

TDK-dolgozat

Szalai Szilárd

Substantive evaluation of vision based on human eye movement [\[1\]](#), [\[2\]](#)



Emberi szemmozgás-kísérleteken alapuló modell
a látvány lényegi kiértékelésére

Szilárd Szalai, IB MSc, 1st year

Supervisors:

Tamás Szirányi D.Sc. professor at PPCU,
Zoltán Vidnyánszky D.Sc., professor at PPCU

Faculty of Information Technology,
Pázmány Péter Catholic University

Budapest, 2011

Abstract

Algorithmic estimation for human eye movement is a dominant area of research. Besides technical knowledge, this kind of problem requires neurobiological and psychophysical competences as well. According to the current knowledge, vision is a complex process that is not only based on the features of the images, but is also dependent on cognitive factors as well. However, state-of-the-art methods are mostly bottom-up approaches.

In this work, I address the statistical inference of saliency features in the images based on cognitive factors. As a first step, training videos were recorded by a head-mounted wearable eye-tracker device, where the position of the eye fixation relative to the recorded image was annotated. From the same video records, artificial saliency points were measured by computer vision algorithms which were clustered to describe the images with a manageable amount of descriptors. The measured human eye-tracking (fixation pattern) and the estimated saliency points are fused in a statistical model, where the eye-tracking supports us with transition probabilities among the possible image feature points.

This HVS-based statistical model results in the estimation of possible tracking paths and region of interest areas of the human vision. The proposed method may help in image saliency analysis, better compression of region of interest areas and in the development of more efficient human-computer-interaction devices. This method can successfully estimate the eye movement, moreover, the parameters of the model can be easily modified according to the needs of a given application area.

Összefoglaló

Az emberi szemmozgás algoritmikus úton történő becslése egy fontos kutatási terület. A jelenleg elfogadott modell értelmében a látás egy olyan összetett folyamat, amelyet a látóterünkben levő kép tulajdonságai mellett belső, kognitív tényezők is befolyásolnak. A folyóiratokban bemutatott módszerek legtöbbje azonban tisztán a kép struktúrájából kiindulva közelíti meg a problémát.

Ebben a munkában megkíséreltem túlmutatni ezeken a módszereken kognitív tényezők integrálásával. A modell emberi szemmozgás-kísérletekre épül, amelyek tanítási és tesztelési célt szolgáltak. Első lépésben egy fejre szerelhető szemkövető készülékkel tanuló videókat rögzítettünk, amelyek képkockáin képi leírókat detektáltam. Ezeket klaszterezve kezelhető mennyiségű adatot kaptam. Az így nyert mesterséges kiugró pontokat emberi szemmozgási mintákkal együtt statisztikai modellbe illesztettem. Ez a megközelítés olyan gráfok bejárásán alapul, amelyeket a humán mérések során nyert valószínűségi adatok alapján építettem fel. A modell becslést ad arra, hogy adott képen melyek azok a részek, amelyek vonzzák a tekintetet. Az eredmények kiértékeléséhez újabb kísérleteket végeztem, amelyben alanyok szemmozgását rögzítettem bizonyos feladatok elvégzése közben. A modellt különböző képi leírók kezelésére készítettem fel.

Munkám célja az volt, hogy egy új megközelítés szerint adjak becslést a tekintet útjára. Modellem széles körű alkalmazási lehetőségeket hordozhat magában. Az eredmények kiértékelésénél azt találtam, hogy a módszer sikeresen ad becslést a tekintet útjára, ráadásul könnyen paraméterezhető egy-egy konkrét alkalmazási terület igényei szerint.

Contents

1	Introduction	1
1.1	The structure of the paper	2
2	Theoretical background - Neurobiology of vision	5
2.1	Human visual system	5
2.2	The retina	6
2.3	Lateral Geniculate Nucleus	8
2.4	The primary visual cortex	9
2.5	Overview of higher visual areas	10
2.6	Eye movement patterns	10
2.7	Top-down and bottom-up mechanisms	12
2.8	Related works on visual perception	13
3	Theoretical background - Computational Saliency	15
3.1	Computational saliency	15
3.2	Bottom-up saliency models	16
3.3	Consequences and the goal of my model	18
4	Human measurements	21
4.1	The eye-tracker	21
4.2	Eye-tracker-based training experiments	22

5	Detection of salient regions	25
5.1	Saliency point detection and clustering	25
5.2	Building transition probability tables and constructing graphs . .	26
5.3	Estimation of the most probable points - Regions of Interest . . .	29
6	Verification	31
6.1	Test experiments by a desktop eye-tracker	31
6.2	Creating fixation density map	34
6.3	Detection of the fixation regions on human experiments	34
6.4	Filtering the estimated fixation patterns	36
6.5	Comparison of the estimated points to real measurements	37
7	Applications	43
7.1	Human-Computer Interaction	43
7.2	Medical applications	44
7.3	Marketing purposes	45
7.4	Eye movement analysis and fatigue detection	45
	References	53

Chapter 1

Introduction

Human-Computer Interaction (*HCI*) is a popular field of research since the influence of biology for computational algorithms is to be on the increase. As HCI is concerned with developing descriptive and predictive theories and models, it is worth to substantially evaluate the human vision, which has a unique role, because it is based on some cognitive factors as well. In the development of automatic (*HCI*) systems, finding the focus of visual attention of human observers without any restraint has an important role. These automatic interfaces might be important in different fields of life, for instance in medical applications, traffic control, adaptive advertisement tables or during driving.

Eye tracking systems may support us with a partly unsupervised training method of human visual attention selection. Here the human visual activity can be annotated with fixation points of eye movement, while the scene is recorded in a video in parallel. This annotated record can be used to describe the image content as the human apprehends it for two main research directions:

- to characterize the image content;
- to characterize the viewer.

The first task concerns machine vision, where the visual preferences of human attention are built in the image features in some training procedure. The second task refers to the computational modeling of the perceptual and neural mechanisms of the human vision.

This work addresses the first task: the goal is to discover the artificial saliency features of images with the statistical properties trained from human eye movement recording experiments. The path of the human gaze, the time of fixation on a selected region and the frequency of switches between the different parts of the images can be typical for an image as well.

