time instants t_i and t_{i+1},

$$\mathbf{u}_{i+1} = \mathbf{F}(\mathbf{u}_i, \Delta\vec{\mu}_i).$$

If illumination is assumed to be constant (i.e., if the intensity of the pixel $I(u)$ does not change between the images), then

$$I(\mathbf{u}_i) = I(\mathbf{u}_{i+1}) = I(F(\mathbf{u}_i, \vec{\Delta\mu_i})).$$

Thus the rigid motion vector can be obtained by minimizing the difference between the two image frames,

$$\min\left(E\left(\vec{\Delta\mu}\right)\right) = \sum_{\mathbf{u}_{i+1}\in\Omega} \{I\left(\mathbf{F}\left(\mathbf{u}_i, \vec{\Delta\mu_i}\right)\right) - I(\mathbf{u}_i)\}^2,$$

where Ω stands for the set of points sampled on F_i, which are still visible on F_{i+1}. The minimization problem is solved by Lucas-Kanade method [53],

$$\vec{\Delta\mu_i} = -\left[\sum_{u\in\Omega}(I_u F_{\vec{\Delta\mu}})^T (I_u F_{\vec{\Delta\mu}})\right]^{-1} \sum_{u\in\Omega}(I_t (I_u F_{\vec{\Delta\mu}})^T),$$

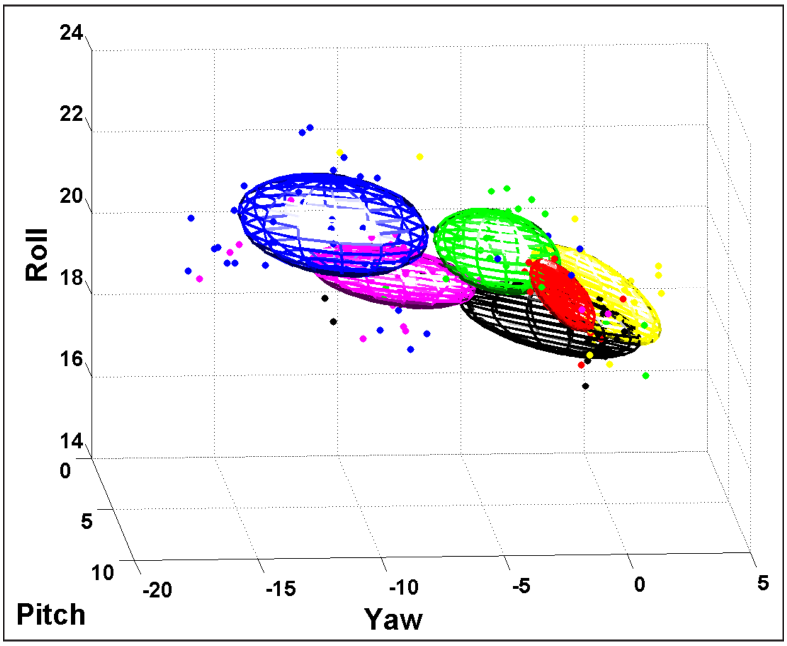
as I_u an I_t are the spatial and temporal image gradients [53]. Thus, the projection of the point at time instant t_{i+1} can be expressed in terms of the 3D location of the point at time instant t_i and the rigid motion vector as,

$$\mathbf{u}_{i+1} = \mathbf{P}(\mathbf{M}(\mathbf{p}_i, \vec{\Delta\mu})).$$

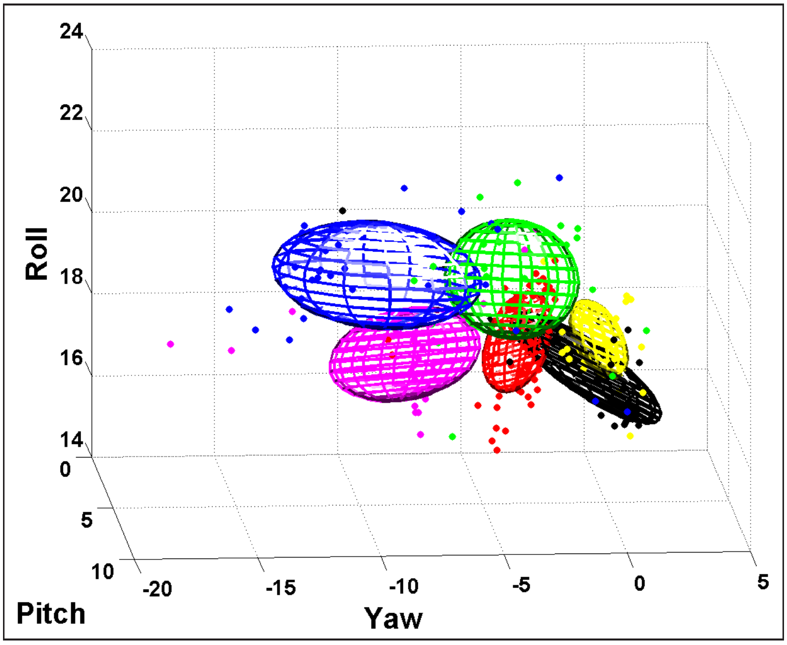
The mapping is thus formulated in a comprehensive way, covering transformations from object space to image plane ($\mathbf{P}$), image plane to object space ($\mathbf{F}$) and inter-frame motion ($\mathbf{M}$).

The pose vectors, which are computed in the above manner, have a distribution as illustrated in Figures 8.4-(a) and 8.4-(b) for two exemplary video sequences. As the pose distributions are modeled with Gaussians with diagonal covariance matrices, the indicated regions in 3D come into view.

The clear distinction in the distribution of head pose angles concerning different objects supports the proposition that there exists a suitable regression scheme to map the pose values to gaze directions.



(a)



(b)

Figure 8.4: Pose distributions for two exemplary video sequences.

8.2 Eye localization

The head pose estimation method described in Section 8.1 is incorporated with the eye center localization method of Valenti et al. [93], which exploits the fact that the eyes are characterized by radially symmetric brightness patterns. In this respect, the computation of the eye center locations is obtained through the curvature of the isophotes on an image frame. The curvature of the isophote curves κ is given as,

$$\kappa = -\frac{L_y^2 L_{xx} - 2L_x L_{xy} L_y + L_x^2 L_{yy}}{(L_x^2 + L_y^2)^{3/2}},$$

where the illumination function is denoted by L, and L_x and L_y stand for its derivatives in the x and y directions. The displacement vector $D(x, y)$ indicates the distance between the estimated center of the circle and any point p on the isophote curve,

$$D(x, y) = -\frac{\{L_x, L_y\}(L_x^2 + L_y^2)}{L_y^2 L_{xx} - 2L_x L_{xy} L_y + L_x^2 L_{yy}}.$$

By using these parameters, an estimated center is calculated for any point on the isophote curve together with the direction and orientation regarding this center. These values are employed in the voting scheme described in [93] and a center map is obtained for the video frame of interest. Since the pixels, which correspond to the eye center locations, receive higher votes, they stand out in the center map. Thereby, the left and right eye locations, E_l and E_r, are resolved.

After the cylinder is initialized in the 3D space, the 2D eye locations are detected on the first frame of video employing the described method. These initial eye locations are considered as reference points for the following video frames. They are projected onto the cylindrical head model and the concerning depth values are calculated. The reference points are then used to estimate the successive eye locations.

8.3 Integrated head pose estimation and eye localization

The CHM head pose estimation and the isophote based eye localization methods have advantages over the other previously reported methods. However, taken separately, they face several drawbacks under certain circumstances. In [93], the authors claim that the system is robust to slight changes in head pose. However, it cannot cope with extreme head poses, since the eye regions is not semi-frontal any more. On the other hand the CHM tracker might erroneously converge to local minimas and might not able to recover the correct track in some cases. By integrating the eye locator with the cylindrical head model we eliminate these drawbacks. Instead of a sequential implementation of the two systems, a deeper integration is proposed by comparing the transformation matrices suggested by these interpenetrated systems. The eye locations and the head pose estimated are adjusted in such a way that they lie in excellent agreement as it is outlined in Algorithm 1.

Steps 1- 4 of Algorithm 1 refer to the initialization process, which is composed of detection of face region, reference eye locations, and initialization of the cylinder model and pose values. As soon as a face image is detected, the video is processed frame by frame (Line 4). For each frame initially the head is assumed to preserve its previous location and pose(Line 6- 9). Subsequently the error in the face region is compared to a certain threshold. If a significant difference is observed, an iteration is carried out to resolve the pose update (Line 10). If the difference is regarded to be insignificant, we go on processing the next frame.

The lines 11- 21 give details of pose update process using the Lukas-Kanade optical flow method explained in Section 8.1. The lines 22- 31 checks whether the resolved pose agrees with the new eye locations. The reference points are mapped onto the cylindrical model and their locations are updated using the resolved pose vector. Subsequently, an area is sampled around the updated reference points and a normalized canonical view is calculated according to the transformation matrix coming from CHM tracker. The eye locations are ascertained on this normalized eye region using a modified version of the method of Valenti et. al. [93]. The main difference lies in the elimination of meanshift algorithm involved in the voting process of [93]. Namely instead of using the meanshift algorithm [24] to estimate the area with the maximum density of votes, the highest peak in the center map, which is closer to the center of the eye region (therefore closer to the reference eye location obtained by pose cues), is selected as estimated eye center. In this way, the localized eyes can be considered to be optimal as long as the CHM tracker is correctly estimating the head pose.

The cylinder is adapted on the new eye locations based on the rigid model assumptions and the corresponding pose vector is calculated. If the condition at Line 29 indicates that this pose vector is in close proximity to the one one resolved by CHM, we go on processing the next video frame. Otherwise, we apply an interpolation on the pose vectors and update the eye locations and the transformation matrix **M** accordingly.

Thereby the cylindrical model is adjusted to an orientation recovering the correct track. A modified version of the eye locator of [93] constantly verifies that the eye location found by pose cues is consistent with the one obtained without pose cues. Thus, as in [58], when reliable evidence (e.g., the eye location in a frontal face) is collected and found to be in contrast with the tracking procedure, the latter is adjusted to reflect this evidence. Therefore, the CHM tracker and the eye locator interact and adjust their own estimations by using each other's information. This synergy between the two systems allows for an initialization-free and self-adjusting system. In [94], a detailed discussion regarding the performance of the proposed scheme is provided with a large collection of experiments.