In my hypothesis, I assume that the transitions between two consecutive fixation points (eye movements, saccades) are also characteristic for the attention selection processes or the image content; the probability that the eye moves to a given point from another may be characteristic of those points and the features behind them. In our proposed model any high number of possible salient regions is possible in the scene, contrary to the model assumptions in Esaliency [3], while the task is not to find one (group) of the object(s), but to find the path as the human gaze may connect them. I also have the goal to find the salient regions, but this is based on the dense sections of the probable tracking paths and not on the saliency of the object itself. In our proposed method, I attempt to find the regions of the images that will be selected by visual attention and thus will be fixated by the human observers.

1.1 The structure of the paper

The two following chapters give some theoretical background to this work. The reason why that part is divided into two parts is that this topic needs multidisciplinary knowledge, from the field of biology and informatics. In the second chapter, the neurobiology of vision is described, and the brain areas are mentioned which are involved in the visual procession. Some notions are defined, for instance the psychophysical saliency, top-down and bottom-up mechanisms, moreover, the results of some studies concerning vision is presented. The topic in Chapter III is the computational saliency. After presenting a recent computational attention model, some bottom-up state-of-the-art methods are described. The consequences are taken from the theoretical background and my goals are described. Chapter IV presents the human measurements and the device which was used for recording the training videos. The steps of the models are detailed

in the fifth chapter. The bottom-up saliency detection in the images, the construction of transition probability graphs, the biology-inspired filtering and the fusion of top-down cognitive fingerprint detection. In Chapter VI, the results are presented and compared to human eye movement patterns. Not only subjective, but also objective comparison is done between my model and state-of-the art models. In the last chapter, some possible application areas are mentioned in which this model can be successfully applied in the near future.

Chapter 2

Theoretical background - Neurobiology of vision

In order to develop methods for inferring human eye movement, it is essential to understand the process of visual perception. This section gives a short neurobiological introduction about human vision and the phase of perception. The information described in these sections is the basis of my model. In these pages, I approach human vision from engineering point of view. The results of some recent studies are presented from the field of computational biology as well.

2.1 Human visual system

The visual system is our most important sense; about one quarter of all nerve cells in our central nervous system (CNS) is related to vision in some way. Therefore it is necessary to perform a substantial analysis of the 2D image as a projection of the 3D world [4]. The front-end visual system measures simultaneously at multiple resolutions, it measures directly (in the scale-space model) derivatives of the image in all directions at least up to fourth order, it measures temporal changes of intensity, the motion and disparity parameters, and the color differential structure.

From engineering aspects, the human visual system is a multi-scale sampling device of the outer world. It exploits this strategy by the creation of so-called receptive fields (RF's) on the retina: groups of receptors assembled in such a way that they form a set of apertures of widely varying size. They together measure

a scale-space of every image. The hierarchical structure of the input image is contained in this multi-scale stack of images measured at a range of scales. This step is called deep structure.

The human visual system does ensemble measurements: for every (perceivable) aspect of the stimulus it has a dedicated set of detector called receptive fields or receptive field pairs. They span the full measurement range of the parameter, i.e. for every location, order of spatial and temporal differentiation of the stimulus, for every orientation, for every velocity in every direction, for every disparity, etc.

The visual system is considered layered: its first stages measure the geometrical structure by multi-scale partial derivatives in space and time, and subsequent layers perform an analysis of the contextual structure, by perceptual grouping and hierarchical topological analysis, the highest stages do the cognitive, highly associative tasks. This rough division in processing layers is also known as front-end, intermediate or high level visual processing [4].

2.2 The retina

The retina is an approximately 0.5 mm thick neural tissue which is adhered to the back of the eyeballs (Fig. 2.1). The role of this complex structure is to perceive the incoming light and the first elementary processing steps are starting here as well.

The cells of the retina are arranged in three layers (Fig. 2.2). The optic nerve contains the ganglion cell axons running to the brain and, additionally, incoming blood vessels that open into the retina to vascularize the retinal layers and neurons. A radial section of a portion of the retina reveals that the ganglion cells (the output neurons of the retina) lie innermost in the retina closest to the lens and front of the eye, and the photo sensors (the rods and cones) lie outermost in the retina against the pigment epithelium and choroid. Light travels through the thickness of the retina before striking and activating the rods and cones. Subsequently the absorption of photons by the visual pigment of the photoreceptors is translated into first a biochemical message and then an electrical message that can stimulate all the succeeding neurons of the retina. The retinal

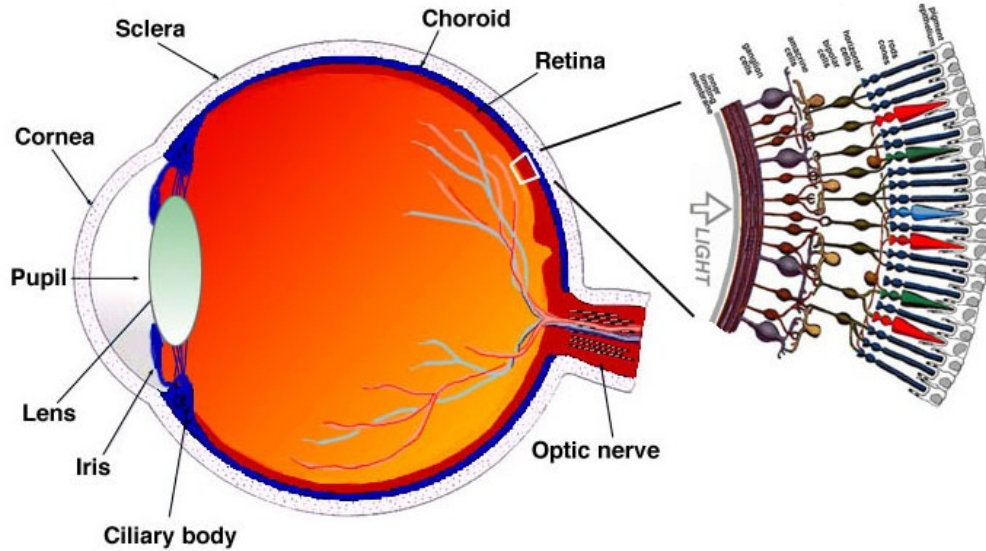


Figure 2.1: Retina

message concerning the photic input and some preliminary organization of the visual image into several forms of sensation are transmitted to the brain from the spiking discharge pattern of the ganglion cells. The middle layer contains horizontal, bipolar and amacrine cells. The front layer contains the about one million ganglion cells, whose axons form together the optic nerve, the output of the retina.