Algorithm 1: Integrated head pose estimation and eye localization.

```
Input: Video sequence
Output: Head pose estimates and eye center locations
/* Initialize parameters                                              */
- Detect face region [96]
- Get initial eye locations [93], i.e. reference points $E_l$ and $E_r$
- Initialize cylinder height and radii [33]
- Initialize $\vec{p_0}$ as in Section 8.1
/* Iterate through all the frames                                     */
for $i \leftarrow 0$ to total frame number do
    - Set $I_{i+1} \leftarrow I_i$ ;                /* Assume no intensity change */
    - Compute the image gradient $\bigtriangledown I_{i+1}$
    - $\vec{p_{i+1}} \leftarrow \vec{p_i}$ ;                        /* Assume no pose change */
    - Initialize $\vec{\Delta\mu} \leftarrow [0, 0, 0, 0, 0, 0]$
    while maximum iterations not reached $\vee$ $\vec{\Delta\mu} <$ threshold do
        - Transform face region points $\mathbf{p}$ of $I_i$ with $\mathbf{M}$
        - Update and normalize face region $\forall \mathbf{u}$
        - Calculate $p_z$ for $\forall \mathbf{p}$ ;                       /* Ray tracing */
        - $p_x \leftarrow \frac{u_x * p_z}{fl}$, $p_y \leftarrow \frac{u_y * p_z}{fl}$ ;      /* Perspective projection */
        - Use inverse motion on $\mathbf{p}$
        - $\mathbf{u} \leftarrow \mathbf{P}(\mathbf{p})$.
        - Compute $T$, Jacobian of $\mathbf{M}$.
        - $H \leftarrow \sum \left[\bigtriangledown I_{i+1} \frac{\partial T}{\partial p}\right]^T \left[\bigtriangledown I_{i+1} \frac{\partial T}{\partial p}\right]$ ;        /* Hessian matrix */
        - $\vec{\Delta\mu} \leftarrow -H^{-1} \sum \left[ \ \bigtriangledown I_{i+1} \frac{\partial T}{\partial p} \ \right]^T \sum \left[ \ I_i - I_{i+1} \ \right]$
        - $\mathbf{p_{i+1}} \leftarrow \mathbf{p_i} * \vec{\Delta\mu}$ ;                            /* Update pose */
        - $\mathbf{M} \leftarrow \mathbf{M} * \vec{\Delta\mu}$ ;                               /* Update M */
        /* Verify the new eye locations in close proximity            */
        - Transform reference points $E_r$ and $E_l$ using $\mathbf{M}$
        - Remap eye regions to pose normalized view
        - $D(x, y) \leftarrow -\frac{\{\frac{\delta I}{\delta x}, \frac{\delta I}{\delta y}\}(\frac{\delta I}{\delta x}^2 + \frac{\delta I}{\delta y}^2)}{\frac{\delta I}{\delta y}^2 \frac{\delta^2 I}{\delta x^2} - 2\frac{\delta I}{\delta x} \frac{\delta^2 I}{\delta x \delta y} \frac{\delta I}{\delta y} + \frac{\delta I}{\delta x}^2 \frac{\delta^2 I}{\delta y^2}}.$ ;                     /* See  [93] */
        - Vote for centers weighted by $\sqrt{\frac{\delta^2 I}{\delta x^2}^2 + 2\frac{\delta^2 I}{\delta x \delta y}^2 + \frac{\delta^2 I}{\delta y^2}^2}$.
        - Select isocenter closer to the center of eye region as eye estimate
        - Remap eye estimate to cylinder reference frame
        - Get an alternative pose vector from eye locations
        if Distance between two poses > threshold then
            - Interpolate pose vectors
            - Update $\mathbf{M}$
```

Chapter 9

Joint Attention Modeling

This chapter describes a transformation scheme that enables derivation of attention fixation points from the head pose estimates and a decision strategy together with a segmentation method, which work in tandem to resolve the attended objects [100, 101, 102].

Before exploring a suitable regression scheme, one should search for reliable indications, which indicate the feasibility of such a mapping. Hence we closely examine the distribution of the pose angles illustrated in Figure 8.4, in relation to the orientation of the objects in Figure 7.3. This comparative inspection points to preservation of the topological connection between the localization of the objects and corresponding head pose values. It is inferred that head pose and gaze direction are closely associated permitting a transformation with reasonable performance.

We also need to identify the factors that characterize the attention fixation points precisely and then seek for a suitable regression scheme to derive those from the head pose estimates. In that respect, gaze direction is considered as the main identifying element of focus of attention. It is denoted with α and is defined as the slope of the vector that connects the head center resolved by

CHM and the center of the object of interest annotated by the user. The initial location of the head center, $\mathbf{p_c^0}$, can be updated at the i^{th} video frame using the pose corresponding pose vector $\vec{p_i}$. Subsequently the transformed location $\mathbf{p_c^i}$ is mapped onto the image plane as,

$$\mathbf{P}\left(\mathbf{p_c^i}\right) = \mathbf{u_c^i}.$$

Let the center of the annotated object of focus for the i^{th} frame be at $\mathbf{o_c^i}$. Gaze direction is formulated in terms of $\mathbf{p_c^i}$ and $\mathbf{o_c^i}$ as,

$$\alpha = \tan \frac{\mathbf{u_c^i}(y) - \mathbf{o_c^i}(y)}{\mathbf{u_c^i}(x) - \mathbf{o_c^i}(x)}$$

However, gaze direction alone provides only a preliminary specifier, which is the course that attention is channeled towards. In order to make a distinction among the set points lying along this direction, depth of field needs to be determined. For that purpose, we define object depth, δ, as the y coordinate of the object center, i.e., $\mathbf{o_c^i}(y)$.

The intersection of these two factors, α and δ, gives an initial estimate for the attention fixation points. Moreover, due to the fact that these estimates lie on certain objects in our application scenario, we can employ saliency analysis for the determination of attended objects. In the next section, we give a brief overview of the regression process and Sections 9.2 and 9.3 elaborate on the derivation of initial estimates and integration saliency computation.

9.1 Gaze direction and target depth estimation

The grouping of the pose angles in Figure 8.4 with respect to the target objects reveals not only a clear clustering but also the nonlinear nature of the relation between head pose and attention direction. Some approaches deal with this issue by incorporating additional assumptions. For instance in [85], the focus of attention is assumed to rest on a person, and the estimated head pose is corrected to select the closest person as the target of the gaze. However, this approach cannot be applied to our scheme directly, due to the differences in the experiment scenario. Therefore, we suggest to remedy this disparity thorough a nonlinear formulation. For this purpose, we employ Gaussian process (GP) regression in this study. By this means, the gaze direction and depth of the attended object are interpolated from head pose estimates through a strict formulation rather than an environment dependent compensation [72]. In what follows, we describe details of GP regression, covariance functions, model selection, and adaptation of free parameters.

9.1.1 Gaussian process regression

Let the variable x denote a pose vector and y denote the corresponding scalar target value, which represents α or δ depending on the formulation. Our aim is to explore the characteristics of the relation between x and y. The details of the explicit association are disclosed adopting a Gaussian process model for the underlying structure.

Assume that the target values y are obtained from the head pose estimates x using a transformation $f(.)$

$$y = f(x). \tag{9.1}$$

The transformation f is assumed to come from a distribution identified as a Gaussian process so that it is completely characterized by the mean function $m(x)$ and the covariance function, $k(x, x')$

$$f(x) \sim \mathcal{G}P(m(x), k(x, x')),$$

where

$$m(x) = E[f(x)],$$

$$k(x, x') = E[(f(x) - m(x))(f(x') - m(x'))],$$

and $E[.]$ denotes the expected value [72]. For notational simplicity, let $m(x) = 0$ at all times, which could easily be achieved by applying an offset.

The observations are also postulated to be noisy, which makes the scenario close to the real life settings. This may be regarded as accounting for the eye movements by considering their effect as additive noise. This statement is formalized by extending Equation 9.1 as

$$y = f(x) + \varepsilon_0, \tag{9.2}$$

where ε_0 stands for the independent and identically distributed (*i.i.d.*) white noise with variance σ_0^2. Suppose that we have a training set $\mathcal{D}$ with n observation pairs,

$$\mathcal{D} = \{(x_i, y_i) | 1 \leq i \leq n\}.$$

As the inputs are aggregated in the matrix X, and the targets are collected in the vector $\mathbf{y}$, the entire information encompassed by $\mathcal{D}$ is

$$\mathcal{D} = (X, \mathbf{y}).$$

The observation set $\mathcal{D}$ is employed in *training* the Gaussian process regression model. This means that model selection and optimization of parameters are carried out utilizing the information provided by $\mathcal{D}$. We adopt a Bayesian approach by which the conditional distribution of the targets given the inputs can be resolved.

Suppose also that X is composed of n training samples and $X^\star$ is composed of $n^\star$ test points. Hence one can state

$$\mathbf{f}^\star \sim \mathcal{N}(0, K(X^\star, X^\star)).$$

as K denotes the covariance matrix. The prior joint distribution of training outputs $\mathbf{y}$ and the test outputs $\mathbf{f}^\star$ is then,

$$\begin{bmatrix} \mathbf{y} \\ \mathbf{f}^\star \end{bmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} K(X,X)+\sigma_0^2 I & K(X,X^\star) \\ K(X^\star,X) & K(X^\star,X^\star) \end{bmatrix}\right).$$

The posterior distribution is obtained by restricting the prior joint distribution to contain only those functions that agree with the observations. Hence, as the prior is conditioned on the observations, we get,

$$\mathbf{f}^\star | X, \mathbf{y}, X^\star \sim \mathcal{N}(\bar{\mathbf{f}}^\star, \mathrm{cov}(\mathbf{f}^\star)),$$

where,

$$\begin{aligned} \bar{\mathbf{f}}^\star &= K(X^\star,X)[K(X,X)+\sigma_0^2 I]^{-1}\mathbf{y}, \\ \mathrm{cov}(\mathbf{f}^\star) &= K(X^\star,X^\star) - K(X^\star,X)[K(X,X)+\sigma_0^2 I]^{-1}K(X,X^\star). \end{aligned} \tag{9.3}$$

Since the mean prediction is a linear combination of the observations, Equation 9.3 is sometimes referred as the *linear predictor* [72].

9.1.2 Model selection

As mentioned in Section 9.1, our aim is to disclose the details of the relationship between x and y, i.e., the structure of the Gaussian process f. In that respect, it is commonly practiced to adopt a hierarchical approach. Let us assume that at the uppermost level is the model structure $\mathcal{H}$ that f belongs to. One level below, stand the hyper-parameters θ, which indicate the distribution of model parameters. The model parameters are denoted with $\mathbf{w}$ and stand at the bottom level.

The term *model selection* refers to discrete choices like model structure $\mathcal{H}$ as well as optimization of hyper-parameters θ, since both of these problems are treated in the same manner, i.e., using a Bayesian approach.

The Bayes' rule gives the posterior distribution at the bottom level, in terms of the prior, likelihood, and marginal likelihood as follows,

$$p(\mathbf{w}|\mathbf{y}, X, \theta, \mathcal{H}_i) = \frac{p(\mathbf{y}|X, \mathbf{w}, \mathcal{H}_i)p(\mathbf{w}|\theta, \mathcal{H}_i)}{p(\mathbf{y}|X, \theta, \mathcal{H}_i)}.$$

Here $p(\mathbf{y}|X, \mathbf{w}, \mathcal{H}_i)$ is the likelihood term, where $p(\mathbf{w}|\theta, \mathcal{H}_i)$ is the prior. The evidence term, which is also termed as *marginal likelihood*, is,

$$p(\mathbf{y}|X, \theta, \mathcal{H}_i) = \int p(\mathbf{y}|X, \mathbf{w}, \mathcal{H}_i)p(\mathbf{w}|\theta, \mathcal{H}_i)d\mathbf{w}.$$

At the next stage of the hierarchical formulation we have,

$$p(\theta|\mathbf{y}, X, \mathcal{H}_i) = \frac{p(\mathbf{y}|X, \theta, \mathcal{H}_i)p(\theta|\mathcal{H}_i)}{p(\mathbf{y}|X, \mathcal{H}_i)},$$

where,

$$p(\mathbf{y}|X, \mathcal{H}_i) = \int p(\mathbf{y}|X, \theta, \mathcal{H}_i)p(\theta|\mathcal{H}_i)d\theta.$$

At uppermost level we have

$$p(\mathcal{H}_i|\mathbf{y}, X) = \frac{p(\mathbf{y}|X, \mathcal{H}_i)p(\mathcal{H}_i)}{p(\mathbf{y}|X)},$$

and the marginal likelihood is

$$p(\mathbf{y}|X) = \sum_i p(\mathbf{y}|X, \mathcal{H}_i)p(\mathcal{H}_i).$$

This approach is employed together with k-fold cross validation scheme. This means the dataset is split into k equal-sized disjoint sets. Validation is carried out on one of these sets, while training is performed using the union of the remaining $k - 1$ sets. This procedure is run k times, where at each run a different set is considered as validation set.

Through the Bayesian model selection procedure, we determine the free parameters of the Gaussian processes with covariance functions as defined below.

The correlation between random variables at two different observations is formulated using the correlation functions listed in Table 9.1. These are employed

Table 9.1: Covariance functions.

Covariance function	Expression
Independent	σ_0^2
Linear	$\sum\limits_{d=1}^{D} \sigma_d^2 x_d {x_d}'$
Rational quadratic	$\left(1 + \frac{r^2}{2\alpha l^2}\right)^{-\alpha}$
Neural network	$\frac{2}{\pi}\sin^{-1}\left(\frac{2\tilde{x}^T \sum \tilde{x}'}{\sqrt{(1+2\tilde{x}^T \sum \tilde{x})(1+2\tilde{x}^T \sum \tilde{x}')}}\right)$

as indicators for dependencies in addition to the formation of the basis of rules for interpolating observed values. In what follows, we list basic properties of these covariance functions. Independent covariance function is used to express the white noise term in the observations. It is assumed to come from a normal distribution of $\mathcal{N}(0, \sigma_0^2)$ and is of the form,

$$k_c\left(x, x'\right) = \sigma_0^2.$$

For estimating depth, we consider a homogeneous linear kernel with the following form,

$$k_L\left(x, x'\right) = \sum_{i=1}^{n} \sigma_i^2 x_i {x_i}',$$

where n is the number of observations. Consider a network with one hidden layer and N units. It takes an input x and linearly combines the outputs of the units with a bias b according to

$$f(x) = b + \sum_{j=1}^{N} v_j h(x; u_j),$$

where v_j's are the hidden-to-output weights, u_j's are the input-to-hidden weights, and $h(x; u_j)$ is the hidden unit transfer function.

Let b and v_j's have zero mean and unit variance σ_b^2 and σ_v^2, respectively, and weights u_j come from an *iid* distribution. Denoting the weights by w,

$$\begin{aligned} E_w[f(x)] &= 0 \\ E_w[f(x)f(x')] &= \sigma_b^2 + \sum_j \sigma_v^2[h(x;u_j)h(x';u_j)] \\ &= \sigma_b^2 + N_H\sigma_v^2 E_u[h(x;u_j)h(x';u_j)]. \end{aligned}$$

Choosing the transfer function as the error function,

$$h(z) = erf(z) = \frac{2}{\pi}\sin\int_0^z e^{-t^2}dt,$$

we obtain the covariance function as,

$$k_{NN}(x,x') = \frac{2}{\pi}\sin^{-1}\left(\frac{2X^T\sum X'}{\sqrt{(1+2X^T\sum X)(1+2X^T\sum X')}}\right),$$

where $X = [1\ x_1\ \ldots\ x_n]^T$is the augmented input vector. If the covariance function is a function of $r = (x - \tilde{x})$, where x and $\tilde{x}$ is an input pair, it called an isotropic function. Rational quadratic covariance function is a commonly-used isotropic covariance function with the form

$$k_{RQ}(r) = \left(1 + \frac{r^2}{2\alpha l^2}\right)^{-\alpha},$$

As α l are positive, this can be seen an infinite scaled sum squared exponential (SE) covariance functions, where SE covariance function has the following form,

$$k_{SE}(r) = \exp\left(\frac{r^2}{2l^2}\right).$$

Parameterizing in terms of $\tau = l^{-2}$, $\beta^{-1} = l^2$, we have

$$k_{RQ}(r) = \int p\,(\tau|\alpha,\beta)k_{SE}(r|\tau)d\tau.$$

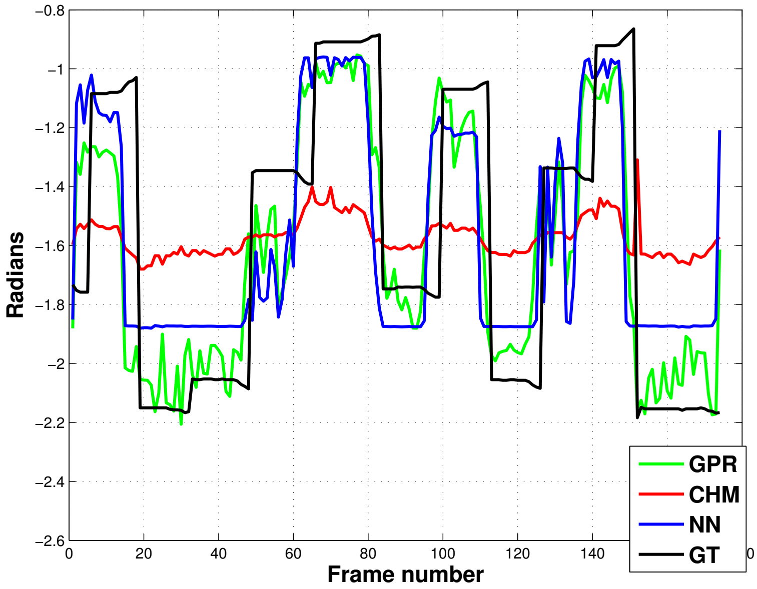
9.2 Target object location estimation

In this section we present some initial results obtained by GP regression, in comparison to an artificial neural network (NN) regressor [16].

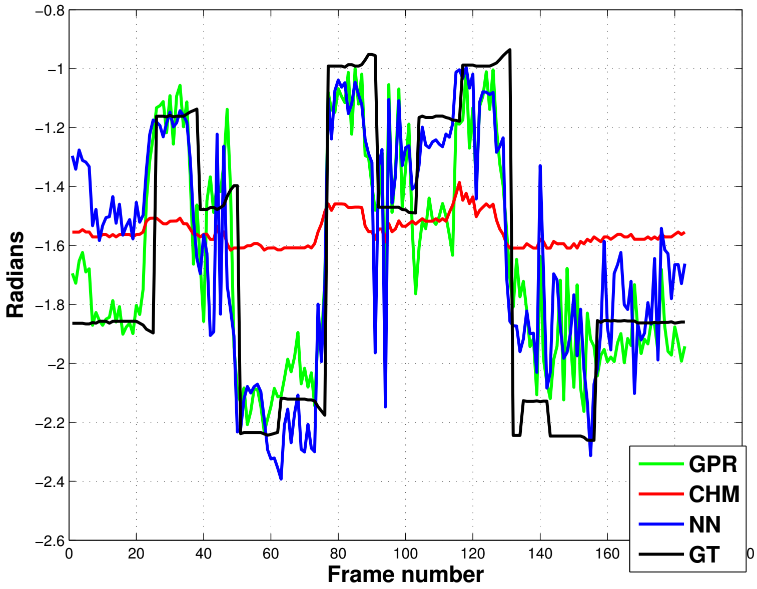
We first examine the details of the NN model. The regression of gaze direction and object depth is carried out using a neural network of one hidden layer. The input layer of the feed-forward artificial neural network receives the three-dimensional estimated pose vector and maps this input to gaze direction, represented by a single angle on the image plane. Therefore, we have an input layer with 3 units and an output layer of one unit. Moreover there are 10 hidden units. The learning rate is initialized to 0.1 and it decreases exponentially during online training. Weights are initialized randomly from the (-0.5,0.5) interval. A validation set is monitored for error decrease to prevent overfitting. The training samples required for the supervised training of the neural network are obtained by manual annotation of the target object location for each frame of the video.

We illustrate the improvement introduced by GP regression by presenting the performance of GP and NN regressors for estimating each of the two identifiers of attention fixation, i.e., gaze direction and target depth. Figures 9.1 and 9.2 illustrate the improvement introduced by GP regression in gaze direction and target depth estimation, respectively, with respect to neural network regression. In Figures 9.1-(a) and (b), it is observed that CHM underestimates the actual gaze direction, while GP and NN regressions follow the ground truth with better accuracy. Moreover GP regressor performs better than the NN regressor in gaze direction estimation. In target depth estimation we observe a similar performance pattern. From Figures 9.2-(a) and (b), it i clear that NN regressor has the possibility of getting stuck at local minimas for some cases.

By imposing target depth estimation on gaze direction estimation, we obtain attention directions and initial estimates for attended object as seen in Figure 9.3. The vector is illustrated such that it starts from the head center, goes along the gaze direction and ends at the point, where it reaches the estimated depth of the object of interest. The end point of the vector is considered to be a reasonable initial estimate for the attended object center. As we carry out this process for



(a)

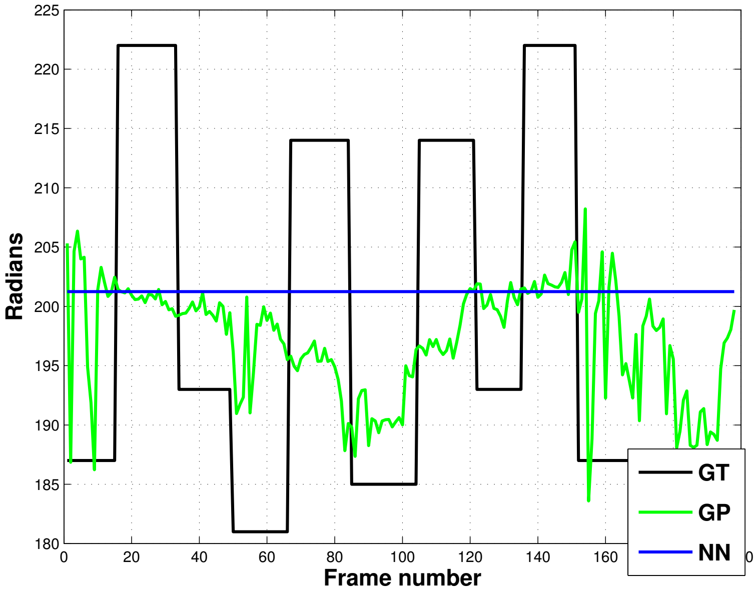


(b)

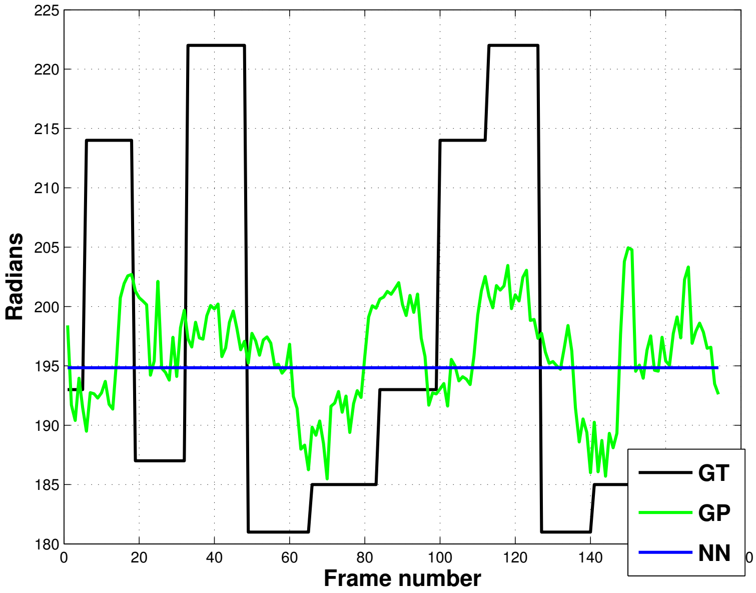
Figure 9.1: Improvement in gaze direction estimation introduced by GP regression, ground truth (GT), gaze estimation from neural network regressor and gaze estimation from cylindrical head model for two video sequences from two different experimenters.

each frame of an exemplary video sequence, we get the initial estimates of focus of attention as presented in Figure 9.4.