The light sensitive receptors, the rods and the cones, are located in the back of the eye behind the other layers. Rods are only used at very dim light levels. This is rod vision, or scotopic vision. In normal daylight lighting conditions, they are in a completely saturated state, and have no role in perception. Rods are single color, so in the dim light we see no colors. Cones are used at normal light levels, i.e. photopic vision. Cones have three types, for long (red), medium (green) and short (blue) wavelength sensitivity. The rhodopsin molecules in the receptors that are bleached by the light can in this way be easily replenished. The estimated number of rods in the human retina is approximately 120 000 000, while we have about 6 400 000 cones [5].

Each cell in the visual cortex has a receptive field, a discrete area in space

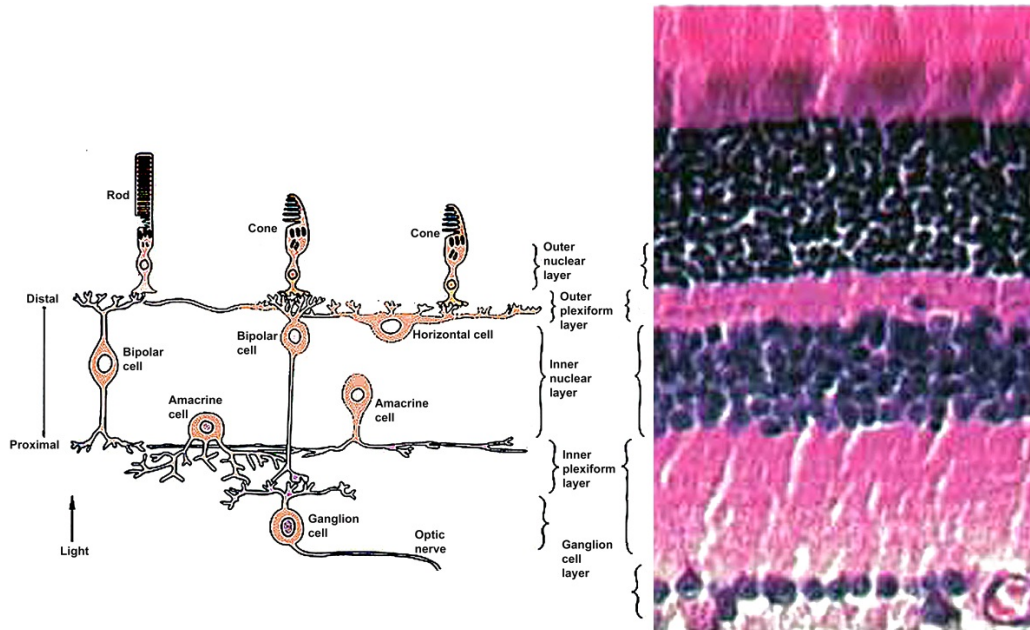


Figure 2.2: The cells of the retina are arranged in layers

relative to the fovea where the presentation or removal of a visual stimulus will cause cellular activation. By definition, stimuli presented outside of this receptive field will neither increase nor decrease the ongoing activity of that individual cell. The location and size of a receptive field can be considered among the most basic of RF properties. The receptive fields have some basic properties: orientation and direction selectivity, binocularity and binocular disparity etc.

2.3 Lateral Geniculate Nucleus

The retina is connected to the brain via the optic nerve which runs into a specific region inside the thalamus, called Lateral Geniculate Nucleus (LGN). In the thalamus the visual perception information meets with other information coming from different modalities (e.g. auditory, tactile). The LGN is a small peanut-shaped nucleus which is located at the dorsal part of the thalamus and consists of six layers. The top four parvocellular layers have small cell bodies, while the bottom two magno-cellular layers have relatively large. The central line, that con-

nects the cells, receives projections from the fovea. Parvo-cellular layers convey information about colors and shapes; motion information is processed through magno-cellular layers.

According to the electrophysiological study of [6], the receptive fields of ganglion cells have a 50% on-center and 50% off-center multi-scale center-surround structure; moreover, LGN cells have a 50% on-center and 50% off-center center-surround structure at a wide range of scales as well. Since the receptive fields are not constant in time, a good model describing this spatio-temporal behavior is the product of a Laplacian of Gaussian for the spatial component, multiplied with a first order derivative of a Gaussian for the temporal domain:

$$\left(\frac{\partial^2 G}{\partial x^2} + \frac{\partial^2 G}{\partial y^2} \right) \frac{\partial G}{\partial t} \quad (2.1)$$

As 50% of the center-surround retinal cells is on-center and 50% is off-center, the foreground and the background are just as important.

The main projection area after the LGN for the primary visual pathway is the primary visual cortex in Brodmann's area 17. Since the topological structure is still kept the map of the visual field is projected to the visual cortex. In this phase, there is a feedback control (geometry-driven diffusion and perceptual grouping) to the early stage of the thalamus. This feedback seems to be essential because it could help to understand local and global image analysis.

2.4 The primary visual cortex

The primary visual cortex is a region in the back of the head, in the area 17 and consists of seven layers which structure is retinotopic and columnar. These cortical columns are repetitive areas and can be considered as visual pixels; moreover, they contain processing filters for local geometrical analysis of the given pixel. The hyper columns are functional units of the cortical structure and can do binocular analysis of the mentioned pixels in the overlapping visual field of both eyes.

The visual cortex contains simple and complex cells as well. The sensitivity of the simple cells can be modeled by Gaussian derivative kernels [7], these receptive fields can be modeled by Gabor filters [8]. Since the structure of the complex cells

is not clear, it is difficult to model them. However, we know that their receptive fields show temporal responses. The cells in the primary visual cortex have a directional sensitivity for motion.

Receptive fields substantially overlap, and they should in order to create a correlation between neighboring fibers. However, they overlap because we have a multi-scale sampling structure. At a single scale, our model presumes a tight hexagonal tiling of the plane. There is a deep notion here of the correlation between different scales, and the sampling at a single location by receptive fields of different scale.

The cortex has to estimate the spatial arrangements of incoming fibers from the mutual relations between the signals they transport. Neighboring geometrical properties have to correspond, such as intensity, contours with the same orientation, etc. [4].