There is of course a possibility to improve the results indicated in Figure 9.4. For that purpose, we propose to position a Gaussian distribution around the



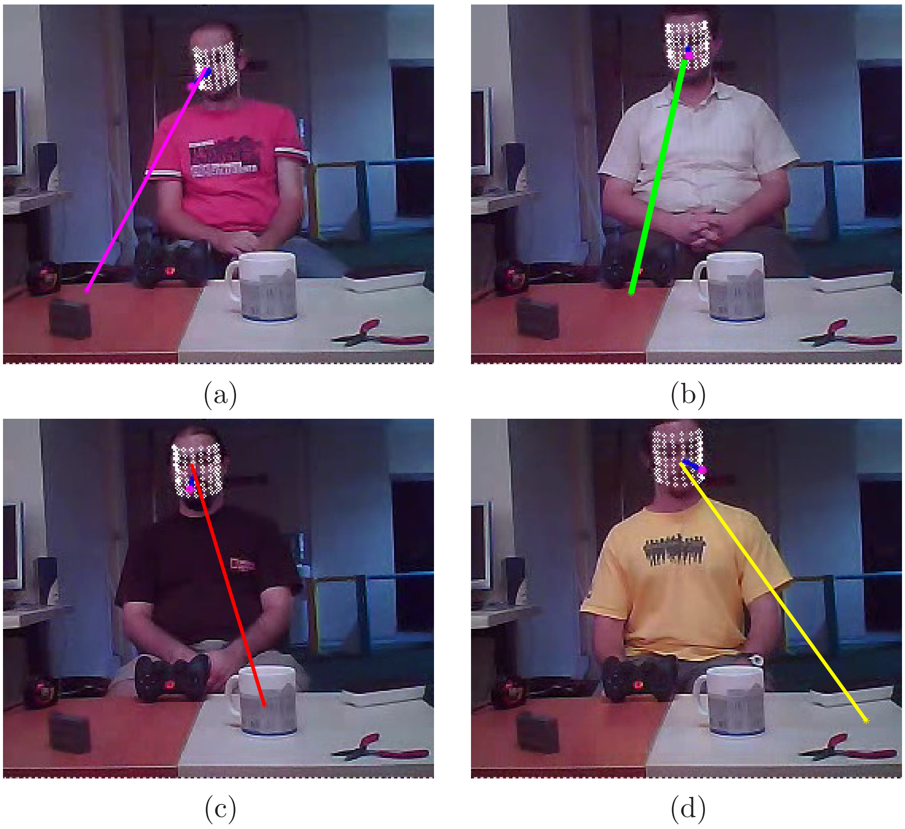
(a)



(b)

Figure 9.2: Improvement in target depth estimation introduced by GP regression, ground truth (GT), depth estimation from neural network regressor for two video sequences from two different experimenters.

initial estimates and compute saliency within this window. We show that the most salient in this window will yield a point that is closer to the actual object center in cases of a deviated initial estimate.



(a)

(b)

(c)

(d)

Figure 9.3: Estimated gaze direction and target depth via GP regression.

It does not however make sense to do this refinement at each frame, since human eye does not focus on different points as frequent as the frame rate of the video (15*fps*). We can turn this fact into a further enhancement by pooling the initial estimates and computing saliency within the accumulated Gaussian windows.

For determining the number of frames to be pooled, we make use of the natural average rate of saccades. Since human eye makes three to five saccades per second, we form bins of consecutive frames by considering 3 consecutive frames to belong to the same bin. This corresponds to accumulating the gaze information for 0.2 seconds, which is a reasonable suggestion in the light of the natural rate of saccades.

The next parameter to determine concerning the Gaussian distributions is their extent, i.e., their standard deviation. A fixed value is picked appropriately

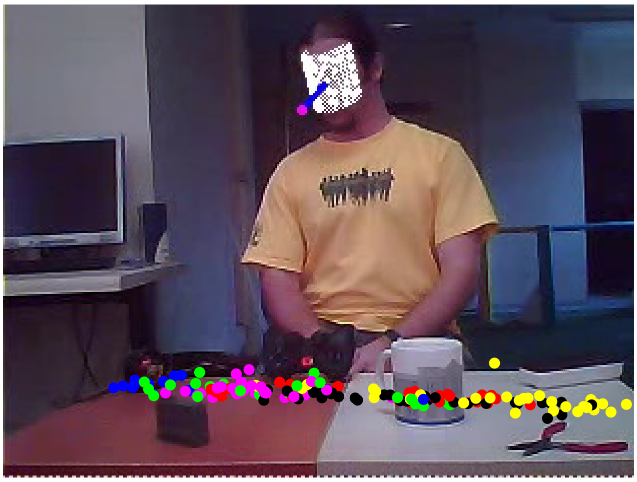
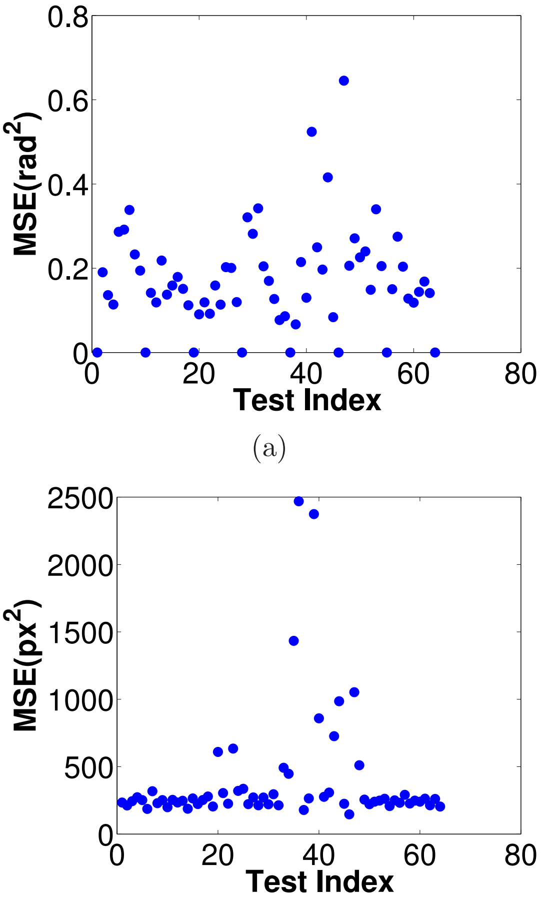


Figure 9.4: Initial estimates for object centers for an exemplary video sequence.

considering the sizes of the objects and the accuracy of estimation. Figures 9.5-(a) and (b) illustrate the mean square error (MSE) for initial estimates regarding gaze direction and target depth estimation, respectively. The mean of MSE concerning gaze direction for the whole set of experiments is calculated to be 29.85 pixels, where the mean of MSE for target depth estimation is 18.47 pixels. However, considering the sizes of the objects, we choose the standard deviation of the Gaussian distributions as 25 pixels. Some resulting prospective target regions illustrated in Figure 9.6 involve particular regions around the estimated gaze, which retain image information and the rest of the visual field is suppressed.

9.3 Saliency model

The masked images, which are obtained by imposing a set of Gaussian distributions on the video images, are used in saliency computation. Using saliency to fixate on the interesting objects serves a two-fold purpose. First, it reduces the uncertainty in the estimation of the gaze direction. In case of slight deviations from the actual object center, saliency computation helps to correct the



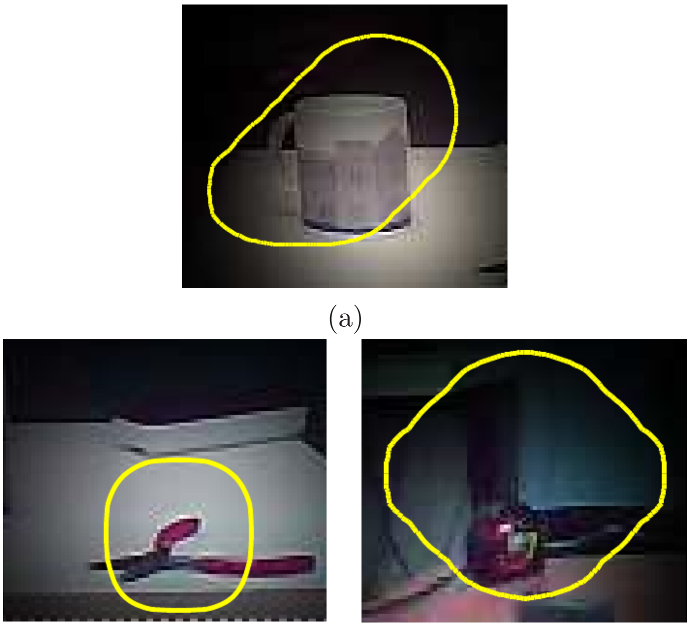
(a)

(b)

Figure 9.5: Mean square error for (a) gaze and (b) target depth estimation.

estimation. Second, saliency-based grafting compensates the discrepancy between intended motor commands and executed physical actions, an issue that is particularly relevant for robotic implementations. The movement of the simulated fovea effectively creates an object-centered coordinate system, which is a precondition of parsimonious mental object representations.

For computing saliency in the prospective region, we employ the popular bottom-up scheme proposed by [45]. Since the agent attempts to determine the final estimation on the masked images, it is forced to attend to the salient parts within this prospective region. The bottom-up scheme is based on the feature integration theory of Treisman and Gelade [89]. Namely, it decomposes



(a)

(b) (c)

Figure 9.6: Saliency computation and segmentation in the prospective region.

the saliency of a scene into separate feature channels, where the presence of illumination intensity, colors, oriented features and motion are indicative of salient locations in the scene. Each feature channel is separately used to determine a feature-specific saliency map, which are then combined to a saliency master map.