2.5 Overview of higher visual areas

The processing of the visual stimuli continues in higher cortical visual areas through neural pathways. From the primary visual cortex, the information runs to the V2, V3, V4 and the mediotemporal layer. In these areas, shape, motion and color features are processed in a more accurate way. Figure 2.3 contains the higher brain areas having roles in visual perception [4].

2.6 Eye movement patterns

To analyze the properties of the eye movement and to understand its behavior, the introduction of some basic notions is essential. The eye movement consists of different type of patterns having the dependent on the period of movement. There are different ways to distinguish them, for instance according to the speed difference.

The visual input is processed during the *fixation* periods. During fixation periods, the movement of the eye is getting slower, the gaze is maintained on a single location. This is the phase of visual perception; the gaze is fixating to the

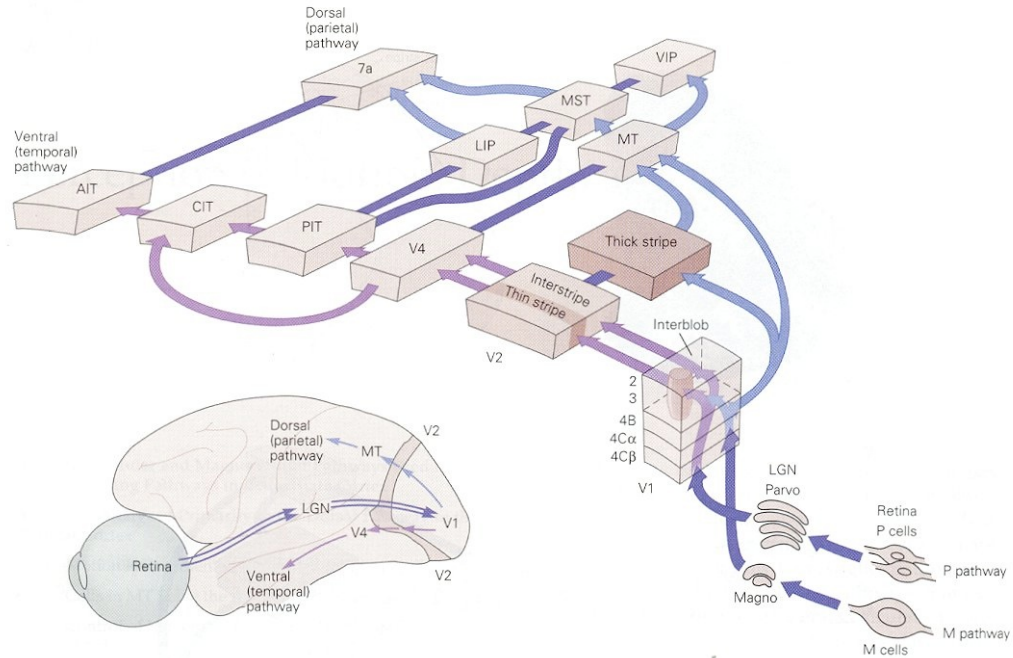


Figure 2.3: Higher brain areas

so-called fixation regions. In the case of fixation, the observer's eye's fixation is on a small-scale region, the eye movement is twitching over this area.

Since humans do not look at a scene in fixed steadiness, there is another type of eye movement. The *saccade* is a fast movement of the eye; it has a short temporal extension since during this time the fixation is moving to another fixation region. Initiated by eye fields in the frontal and parietal lobes of the brain, saccades serve as a mechanism for fixation, rapid eye movement and the fast phase of optokinetic nystagmus. Saccades are quick, simultaneous movements of both eyes, the direction of the eye movement is about the same. There is no visual information processing during the saccades. If the saccades are underway, they cannot be altered by the will until the eye fixates. In an unexpected situation, saccades take about 200 milliseconds.

Glissade is a gliding involuntary movement of the eye in changing the point of fixation. It is a slower, smoother movement than the saccade movement. In this model I neglected the so-called glissade patterns and assumed that eye movement

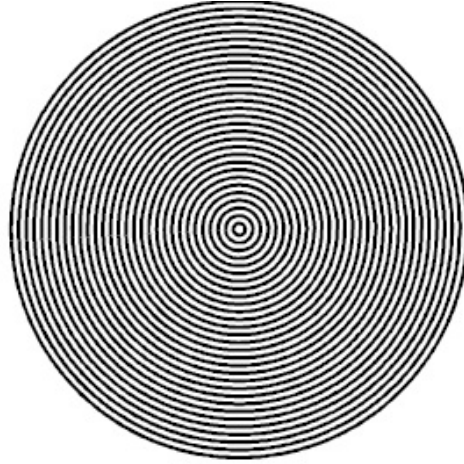


Figure 2.4: Involuntary eye movements (e.g. micro-saccade) can cause a shimmering effect on this stimulus pattern.

consists of only fixation and saccade movement. The explanation lies in the fact, that our eye tracker devices distinguishes only this two type of parts since their sampling rate is not enough the detect glissades.

We continuously make very small eye movements, possibly in order to keep seeing. These small eye movements are substantially larger than single receptors, and are made synchronously by both eyes. There is also small drift. Due to the continuously micro-saccade eye movement, some stimulus patterns (Fig. 2.4) continuously seem to be shimmering which is due to the involuntary eye movements (drift, tremor and micro-saccades).

2.7 Top-down and bottom-up mechanisms

The most important function of selective visual attention is to direct our gaze rapidly towards objects of interest in our visual environment. In [9] it is suggested that subjects selectively direct attention to objects in a scene using both bottom-up, image-based saliency cues and top-down, task-dependent cues.

Bottom-up mechanisms are thought to derive from raw sensory input; this mechanism is based on the image content. Because of the basic visual inputs, bottom-up stimuli originates in the retina and other low level brain areas. As a result, our visual attention is rapidly and involuntarily shifted to salient visual features of potential importance.

Top-down or task-dependent mechanisms are based on our cognitive strategies. The currently given tasks and cognitive factors are biasing our attention toward something we are interested in. Task-dependency is derived from higher brain regions and they project to low visual regions including retina [10]. Since this kind of behavior is voluntary, the process can be influenced.

Since human vision is a rather complex system, it is possible that these two mechanisms work together, in a parallel way.