The saccadic eye movements are simulated by directing a foveal window to the most salient location, determined by a dynamic and competitive Winner-Take-All (WTA) network [45]. Once a location is selected, it is suppressed by an inhibition-of-return mechanism to allow the next most-salient location to receive attention in the next saccade.

If there is more information available as to the experimenters intentions, or an instruction history that can provide background probabilities with regards to which objects are more likely to receive attention, these can be integrated into the saliency computation in a top-down manner, by for instance modulating the responses of individual feature channels appropriately. In [38], the probability that an experimenter selects a particular object is learned by fitting a Gaussian mixture model on the pixel distribution. We do not model the top-down influence at this stage, simply because in the absence of specific contextual models, this additional information presented to the system would optimistically bias the results.

Chapter 10

Experimental Results

This chapter presents the descriptions of the performance quantifiers employed for our purposes, the details about the training and test schemes, and comparative evaluation of performance rates of the proposed and modified methods. The performance quantifiers are described in Section 10.1 and the features of the training and testing processes are detailed in Section 10.2. Evaluation of performance for the proposed scheme is provided in Section 10.3 in comparison to the modified NN based method.

10.1 Quality measures

Quantification of performance is obtained in terms of two measures, Q_1 and Q_2. While Q_1 indicates at which rate an estimated center falls into the bounding box of the target object of interest, Q_2 shows the rate at which the estimated point is at shortest distance to the true center.

Let $\mathbf{u}_e$ denote the pixel locations of the estimated object centers for a set of frames, which are labeled with object number i. Let B_i be the bounding box of this object on image plane. Then, we have

$$Q_1(i) = \mathcal{C}(\mathbf{u}_e \in B_i)/\mathcal{C}(\mathbf{u}_e),$$

where $\mathcal{C}(.)$ denotes the cardinality of a set.

Since Q_2 assigns an estimated point to the object, whose center lies in shortest distance, it follows that,

$$Q_2(i) = \mathcal{C}(\{\mathbf{u}_e | d(\mathbf{u}_e, \mathbf{c}_i) < d(\mathbf{u}_e, \mathbf{c}_j), \forall j = 1, \cdots, 6, j \neq i\})/\mathcal{C}(\mathbf{u}_e),$$

where $d(\mathbf{a}, \mathbf{b})$ denotes the Euclidean distance between points $\mathbf{a}$ and $\mathbf{b}$ in 2D, and $\mathbf{c}_i$ stands for the object center concerning object i.

In evaluation of Q_1, some estimations may not be assigned to any object, in case they do not fall into any bounding boxes. However, in evaluation of Q_2, the sum of the rows of the confusion tables always add up 1, since each estimation is assigned to an object.

10.2 Training and test schemes

Performance rates are investigated for several configurations of the data. We investigate the behavior of Q_1 and Q_2 for both initial and final object location estimates. Thus, the effects introduced by pooling and saliency computation are pointed out. In order to evaluate the contribution of GP regression, we implement the proposed method replacing GP regression with NN regression and present performance rates for the modified implementation as well. For the nonlinear competing NN model, we prefer a two-layer back-propagation neural network to interpolate the gaze direction and object depth from the head pose estimates without any restrictive assumptions with regards to the context of the application as explained in Section 9.2.

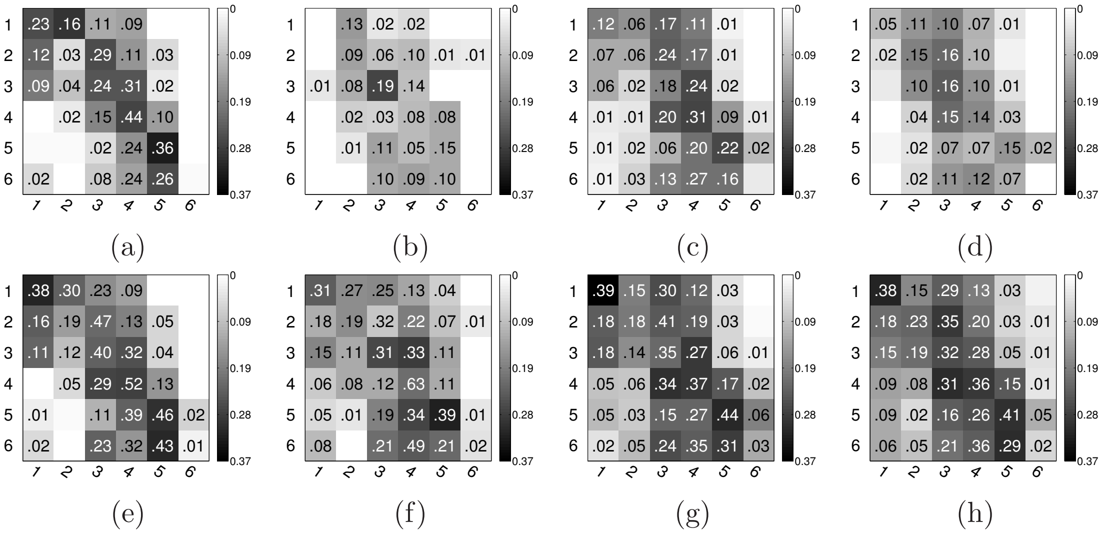
Performance results for self-, and cross-referencing are reported in order to account for the generalization capabilities. In self-referencing, we train the GP regression model with the information obtained from one video of one of the experimenters and run the test on the other video of the same experimenter. In cross-referencing, the process is trained by one video of one of the experimenters, but test is carried out on the videos of other experimenters. The comparison of self- and cross-referencing results help us to understand up to which extent people present personal characteristics in terms of attention direction and how generalizable the proposed scheme is.

10.3 Evaluation of performance

In presentation of the performance rates, the confusion table method described in Chapter 5 is used together with one modification. As in Chapter 5, the numbers in the cells indicate the rate at which the true object is classified to be in the assigned class. The intensity of the colors of the cells refers to the standard deviation of the performance rates evaluated for different combinations of the test and training sets. The scale of the color range is shown next to each table (see Tables 10.1(a-h)).

We present performance rates for the proposed GP regression based method in Tables 10.1 (a-h), competing NN regression based method in Tables 10.2 (a-h) and the relative rates of true positives of those in Tables 10.3 (a-h). Each of these tables have a certain organization. Namely, the top row, i.e., Tables (a-d), gives performance results in terms of Q_1, where the bottom row, i.e., Tables (e-h), gives results for Q_2. Tables (a-b) and (e-f) are calculated considering self-referencing, whereas Tables (c-d) and (g-h) consider cross-referencing. For each of these sets, the first table presents results for initial estimates and the second one for final estimates. As we consider each object individually, the tables are organized in a 6 × 6 fashion.

Table 10.1: GPR Performance quantification for individual objects. Tables (a-d) indicate evaluation of Q_1, (e-h) are for Q_2. Initial and final estimate means for self-referencing are given at (a-b). Initial and final estimate means for cross-referencing are given at (c-d). Tables (e-h) are presented similarly for Q_2.


0
0.09
0.19
0.28
0.37

	1	2	3	4	5	6
1	.23	.16	.11	.09		
2	.12	.03	.29	.11	.03	
3	.09	.04	.24	.31	.02	
4		.02	.15	.44	.10	
5			.02	.24	.36	
6	.02		.08	.24	.26	

(a)

	1	2	3	4	5	6
1		.13	.02	.02		
2		.09	.06	.10	.01	.01
3	.01	.08	.19	.14		
4		.02	.03	.08	.08	
5		.01	.11	.05	.15	
6			.10	.09	.10	

(b)

	1	2	3	4	5	6
1	.12	.06	.17	.11	.01	
2	.07	.06	.24	.17	.01	
3	.06	.02	.18	.24	.02	
4	.01	.01	.20	.31	.09	.01
5	.01	.02	.06	.20	.22	.02
6	.01	.03	.13	.27	.16	

(c)

	1	2	3	4	5	6
1	.05	.11	.10	.07	.01	
2	.02	.15	.16	.10		
3		.10	.16	.10	.01	
4		.04	.15	.14	.03	
5		.02	.07	.07	.15	.02
6		.02	.11	.12	.07	

(d)

	1	2	3	4	5	6
1	.38	.30	.23	.09		
2	.16	.19	.47	.13	.05	
3	.11	.12	.40	.32	.04	
4		.05	.29	.52	.13	
5	.01		.11	.39	.46	.02
6	.02		.23	.32	.43	.01

(e)

	1	2	3	4	5	6
1	.31	.27	.25	.13	.04	
2	.18	.19	.32	.22	.07	.01
3	.15	.11	.31	.33	.11	
4	.06	.08	.12	.63	.11	
5	.05	.01	.19	.34	.39	.01
6	.08		.21	.49	.21	.02

(f)

	1	2	3	4	5	6
1	.39	.15	.30	.12	.03	
2	.18	.18	.41	.19	.03	
3	.18	.14	.35	.27	.06	.01
4	.05	.06	.34	.37	.17	.02
5	.05	.03	.15	.27	.44	.06
6	.02	.05	.24	.35	.31	.03

(g)

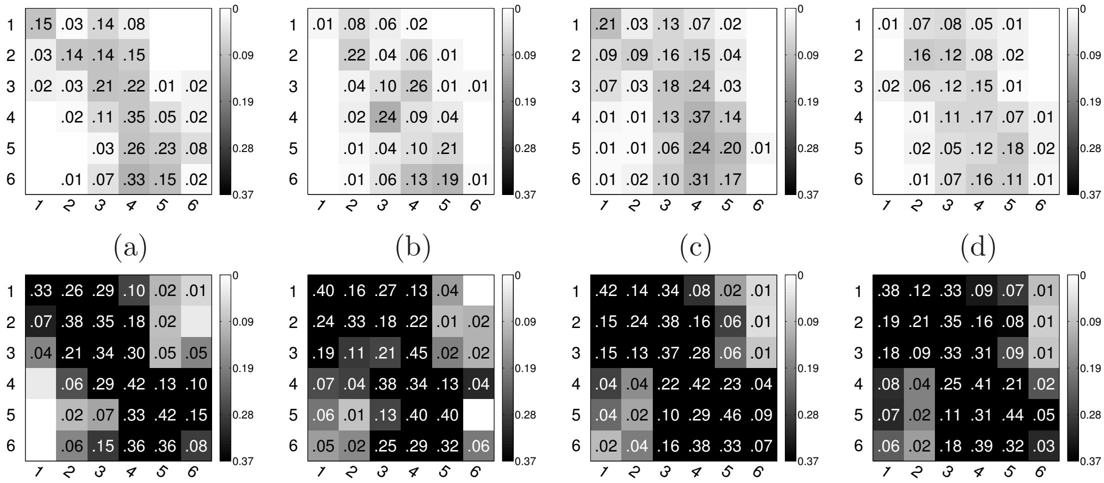
	1	2	3	4	5	6
1	.38	.15	.29	.13	.03	
2	.18	.23	.35	.20	.03	.01
3	.15	.19	.32	.28	.05	.01
4	.09	.08	.31	.36	.15	.01
5	.09	.02	.16	.26	.41	.05
6	.06	.05	.21	.36	.29	.02

(h)

Let us focus on the Tables 10.1(a-h) and discuss the results obtained by the proposed GP based method. Tables 10.1(a-h), indicate that performance rates obtained by using Q_2 are usually higher than those obtained using Q_1. However, the standard deviation of the results obtained employing Q_2 are also higher than the deviation of the results obtained employing Q_1. Moreover, the final estimations are slightly worse than initial estimates for both self- and cross-referencing quantified by Q_1 and Q_2. Regarding the generalization capabilities, we observe that for initial estimates switching from self-referencing to cross-referencing degrades the results slightly, whereas for final estimates it leads to an improvement.

Focusing on Tables 10.2(a-h), we see that most of the remarks we made for Tables 10.1(a-h), hold for the competing NN based method as well. Namely,

Table 10.2: NN Performance quantification for individual objects. Table (a-d) indicate evaluation of Q_1, (e-h) are for Q_2. Initial and final estimate means for self-referencing are given at (a-b). Initial and final estimate means for cross-referencing are given at (c-d). Tables (e-h) are presented similarly for Q_2.


0
0.09
0.19
0.28
0.37

	1	2	3	4	5	6
1	.15	.03	.14	.08		
2	.03	.14	.14	.15		
3	.02	.03	.21	.22	.01	.02
4		.02	.11	.35	.05	.02
5			.03	.26	.23	.08
6		.01	.07	.33	.15	.02

(a)

	1	2	3	4	5	6
1	.01	.08	.06	.02		
2		.22	.04	.06	.01	
3		.04	.10	.26	.01	.01
4		.02	.24	.09	.04	
5		.01	.04	.10	.21	
6		.01	.06	.13	.19	.01

(b)

	1	2	3	4	5	6
1	.21	.03	.13	.07	.02	
2	.09	.09	.16	.15	.04	
3	.07	.03	.18	.24	.03	
4	.01	.01	.13	.37	.14	
5	.01	.01	.06	.24	.20	.01
6	.01	.02	.10	.31	.17	

(c)

	1	2	3	4	5	6
1	.01	.07	.08	.05	.01	
2		.16	.12	.08	.02	
3	.02	.06	.12	.15	.01	
4		.01	.11	.17	.07	.01
5		.02	.05	.12	.18	.02
6		.01	.07	.16	.11	.01

(d)

	1	2	3	4	5	6
1	.33	.26	.29	.10	.02	.01
2	.07	.38	.35	.18	.02	
3	.04	.21	.34	.30	.05	.05
4		.06	.29	.42	.13	.10
5		.02	.07	.33	.42	.15
6		.06	.15	.36	.36	.08

(e)

	1	2	3	4	5	6
1	.40	.16	.27	.13	.04	
2	.24	.33	.18	.22	.01	.02
3	.19	.11	.21	.45	.02	.02
4	.07	.04	.38	.34	.13	.04
5	.06	.01	.13	.40	.40	
6	.05	.02	.25	.29	.32	.06

(f)

	1	2	3	4	5	6
1	.42	.14	.34	.08	.02	.01
2	.15	.24	.38	.16	.06	.01
3	.15	.13	.37	.28	.06	.01
4	.04	.04	.22	.42	.23	.04
5	.04	.02	.10	.29	.46	.09
6	.02	.04	.16	.38	.33	.07

(g)

	1	2	3	4	5	6
1	.38	.12	.33	.09	.07	.01
2	.19	.21	.35	.16	.08	.01
3	.18	.09	.33	.31	.09	.01
4	.08	.04	.25	.41	.21	.02
5	.07	.02	.11	.31	.44	.05
6	.06	.02	.18	.39	.32	.03

(h)

Q_2 values are higher than Q_1 values, final estimation results are not as good as the initial estimates, and cross-referencing results are slightly better than self-referencing for final estimates. However, for standard deviations we observe a similar pattern with more extreme behavior. Observing Tables 10.1 and 10.2, one immediately notices that the proposed GP based method leads to a larger standard deviation in comparison to the competing NN based method, when performance is evaluated in terms of Q_1. However, evaluations in terms of Q_2 indicates that GP based method is more consistent compared to NN based method. This means, compared to Tables 10.1, deviations in Q_1 results are minute, while deviations in Q_2 results are more prominent.

Tables 10.3(a-h) provide a compact presentation of comparison of Tables 10.1 and 10.2. In Tables 10.3, we focus on true positives and present the relative rates in terms of percentages. The results indicate that we usually observe a slight decrease in performance with the proposed GP based method in comparison to the NN based method. But one should also take the actual values into account.

Table 10.3: Relative performance quantification for individual objects. Tables (a-d) indicate evaluation of Q_1, (e-h) are for Q_2. Initial and final estimate means for self-referencing are given at (a-b). Initial and final estimate means for cross-referencing are given at (c-d). Tables (e-h) are presented similarly for Q_2.

	1	2	3	4	5	6
1	.53					
2		−.76				
3			.16			
4				.26		
5					.60	
6						−.76

(a)

	1	2	3	4	5	6
1	−1.0					
2		−.59				
3			.91			
4				−.10		
5					−.27	
6						−1.0

(b)

	1	2	3	4	5	6
1	−.45					
2		−.40				
3			.01			
4				−.18		
5					.06	
6						1.45

(c)

	1	2	3	4	5	6
1	3.7					
2		−.06				
3			.35			
4				−.16		
5					−.15	
6						−1.0

(d)

	1	2	3	4	5	6
1	.17					
2		−.51				
3			.17			
4				.22		
5					.08	
6						−.91

(e)

	1	2	3	4	5	6
1	−.23					
2		−.42				
3			.44			
4				.86		
5					−.01	
6						−.76

(f)

	1	2	3	4	5	6
1	−.06					
2		−.22				
3			−.06			
4				−.14		
5					−.05	
6						−.57

(g)

	1	2	3	4	5	6
1	−.01					
2		.06				
3			−.04			
4				−.11		
5					−.06	
6						−.36

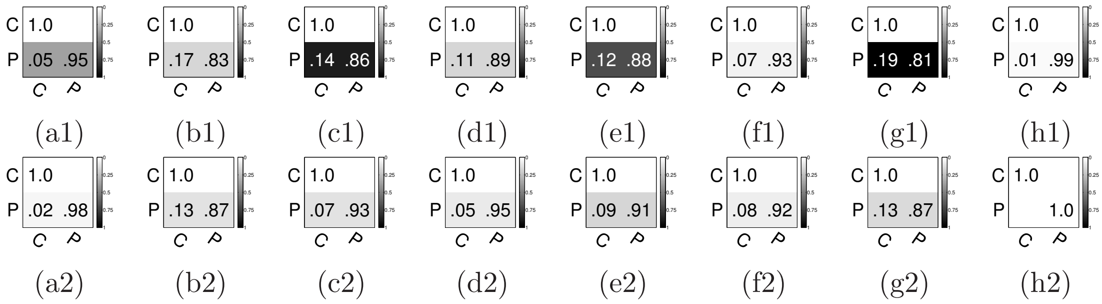
(h)

For example, a decrease from 0.02 to 0.03 leads a relative rate of -0.36 in the last entry of Table 10.3-(h).

In addition to this individual approach, we cluster the objects as *Central* (C) and *Peripheral* (P), depending on their localization on the table. The ones in the middle are considered to be in cluster C, whereas the ones lying on the sides are considered to be in cluster P. For assignment of the objects to either of these clusters, two different schemes are employed. In clustering scheme 1, the separation is obtained by considering the light dashed lines in Figure 7.3 as separating borders, whereas clustering scheme 2 uses heavy dashed lines of Figure 7.3 as decision ground. The confusion tables for the cluster based approach have a 2×2 structure and are given in Tables 10.4.

Tables 10.4(a1-h2) indicate that C is classified correctly at each configuration with 100% accuracy. The peripheral objects are classified as C in some cases but

Table 10.4: Performance quantification for object clusters center (C) and peripheral (P) with (a1-h1) proposed GP based method and (a2-h2) competing NN based method. Figures (a1-d1) are for clustering scheme 1, (e1-h1) are for clustering scheme 2. Initial and final estimate means for self-referencing are given at (a1-b1). Initial and final estimate means for cross-referencing are given at (c1-d1). Figures (e1-h1) are presented in a similar way to (a1-d1). Same operations are carried out for NN regression and the results are reported in a similar way in Tables (a2-h2).



(a1) (b1) (c1) (d1) (e1) (f1) (g1) (h1)

(a2) (b2) (c2) (d2) (e2) (f2) (g2) (h2)

the false positives occur usually at minor rates. Moreover, the proposed GP based method has in general slightly worse performance rate compared to the competing NN based method.

Generally speaking, there are several factors leading to degradation in performance. First of all, as one would intuitively see it is harder to find small objects in terms of Q_1, since they are defined by smaller bounding boxes. On the other hand location of the object on the table affects the performance rate as well. As the yaw and pitch angles increase, the head pose is harder to determine since the view is less similar to the template obtained from a frontal view. In that case, it is more probable that the gaze direction and thus the estimated object location deviates from the correct localization. Moreover for the objects lying at extreme locations, it is harder to run interpolation since they do not have neighboring values beyond them.

Chapter 11

Conclusions

This thesis studies recognition and understanding of behavior using automated decision schemes based on the visual inputs. We exploit the advantages of digital visual behavior analysis and propose several methods that are carried out on a wide scope ranging from animal behavior to human behavior.

Part I of the thesis focuses on animal behavior analysis for drug screening purposes. Experiments, which investigate the effects of drugs on mice and rats, constitute an important research area in pharmacology. The observation and discrimination of drug effects by a skillful authority is on the other hand quite time and resource consuming. Although motion tracking-based computer analysis for behavioral responses has been used for years, the previous approaches were not intended to discriminate a particular drug among other psychoactive agents employed. Automation of the analysis of locomotor activity renders drug screening and behavioral phenotyping of experimental animal studies much easier and faster, consequently this will increase the experimental throughput.

In Part I of the thesis, a new algorithm, which determines whether a test subject is drug-naive or drug-treated automatically and classifies these patterns according to the drug effects, in case they are detected to be drug-treated is presented. We give a motion tracking algorithm and process its results to provide the feature vectors for classification phase. It is observed that the proposed feature vectors are capable of representing the distinctions between the drug effects and the classifiers of SVC and LDC, which are fed with those feature vectors, give satisfactory results.

Part II of the thesis carries out a human behavior analysis with specific focus on visual attention and proposes a system that can comprehend human behavior from an attention point of view. We apply this scheme in a human-robot interaction framework and provide a detailed insight into the design and training of naturally interacting robotic agents by first giving an overview of evolution of joint attention in infants from a developmental psychology point of view and then describing the decomposition of progression in cognitive skills from a robotic implementation perspective. Several cognitively-inspired intelligent agent designs are studied and an alternative algorithm, which employs 3D elliptic cylindrical head models to estimate head pose, is described. Our model uses estimation of head pose, correction for gaze direction, and attention based selection for finding objects attended by an experimenter. We point out to a shortcoming in the literature, in which the head pose is used for specifying the focus of attention. We remedy this by employing a Gaussian process regressor that interpolates the gaze direction and target object depth from the head pose estimates. By this means, we provide a first approximation to an otherwise complex cognitive phenomenon. The proposed scheme has been shown to work with a considerably high performance rate.