2.8 Related works on visual perception

In recent studies concerning visual perception [11], [12] the goal is to find numerical correlation between different factors of vision. According to human eye-tracker studies, the work in [11] is based on the discovered fact that gaze fixations are mostly in the center of the images. Moreover, top-down phase can be categorized into a strong and a weak type, and these types can influence the viewing behavior in a different way; according to [12], the two phases have different roles in vision. Other research teams analyze the images specifically according to their structures [3], [13]. Noise effects in the visual perceptions were investigated in [14], finding noise-related neural processes when dissociating the effect of noise on sensory processing from that of the processes associated with overall task difficulties by recording EEG and measuring fMRI responses separately in the same task conditions.

In [9], some important trends have emerged on computational models of focal visual attention that emphasize the bottom-up, image-based control of attentional deployment. First, perceptual saliency of stimuli is derived from low-level visual features but, more than absolute feature strength or other detailed characteristics of the features, what seems to be important for the computation of saliency is feature contrast with respect to the contextual surround. Second, a unique saliency map that topographically encodes for stimulus conspicuity over the visual scene has proved to be an efficient and plausible bottom-up control strategy. Third, attention will not shift unless the currently attended (most salient) location is somehow disabled (otherwise, any model looking for saliency will keep coming back to the most salient location). Inhibition of return is consequently

an essential computational component of attention and it has been recently described as a complex, object-based and dynamically adaptive process that needs to be better modelled. Fourth, attention and eye movements tightly interplay, posing computational challenges with respect to the coordinate system used to control attention. And last, scene understanding and object recognition strongly constrain the selection of attended locations.

An important conclusion of [9] is, that although several computer vision models have been proposed that integrate both attentional orientating and object identification, many exciting research challenges still lie in attempting to provide a more complete account of the interactions between the dorsal and ventral processing streams in primate brains. Controlling where attention should be deployed is not an autonomous feedforward process. According to this article, important future directions for modelling work include modelling of interactions between task demands and top-down cues, bottom-up cues, mechanistic constraints and neuroanatomical constraints such as feedback modulation.

Chapter 3

Theoretical background - Computational Saliency

In this chapter, the notion of computational saliency is introduced and some previously published works are reviewed. These methods attempt to give estimations to the salient objects in pictures or in video streams. There are different approaches for this kind of estimation; however, most of them are based on only image content-based approaches.

3.1 Computational saliency

According to the latest works concerning vision, human fixation behavior is driven by bottom-up and by top-down mechanisms. The phenomenon referring to bottom-up phase called visual saliency. Although human fixation behavior consist of this two type approaches, visual saliency does influence human fixation behavior in a large extent, therefore saliency estimation is an efficient way to give prediction for the path of the gaze.

Computational saliency models transform the input image into a saliency map. Koch and Ullman [15] introduced the concept of the saliency map. This is a two-dimensional topographic representation of saliency for each pixel in an image. The background and the unattractive parts of the images are black while the attractive parts are shades of white: the more whiter an image part is, the more salient the given region is. Human eye movements can be transformed to a similar map, which I refer in the work as fixation density map.

The state-of-the-art bottom-up saliency models mostly consist of three steps. As a first step, some image content-based features are collected which are relevant in saliency detection. According to [9], these features are orientation, motion, luminance, center-surround histograms, multi-scale contrast, color spatial distribution. After collecting these properties for all image scale, some learning method is applied to detect the salient regions: Bayesian framework, Markov chains, conditional random fields, graph-based learning algorithms or data mining-based approaches. The third step is the construction of the ground truth data and the verification. Usually human eye tracking experiments generates this data set. The verification is usually based on statistical correlation or distances.

As it is mentioned in the previous chapter, from engineering point of view, human vision can be considered as a multi-scale sampling device. Therefore some effective bottom-up approaches are multi-scaled based: images or videos are processed in a pyramidal image representation. The input image is transformed and resized into different sizes.

The algorithmic detection of salient objects is a dominant area of research, science and practical applications needs automatic saliency detection. For instance, saliency is for object recognition; image or video quality assessment; detecting region of interest in medical images, in video surveillance systems. More about possible application areas can be read in the last chapter.

3.2 Bottom-up saliency models

In [16] a visual attention system was presented which was inspired by the behavior and the neuronal architecture of the early primate visual system. Multiscale image features were then collected and combined into a saliency map. Attended locations are selected by a neural network, in order of decreasing saliency. The complex problem of scene understanding is broken down by rapidly selecting conspicuous locations to be analyzed in detail. The Graph-based Visual Saliency (GBVS) [17] is a bottom-up visual saliency model. As a first step, some activation maps are constructed on certain feature channels, and conspicuity is highlighted by normalizing them. The results of the model were mostly better than of the previous approach.

In [10], a bio-inspired bottom-up attentional model was introduced which was implemented on Cellular Neural/Nonlinear Network (CNN). This work is unique since it starts with a short overview of the vision of human creatures, the related brain areas and some psychophysical information. After understanding the basic neurobiological processes, the human retina-like model is introduced and successfully tested. Some possible application areas are mentioned, for instance the Bionic Eyeglass project, which is implemented to help to handicapped (blind) people.

In [3] the Esaliency Algorithm was introduced, consisting of a pre-attentive segmentation and an attentive graphical model approximation phase. The probability that an image part is of interest (saliency) was defined using some underlying assumptions: the number of target candidates was small, there was a correlation between visual similarity and target-nontarget labels and natural scenes are composed of clustered structural units. This algorithm started with a pre-selection of possible candidates using some segmentation processes and assigning initial probability for them, and then a similarity based Bayesian network was built, finding the most likely joint assignment. In the processing period, the estimated salient objects were determined by using the spanning-tree of graphs. For the verification, real salient objects were annotated by humans as ground truth in the training phase.

Liu et al. in [13] studied visual attention by detecting a salient object in an input image. In their approach salient object detection is formulated as an image segmentation problem by separating the salient object from the image background. A set of novel features (multi-scale contrast, center-surround histogram, and color spatial distribution) is proposed to describe a salient object locally, regionally, and globally. A Conditional Random Field is learned to effectively combine these features for salient object detection. They constructed a large image database containing tens of thousands of labeled images by multiple users. This database is the first large image database for quantitative evaluation of visual attention algorithms, on which the authors' approach is validated.

3.3 Consequences and the goal of my model

The goal of my research is multi-purpose. First of all, I want to analyze the human fixation patterns and reveal those specific features that could be taken into account when constructing our human vision model. Secondly, I estimate statistical models to describe the eye motion behavior during different experiments on different subjects. Moreover, my goal is to develop a new algorithm for estimating the saliency points in the possible path of the user's eye.

In this model, I wanted to combine the top-down and bottom-up mechanisms mentioned in the previous chapter. The top-down method is integrated by the training videos where a kind of behavior fingerprint about the subject's vision is recorded and used. In my hypothesis, the path of the human gaze is influenced by salient image features as well. The bottom-up approach is applied by processing the recorded images by computer vision algorithms. The results of our algorithm are continuously compared to real human eye movement data derived from our experiments so that I could permanently train our model by tuning the parameters.

Generally we consider two types of salient features: in human vision experiments saliency points are the most attractive regions of the fixation patterns, while the artificial saliency points are the most important characterizing points as image features. The paper describes the fusion of the human and the artificial saliencies in one, trained model.

Recent studies [3], [13] mention object or pattern based methods for the detection of saliency. Contrary to these articles, our proposed algorithm uses the connections' graph of low-level salient points; the artificial points are found in a relative fast and simple way by applying the Scale- Invariant Feature Transform (SIFT) developed by David G. Lowe [18]. The usage of a high number of saliency points is more advantageous than using saliency objects of low number, since the task is the annotation of eye tracking. Practically, estimation and analysis of the eye movement are useful in various fields of science where human-computer interaction is needed. Thus the results of these experiments can be widely used in the future, for instance for improving video surveillance systems or for marketing

purposes. In another application, it can be used as a possible pre-filter for the preparation of medical images in the medical operation theater.

This research presents a new approach by using some basic biological information about the process of human visual exploration and orientation behavior. Personal position information of the human training phase is taken into account: the position and the movement of the head and the distribution of the fixation points on the frames captured from the training videos upgrade the model. The saliency points which seem to be irrelevant (caused by e.g. *twinkle*) are filtered out by using the results of HVS-based statistical models as they do not convey important information.

I present a saliency track building algorithm trained by human data, where the measured human behavior characterized by statistical biological data and saliency features of computer vision algorithms are fused to build the human eye movement into a model. Human behavior is measured by a head-mounted eye-tracker and statistical tables were generated describing the gaze path. Using these data, graphs are constructed and their spanning trees were used for estimating the salient tracks on specific images. The result of the estimations is improved by integrating the statistical knowledge of some models about human vision perception.

In my hypothesis, I assume that the transitions between two consecutive fixation points (eye movements, saccades) are also characteristic of the attention selection processes or the image content; the probability that the eye moves to a given point from another may be characteristic of those points and the features behind them. In this proposed model any high number of possible salient regions is possible in the scene, contrary to the model assumptions in Esaliency [3], while the task is not to find one (group) of the object(s), but to find the path as the human gaze may connect them. I also have the goal to find the salient regions, but this is based on the dense sections of the probable tracking paths and not on the saliency of the object itself. In my proposed method, I attempt to find the regions of the images that will be selected by visual attention and thus will be fixated by the human observers. In my model, according to the perspectives of [9], some top-down behavioral patterns were taken into account. Moreover,

during the recordings, the position of the head was taken into consideration by influencing the training phase.

Chapter 4

Human measurements

This chapter presents how the training videos were recorded by a wearable head-mounted eye-tracker. These videos can be considered as behavioral fingerprints; therefore, some top-down factors are integrated to the model this way.

4.1 The eye-tracker

For the human measurements of training videos, a special wearable, head-mounted eye-tracker device was used (Fig. 4.1). The GigE-60 [19] manufactured by Arrington Research is able to monitor and record videos in which the movement of the users' eye is recorded using two cameras. This device is basically a wearable spectacle frames which consists of two cameras and an infra LED.

The first camera is watching the right eye flashed by an infra LED revealing the detailed texture of the iris. The camera watching the eye does not record a video, it records only some information about the position of the eye, including the location, the size etc. of the pupil. The other camera, the so-called scene camera is located at the center of the frame. The scene camera records the scene in the front of the observer. This camera actually records a video which is real time transferred to a notebook via gigabit network. The observer's gaze can be fused into the video or only simply marked on the data file. During the processing, this video-fused marker was switched off. The sampling rate of the eye camera is quite good, 240 Hz, while the scene camera operates in 30 FPS.

Figure 4.1: A wearable head-mounted eye-tracker device manufactured by Arrington Research. This frame consists of two cameras and an infra LED. The first camera records the scene in front of the subject, while the other analyzes the movement of the eye. The data of the two cameras are integrated and as a result a video is recorded which contains the actual point of the gaze on each time slot. These records have a fundamental role in our model as the training human data for our algorithm by giving information about the subject's vision.



4.2 Eye-tracker-based training experiments

Since we recorded several videos under different conditions, mostly in streets, but some videos were recorded inside buildings as well, the most important fact was to calibrate the eye-tracker precisely, since this problem depends on the distance to a large extent and since it is unrealistic to use a fixed distance between the moving user and the landmarks. The distance can be considered as constant inside the buildings. However, I met the problem of the distance in the cases of the video streams recorded on the streets, since it is unrealistic to use a fixed distance between the moving user and the landmarks. According to experiences, it was better to overestimate the distance during the calibration rather than calibrating the device too close. By the precise calibration, the inaccuracy could be reduced below an appropriate level in the cases when the subjects looked at objects which are either too far or too close. Without the accurate calibration of the eye-tracker, the estimation can be quite rough and useless for training our algorithm. A flat text file is generated during the records which contains the mentioned information about the measurement.

During these human measurements, I had to take into consideration that this device can be strange for its user in the beginning of the tests. In these cases, the behavior of the subjects did not seem to be natural. In order to improve the quality and the reliability of the results, we recorded some test videos so that the subject gets familiar with the eye-tracker. In other cases, I simply removed the first minutes of the recorded videos. The reliability of our algorithm depends on the precise arrangement of the human measurements. According to our experiments, it is enough to record videos with small resolution, since large resolution needs more computation time whereas the results are about the same. Nevertheless, the video streams were recorded with 30 FPS sampling frame rate which was acceptable for us because of producing a training set with a passable size.