Possible future directions of research include direct gaze estimation by using a higher-resolution camera to inspect the eyes of the experimenter, as additional top-down influences. Such systems are considered to have a more flexible nature and thus present a suitable environment for a testbed for complex interaction models, social patterns, alternative teaching techniques, analysis developmental disorders, and running social simulations. Yet one should not forget the contribution of context in the interaction. As Kaplan et al. point out in [49], the existence of top-down influences and the considerations imposed by higher-level cognitive functions make achievement joint attention a very difficult problem.

Bibliography

[1] Aldebaran Nao Humanoid Robot. http://www.aldebaran-robotics.com/eng/Nao.php.

[2] S. Aksoy. Course Notes on Pattern Recognition, 2008. http://www.cs.bilkent.edu.tr/~saksoy/teaching.html.

[3] K. H. An and M.J. Chung. 3D head tracking and pose-robust 2D texture map-based face recognition using a simple ellipsoid model. In *Proceedinga of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 307–312, 2008.

[4] P. Andrews, H. Wang, D. Valente, J. Serkhane, P.P. Mitra, S. Saar, O. Tchernichovski, and I. Golani. Multimedia signal processing for behavioral quantification in neuroscience. In *Proceedings of the* 14^{th} *Annual ACM International Conference on Multimedia*, pages 1007–1016, 2006.

[5] M. Asada, K.F. MacDorman, H. Ishiguro, and Y. Kuniyoshi. Cognitive developmental robotics as a new paradigm for the design of humanoid robots. *Robotics and Autonomous Systems*, 37(2-3):185–193, 2001.

[6] S. Asteriadis, N. Nikolaidis, A. Hajdu, and I. Pitas. An eye detection algorithm using pixel to edge information. In *Proceedings of the International Symposium on Control, Communications, and Signal Processing*, 2006.

[7] K.B. Austin and G.M. Rose. A new technology for quantifying behavioral activation in the rodent [using Doppler radar]. In *Proceedings of the* 19^{th}

Annual International Conference of the IEEE Engineering in Medicine and Biology Society, pages 2411 – 2414.

[8] K.B. Austin and G.M. Rose. Automated behavior recognition using continuous-wave Doppler radar and neural networks. In *Proceedings of the* 19th *Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, volume 4, pages 1458–1461, 1997.

[9] S. Ba and J.M. Odobez. From camera head pose to 3D global room head pose using multiple camera views. In *Proceedings of the International Workshop Classification of Events Activities and Relationships*, 2007.

[10] S.O. Ba and J.M. Odobez. A probabilistic framework for joint head tracking and pose estimation. In *Proceedings of the* 17th *International Conference on Pattern Recognition*, volume 4, pages 264–267, 2004.

[11] L. Bai, L. Shen, and Y. Wang. A novel eye location algorithm based on radial symmetry transform. In *Proceedings of the* 18th *International Conference on Pattern Recognition*, pages 511–514, Washington, DC, USA, 2006. IEEE Computer Society.

[12] S. Balakrishnama and A. Ganapathiraju. Linear discriminant analysis-a brief tutorial. *Mississippi State University Institute for Signal and Information Processing*, 1998.

[13] R. Bates, H. Istance, L. Oosthuizen, and P. Majaranta. Survey of defacto standards in eye tracking. In *Proceedings of the COGAIN Conference on Communication by Gaze Interaction*, 2005. IST-2003-511598: `http://www.cogain.org/results/reports/COGAIN-D2.1.pdf`.

[14] R.J. Beninger, T.A. Cooper, and E.J. Mazurski. Automating the measurement of locomotor activity. *Neurobehavioral Toxicology and Teratology*, 7(1):79–85, 1985.

[15] M.T. Bergen, S.A. Soldan, S.S. Reisman, and J.E. Ottenweller. Low cost rodent activity monitoring instrumentation. In *Proceedings of the IEEE 23^{rd} Northeast Bioengineering Conference*, pages 37–38, 1997.

[16] C.M. Bishop. *Neural Networks for Pattern Recognition*. Oxford University Press, 2005.

[17] K. Branson, V. Rabaud, S. Belongie, and U.C. San Diego. Three brown mice: See how they run. In *Proceedings of the Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, pages 78–85, 2003.

[18] L.M. Brown. 3D head tracking using motion adaptive texture-mapping. volume 1, pages 998–1003, Los Alamitos, CA, USA, 2001. IEEE Computer Society.

[19] O. Buresova, J.J. Bolhuis, and J. Bures. Differential effects of cholinergic blockade on performance of rats in the water tank navigation task and in a radial water maze. *Behavioral Neuroscience*, 100:476–482, 1986.

[20] G. Butterworth and N. Jarrett. What minds have in common is space: Spatial mechanisms serving joint visual attention in infancy. *British Journal of Developmental Psychology*, 9(1):55–72, 1991.

[21] P. Campadelli, R. Lanzarotti, G. Lipori, and U.S. di Milano. Precise eye localization through a general-to-specific model definition. In *Proceedings of the British Machine Vision Conference*, volume 1, pages 187–196, 2006.

[22] R.L. Clarke, R.F. Smith, and D.R. Justesen. An infrared device for detecting locomotor activity. *Behavior Research Methods, Instruments & Computers*, 17(5):519–525, 1985.

[23] J.L. Coatrieux. Shape and function from motion in medical imaging: part 2. *IEEE Engineering in Medicine and Biology Magazine*, 25(1):6–21, 2006.

[24] D. Comaniciu, V. Ramesh, and P. Meer. Kernel-based object tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(5):564–577, 2003.

[25] Ethical committee of Faculty of Medicine. Hacettepe university. Documentation, 2008. http://www.etikkurul.hacettepe.edu.tr/.

[26] T. Cootes, G. Edwards, and C. Taylor. Active appearance models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(6):681–685, 2001.

[27] V. Corkum and C. Moore. *Development of joint visual attention in infants*, pages 61–83. Joint Attention: Its Origins and Role in Development. Lawrence Erlbaum Associates Inc., Hillsdale, NJ, 1995.

[28] D. Cristinacce, T. Cootes, and I. Scott. A multi-stage approach to facial feature detection. In *Proceedings of the British Machine Vision Conference*, pages 277–286, 2004.

[29] R.A. Dielenberg, P. Halasz, and T.A. Day. A method for tracking rats in a complex and completely dark environment using computerized video analysis. *Journal of Neuroscience Methods*, 158(2):279–286, 2006.

[30] D. Drai, Y. Benjamini, and I. Golani. Statistical discrimination of natural modes of motion in rat exploratory behavior. *Journal of Neuroscience Methods*, 96(2):119–131, 2000.

[31] R.O. Duda, P.E. Hart, and D.G. Stork. *Pattern Classification*. Wiley-Interscience, 2001.

[32] D.A. Forsyth and J. Ponce. *Computer Vision: A Modern Approach*. Prentice Hall Professional, 2002.

[33] C.C. Gordon, T. Churchill, CE Clauser, B. Bradtmiller, J.T. McConville, I. Tebbetts, and R.A. Walker. Anthropometric Survey of US Army Personnel: Methods and Summary Statistics. *US Army Natick Research Development and Engineering Center Natick Massachusetts Technical Report*, 225:94, 1989.

[34] M. Hamouz, J. Kittler, JK Kamarainen, P. Paalanen, H. Kalviainen, and J. Matas. Feature-based affine-invariant localization of faces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(9):1490–1495, 2005.

[35] D.W. Hansen and Q. Ji. In the eye of the beholder: A survey of models for eyes and gaze. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 99(1), 2009.

[36] M. Hayhoe, M. Land, and A. Shrivastava. Coordination of eye and hand movements in a normal environment. *Investigative Opthalmology & Vision Science*, 40, 1999.

[37] M. A. Hearst. Support vector machines. *IEEE Intelligent Systems*, 13(4):18–28, 1998.

[38] M.W. Hoffman, D.B. Grimes, A.P. Shon, and R.P.N. Rao. A probabilistic model of gaze imitation and shared attention. *Neural Networks*, 19(3):299–310, 2006.

[39] S. Hongeng, R. Nevatia, and F. Bremond. Video-based event recognition: activity representation and probabilistic recognition methods. *Computer Vision and Image Understanding*, 96(2):129–162, 2004.

[40] K. Hosoda, H. Sumioka, A. Morita, and M. Asada. Acquisition of human-robot joint attention through real-time natural interaction. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, volume 3, pages 2867–2872, 2004.

[41] Y. Hu, L. Chen, Y. Zhou, and H. Zhang. Estimating face pose by facial asymmetry and geometry. In *Proceedings of the* 6th *IEEE International Conference on Automatic Face and Gesture Recognition*, pages 651–656, 2004.

[42] K.S. Huang and M.M. Trivedi. Robust real-time detection, tracking, and pose estimation of faces in video streams. In *Proceedings of the* 17th *International Conference on Pattern Recognition*, volume 3, pages 965–968, Los Alamitos, CA, USA, 2004. IEEE Computer Society.

[43] M. Imai, T. Ono, and H. Ishiguro. Physical relation and expression: Joint attention for human-robot interaction. *IEEE Transactions on Industrial Electronics*, 50(4):636–643, 2003.

[44] H. Ishiguro, T. Ono, M. Imai, T. Maeda, T. Kanda, and R. Nakatsu. Robovie: an interactive humanoid robot. *Industrial robot: An international journal*, 28(6):498–503, 2001.

[45] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11):1254–1259, 1998.

[46] Q. Ji and X. Yang. Real-Time Eye, Gaze, and Face Pose Tracking for Monitoring Driver Vigilance. *Real-Time Imaging*, 8(5):357–377, 2002.

[47] T. Kanda, T. Hirano, D. Eaton, and H. Ishiguro. Interactive robots as social partners and peer tutors for children: A field trial. *Human-Computer Interaction*, 19(1):61–84, 2004.

[48] T. Kanda, H. Ishiguro, M. Imai, T. Ono, and K. Mase. A constructive approach for developing interactive humanoid robots. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 1265–1270, 2002.

[49] F. Kaplan and V. Hafner. The challenges of joint attention. In *Proceedings of the* 4th *International Workshop on Epigenetic Robotics*, pages 67–74, 2004.

[50] B. Kroon, S. Boughorbel, and A. Hanjalic. Accurate eye localization in low and standard definition content. In *Proceedings of the* 8th *IEEE International Conference on Automatic Face & Gesture Recognition*, pages 1–6, 2008.

[51] M. La Cascia, S. Sclaroff, and V. Athitsos. Fast, reliable head tracking under varying illumination: An approach based on registration of texture-mapped 3D models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(4):322–336, 2000.

[52] S.R. Langton, H. Honeyman, and E. Tessler. The influence of head contour and nose angle on the perception of eye-gaze direction. *Perception & Psychophysics*, 66(5):752, 2004.