This device also includes a professional software environment, namely the ViewPoint EyeTracker [20], to record the videos. For the processing, this software environment was used since the eye movement on the recorded videos can be easily evaluated: the different kind of eye movement patterns (saccade and fixation) could be separated and exported. This separation is based on the speed-difference between the eye movement patterns, as it is described in Chapter 2. The software supports the precise evaluation by playing the video streams slowly.



Figure 4.2: Screenshots from human measurements recorded by the eye-tracker.

The next chapter describes, how the estimation model is generated using these training database.

Chapter 5

Detection of salient regions

In this chapter, the artificial saliency detection is described which is based on different image descriptors. After generating the bottom-up salient regions, some points are labeled as salient using the training data. This chapter presents the steps of the model.

5.1 Saliency point detection and clustering

Here we used the most common salient point detection methods. Salient features of an image are detected and described by applying image descriptors. Image descriptors label some points on the image as salient, based on different features. In this work, I used the Scale-Invariant Feature Transform from David G. Lowe [18] and the Speed Up Robust Feature descriptor (SURF) from Herbert Bay [21].

The generated artificial saliency points will be allocated to the human interest regions in the fusion phase. I analyzed the images by my algorithm developed in MATLAB and in C++ using the OpenCV library [22], which was included for the basic task needed during the development. As the resolution of the frame images was small (320 x 240 pixels), I got an average of 300 points on each frame. I followed the clustering of the SIFT points similarly to that in Bag-of-Features methods [23] and [24], by using the k-means. However, to avoid the problem of random initialization in the sparse data representation in the 128 (SIFT), 64 (SURF) dimensional space, a partly supervised initialization was done by manual center-assignment randomly chosen from the coordinates of the detected SIFT

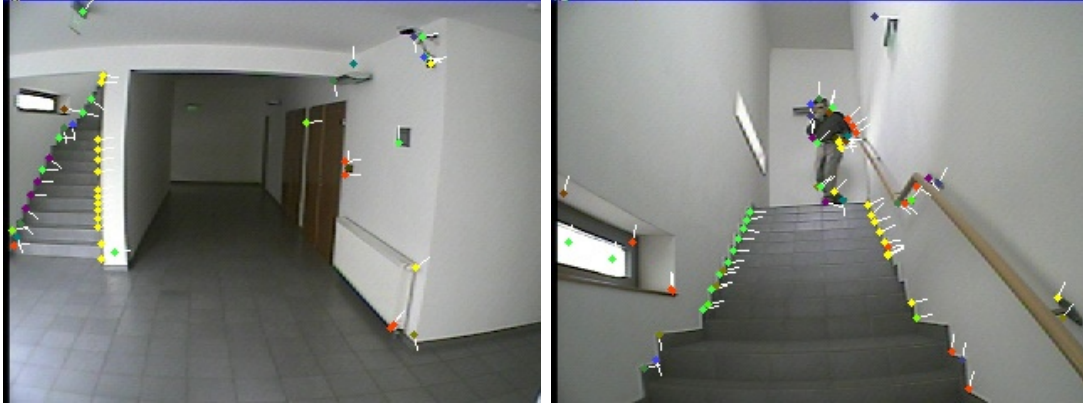


Figure 5.1: SIFT points clustered by k-means algorithm with supervised initialization. The two images are examples for clustering the images by a partly supervised k-means algorithm, where instead of using randomly generated centers for the clusters, we manually assigned these centers taken from the SIFT points. The different frames captured from the same video show that the same parts were generally clustered into the same clusters, for instance the edges of the steps or the objects on the wall.

points. After testing the results of the clustering procedure, I found that the modification was favorable since the same points in different image frames taken in the same place were clustered into the same cluster (Fig. 5.1.).

As this figure shows, the clustered artificial points are able to characterize the images. The optimal value of the clusters was estimated based on the human teaching tests according to our experiences. I discovered that this value depends on the circumstances of the recording and on the subjective features including socialization, interest and feelings. For a few minute-length video, it was enough to define 15-20 clusters; otherwise the values of the transition probability tables were balanced.

5.2 Building transition probability tables and constructing graphs

In the training step, the artificial feature points which were found closest to the current eye position of the human gaze were chosen on each clustered image

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	0.12	0.09	0.05	0.07	0.09	0.03	0.03	0.09	0.12	0.05	0.03	0.10	0.00	0.05	0.07
2	0.06	0.11	0.05	0.06	0.04	0.03	0.01	0.13	0.09	0.08	0.14	0.06	0.03	0.06	0.04
3	0.06	0.05	0.19	0.05	0.09	0.04	0.03	0.11	0.05	0.08	0.08	0.06	0.04	0.03	0.02
4	0.03	0.11	0.07	0.11	0.06	0.03	0.03	0.06	0.13	0.07	0.07	0.11	0.03	0.04	0.04
5	0.07	0.06	0.11	0.04	0.15	0.01	0.02	0.11	0.12	0.08	0.06	0.07	0.02	0.05	0.01
6	0.07	0.00	0.04	0.18	0.04	0.04	0.04	0.11	0.07	0.11	0.11	0.00	0.07	0.11	0.04
7	0.11	0.08	0.08	0.11	0.08	0.00	0.00	0.03	0.08	0.00	0.11	0.05	0.03	0.11	0.11
8	0.02	0.11	0.09	0.04	0.07	0.03	0.05	0.10	0.13	0.09	0.09	0.02	0.06	0.06	0.03
9	0.09	0.05	0.10	0.07	0.11	0.02	0.06	0.10	0.11	0.04	0.03	0.06	0.05	0.08	0.03
10	0.05	0.08	0.08	0.08	0.06	0.06	0.05	0.09	0.13	0.06	0.05	0.06	0.06	0.03	0.05
11	0.04	0.09	0.09	0.08	0.12	0.03	0.08	0.05	0.04	0.05	0.12	0.08	0.03	0.09	0.01
12	0.09	0.05	0.09	0.13	0.09	0.00	0.03	0.08	0.13	0.06	0.03	0.09	0.05	0.02	0.05
13	0.05	0.07	0.09	0.05	0.09	0.07	0.05	0.17	0.03	0.00	0.12	0.03	0.14	0.05	0.00
14	0.00	0.03	0.14	0.03	0.16	0.03	0.00	0.17	0.12	0.07	0.07	0.07	0.02	0.03	0.05
15	0.00	0.06	0.09	0.06	0.15	0.00	0.03	0.12	0.12	0.03	0.06	0.09	0.06	0.15	0.00