[53] B.D. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *Proceedings of the International Joint Conference on Artificial Intelligence*, volume 3, pages 674–679, 1981.

[54] Y. Matsumoto and A. Zelinsky. An algorithm for real-time stereo vision implementation of headpose and gaze direction measurement. In *Proceedings of the* 4th *IEEE International Conference on Automatic Face and Gesture Recognition*, pages 499–504, 2000.

[55] A.N. Meltzoff and K.M. Moore. Explaining facial imitation: a theoretical model. *Early Development and Parenting*, 6(3-4):179–192, 1997.

[56] R. Moratz and T. Tenbrink. Affordance-based human-robot interaction. *Lecture Notes in Computer Science*, 4760:63, 2008.

[57] L.P. Morency and T. Darrell. Head gesture recognition in intelligent interfaces: the role of context in improving recognition. In *Proceedings of the*

11th *ACM International Conference on Intelligent User Interfaces*, page 38, New York, NY, USA, 2006. ACM.

[58] L.P. Morency, A Rahimi, and T. Darrell. Adaptive view based appearance models. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 1, pages 803–810. IEEE Computer Society, 2003.

[59] A. Morita, Y. Yoshikawa, K. Hosoda, and M. Asada. Joint attention with strangers based on generalization through joint attention with caregivers. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, volume 4, pages 3744–3749, 2004.

[60] E. Murphy-Chutorian and M.M. Trivedi. Head pose estimation in computer vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(4):607–626, 2009.

[61] Y. Nagai. Joint attention development in infant-like robot based on head movement imitation. In *Proceedings of the 3rd International Symposium on Imitation in Animals and Artifacts*, pages 87–96, 2005.

[62] Y. Nagai. The role of motion information in learning human-robot joint attention. In *Proceedings of the IEEE International Conference on Robotics and Automation*, pages 2069–2074, 2005.

[63] Y. Nagai, M. Asada, and K. Hosoda. Developmental learning model for joint attention. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 932–937, 2002.

[64] Y. Nagai, K. Hosoda, and M. Asada. How does an infant acquire the ability of joint attention?: A Constructive Approach. In *Proceedings of the 3rd International Workshop on Epigenetic Robotics*, pages 91–98, 2003.

[65] Y. Nagai, K. Hosoda, and M. Asada. Joint attention emerges through bootstrap learning. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, volume 1, pages 168–173, 2003.

[66] Y. Nagai, K. Hosoda, A. Morita, and M. Asada. A constructive model for the development of joint attention. *Connection Science*, 15(4):211–229, 2003.

[67] Y. Nagai, K. Hosoda, A. Morita, and M. Asada. Emergence of joint attention based on visual attention and self learning. In *Proceedings of the* 2^{nd} *International Symposium on Adaptive Motion of Animals and Machines*, 2003.

[68] R. Newman, Y. Matsumoto, S. Rougeaux, and A. Zelinsky. Real-time stereo tracking for head pose and gaze estimation. In *Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition Gesture Recognition*, pages 122–128, 2000.

[69] L.P. Noldus, A.J. Spink, and R.A. Tegelenbosch. EthoVision: a versatile video tracking system for automation of behavioral experiments. *Behavior Research Methods, Instruments, & Computers*, 33(3):398–414, 2001.

[70] T. Ono, M. Imai, and H. Ishiguro. A model of embodied communications with gestures between humans and robots. In *Proceedings of the* 23^{rd} *Annual Meeting of the Cognitive Science Society*, pages 732–737, 2001.

[71] V. Pasquali and P. Renzi. On the use of microwave radar devices in chronobiology studies: an application with Periplaneta Americana. *Behavior Research Methods*, 37(3):522, 2005.

[72] C.E. Rasmussen and C.K.I. Williams. *Gaussian Processes for Machine Learning.* MIT Press, 2006.

[73] D. Reisfeld, H. Wolfson, and Y. Yeshurun. Context-free attentional operators: the generalized symmetry transform. *International Journal of Computer Vision*, 14(2):119–130, 1995.

[74] E. Robles. A method to analyze the spatial distribution of behavior. *Behavior Research Methods, Instruments & Computers*, 22(6):540–549, 1990.

[75] FESTO Robotino robot platform. `http://www.festo-didactic.com/int-en/learning-systems/education-and-research-robots-robotino/`.

[76] D.B. Russakoff and M. Herman. Head tracking using stereo. *Machine Vision and Applications*, 13:164–173, 2002.

[77] B. Scassellati. *Imitation and Mechanisms of Joint Attention: A Developmental Structure for Building Social Skills on a Humanoid Robot*, pages 176–195. Computation for Metaphors, Analogy, and Agents. Springer Verlag. `http://www.springerlink.com/content/wljp04e4h5b4lthh`.

[78] B. Scassellati. Theory of mind for a humanoid robot. *Autonomous Robots*, 12(1):13–24, 2002.

[79] B. Schölkopf, C.J.C. Burges, and A.J. Smola. *Introduction to Support Vector Learning*. MIT Press, Cambridge, MA, USA, 1999.

[80] Y.H. Shih and M.S. Young. Integrated digital image and accelerometer measurements of rat locomotor and vibratory behaviour. *Journal of Neuroscience Methods*, 166(1):81–88, 2007.

[81] A.P. Shon, D.B. Grimes, C.L. Baker, M.W. Hoffman, S. Zhou, and R.P.N. Rao. Probabilistic gaze imitation and saliency learning in a robotic head. In *Proceedings of the IEEE International Conference on Robotics and Automation*, pages 2865–2870, 2005.

[82] A.P. Shon, J.J. Storz, A.N. Meltzoff, and R.P.N. Rao. A cognitive model of imitative development in humans and machines. *International Journal of Humanoid Robotics*, 4(2):387–406, 2007.

[83] B.M. Spruijt, M.O.S. Buma, P.B.A. van Lochem, and J.B.I. Rousseau. Automatic behavior recognition: What do we want to recognize and how do we measure it. In *Proceedings of the* 2^{nd} *International Conference on Methods and Techniques in Behavioral Research, Measuring Behavior*, volume 98, pages 264–6, 1998.

[84] B.M. Spruijt and W.H. Gispen. Prolonged animal observation by use of digitized video displays. *Pharmacology, Biochemistry, and Behavior*, 19(5):765, 1983.

[85] R. Stiefelhagen, J. Yang, and A. Waibel. Modeling focus of attention for meeting indexing. In *Proceedings of the* 7^{th} *ACM international conference on Multimedia (Part 1)*, pages 3–10, 1999.

[86] H. Sumioka, K. Hosoda, Y. Yoshikawa, and M. Asada. Acquisition of joint attention through natural interaction utilizing motion cues. *Advanced Robotics*, 21(9):983–1000, 2007.

[87] J. Sung, T. Kanade, and D. Kim. Pose robust face tracking by combining active appearance models and cylinder head models. *International Journal of Computer Vision*, 80(2):260–274, 2008.

[88] M.H. Teicher, S.L. Andersen, P. Wallace, D.A. Klein, and J. Hostetter. Development of an affordable hi-resolution activity monitor system for laboratory animals. *Pharmacology, Biochemistry and Behavior*, 54(2):479–484, 1996.

[89] A.M. Treisman and G. Gelade. A feature-integration theory of attention. *Cognitive Psychology*, 12(1):97–136, 1980.

[90] J. Triesch, H. Jasso, and G.O. Deák. Emergence of mirror neurons in a model of gaze following. *Adaptive Behavior*, 15(2):149–165, June 2007.

[91] J. Tu, T. Huang, and H. Tao. Accurate head pose tracking in low resolution video. In *Proceedings of the 7th International Conference on Automatic Face and Gesture Recognition*, pages 573–578, 2006.

[92] M. Türkan, M. Pardás, and A.E. Çetin. Human eye localization using edge projection. In *Proceedings of the International Conference on Computer Vision, Theory and Applications*, 2007.

[93] R. Valenti and T. Gevers. Accurate eye center location and tracking using isophote curvature. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2008.

[94] R. Valenti, Z. Yücel, and T. Gevers. Robustfying Eye Center Localization Using Head Pose Cues. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 612–618.

[95] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1:511, 2001.

[96] P. Viola and M.J. Jones. Robust real-time face detection. *International Journal of Computer Vision*, 57(2):137–154, 2004.

[97] C.V. Vorhees, K.D. Acuff-Smith, D.R. Minck, and R.E. Butcher. A method for measuring locomotor behavior in rodents: contrast-sensitive computer-controlled video tracking activity assessment in rats. *Neurotoxicology and Teratology*, 14(1):43, 1992.

[98] J. Xiao, T. Kanade, and J.F. Cohn. Robust full-motion recovery of head by dynamic templates andre-registration techniques. In *Proceedings of the 5th IEEE International Conference on Automatic Face and Gesture Recognition*, pages 156–162, 2002.

[99] H. Yamazoe, A. Utsumi, T. Yonezawa, and S. Abe. Remote gaze estimation with a single camera based on facial-feature tracking without special calibration actions. In *Proceedings of the Symposium on Eye Tracking Research & Applications*, pages 245–250, 2008.

[100] Z. Yücel and A. A. Salah. Head pose and neural network based gaze direction estimation for joint attention modeling in embodied agents. In *Proceedings of the Annual Meeting of Cognitive Science Society*, pages 3139–3144, 2009.

[101] Z. Yücel and A.A. Salah. Resolution of focus of attention using gaze direction estimation and saliency computation. In *Proceedings of the International Conference on Affective Computing and Intelligent Interfaces*, 2009.

[102] Z. Yücel, A.A. Salah, C. Meriçli, and T. Meriçli. Joint Visual Attention Modeling for Naturally Interacting Robotic Agents. In *Proceedings of the 24th International Symposium on Computer and Information Sciences*, pages 242–247, 2009.

[103] Z. Yücel, Y. Sara, P. Duygulu, R. Onur, E. Esen, and A. B. Özgüler. Automated discrimination of psychotropic drugs in mice via computer vision-based analysis. *Journal of Neuroscience Methods*, 180(2):234 – 242, 2009.

[104] Y. Zhang and C. Kambhamettu. 3D head tracking under partial occlusion. *Pattern Recognition*, 35:1545–1557, 2002.

[105] Z.H. Zhou and X. Geng. Projection functions for eye detection. *Pattern Recognition*, 37(5):1049–1056, 2004.

[106] J. Zlatev and C. Balkenius. Introduction: Why epigenetic robotics. In *Proceedings of the 1st International Workshop on Epigenetic Robotics: Modeling Cognitive Development in Robotic Systems*, volume 85, pages 1–4, 2001.

[107] J.B. Zurn, D. Hohmann, S.I. Dworkin, and Y. Motai. A real-time rodent tracking system for both light and dark cycle behavior analysis. In *Proceedings of the IEEE Workshop on Applications of Computer Vision*, pages 87–92, 2005.

[108] J.B. Zurn, X. Jiang, and Y. Motai. Video-based rodent activity measurement using near-infrared illumination. In *Proceedings of the IEEE Instrumentation and Measurement Technology Conference*, volume 3, pages 1928–1931, 2005.