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	0.32	0.04	0.06	0.05	0.00	0.01	0.04	0.01	0.01	0.04	0.14	0.01	0.04	0.05	0.19
2	0.04	0.26	0.04	0.07	0.04	0.07	0.06	0.02	0.07	0.05	0.10	0.02	0.05	0.03	0.10
3	0.05	0.06	0.19	0.06	0.01	0.07	0.03	0.02	0.02	0.03	0.19	0.02	0.05	0.03	0.16
4	0.05	0.06	0.06	0.20	0.02	0.05	0.03	0.02	0.03	0.06	0.17	0.03	0.05	0.04	0.14
5	0.04	0.12	0.08	0.04	0.15	0.08	0.12	0.04	0.08	0.04	0.04	0.04	0.04	0.00	0.12
6	0.01	0.05	0.06	0.05	0.02	0.37	0.06	0.08	0.08	0.02	0.07	0.03	0.02	0.03	0.03
7	0.03	0.08	0.03	0.06	0.03	0.06	0.28	0.01	0.03	0.04	0.13	0.01	0.04	0.04	0.14
8	0.02	0.07	0.04	0.07	0.04	0.11	0.02	0.24	0.04	0.07	0.07	0.04	0.07	0.02	0.07
9	0.01	0.09	0.03	0.04	0.03	0.10	0.03	0.04	0.38	0.04	0.05	0.05	0.04	0.01	0.06
10	0.06	0.06	0.04	0.08	0.01	0.05	0.04	0.03	0.03	0.15	0.15	0.01	0.06	0.06	0.16
11	0.04	0.05	0.05	0.09	0.01	0.02	0.03	0.02	0.02	0.04	0.36	0.01	0.04	0.05	0.19
12	0.05	0.05	0.05	0.03	0.03	0.05	0.03	0.03	0.08	0.05	0.08	0.26	0.08	0.05	0.08
13	0.04	0.05	0.05	0.09	0.00	0.02	0.04	0.02	0.05	0.06	0.16	0.01	0.22	0.04	0.15
14	0.06	0.04	0.03	0.06	0.01	0.03	0.04	0.03	0.03	0.09	0.16	0.01	0.07	0.16	0.16
15	0.05	0.04	0.05	0.05	0.01	0.02	0.03	0.02	0.02	0.04	0.22	0.02	0.03	0.05	0.35

Table 5.1: An example for the measured transition probability tables on the machine defined saliency points. This data can be considered as a visual fingerprint about different people. Nevertheless, these tables are dependent on the circumstances of the records, the different parts of the videos and other features. After clustering the salient SIFT or SURF points by k-means, the transition probability table is generated for the points of different clusters as the eye moves from one point to another. The first table shows the human transition for SIFT descriptors, while the second contains the SURF descriptors for the same video stream.

captured from the video clips. This process is a kind of fusion of the artificial and human vision measurements. The number of the switches between the clusters on each consecutive image pairs was counted, and tables were constructed as follows.

For the image pairs, the cluster of the current image and the cluster of the next image were identified, and the number of the steps between these cluster-pair was counted. For our purpose, transition probability tables were constructed about the given videos by normalizing the transition tables (Table 5.1). During the construction, not only those artificial points were taken into account which has the smallest distance to the position of gaze on the given image, but with smaller probability also some feature points around the closest artificial points.

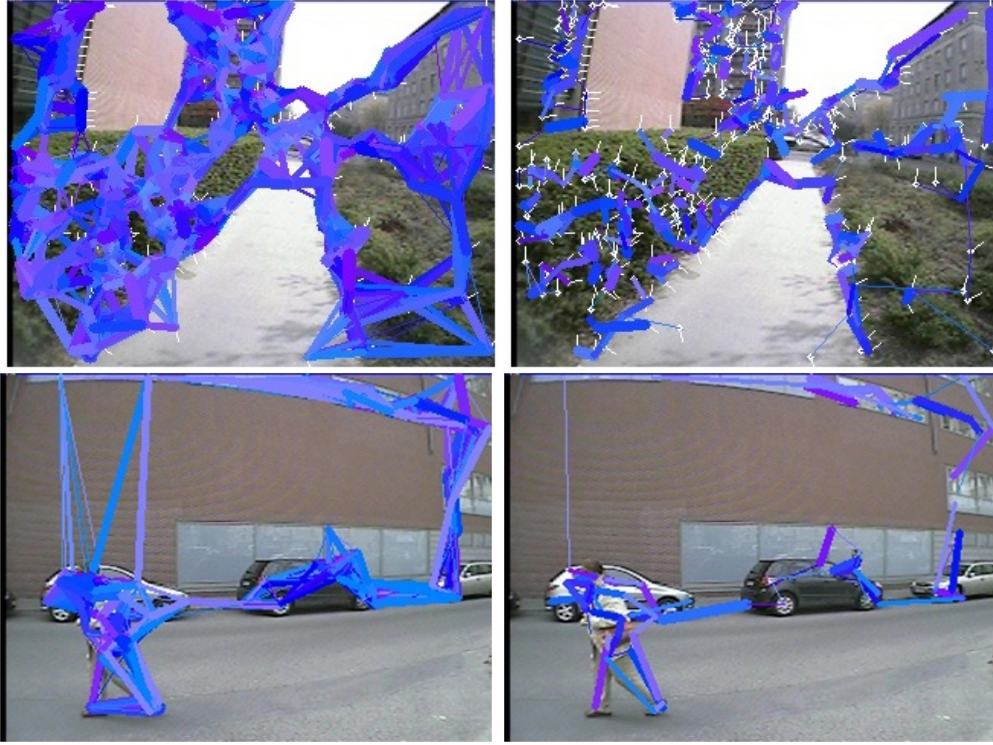


Figure 5.2: Comparison of the two graphs construction approaches. The weights of the graphs are derived from the transitional probability tables. The images on the left show the graphs, in which every vertex is connected to the 8 neighbor vertices. The images on the right show the graphs where the vertices are connected to the 3 neighboring vertices. The images in the upper row show the detected SIFT points as well.

According to the tables, the so-called transition probability graphs were constructed using the possible artificial feature points and the related cluster data of the transition probability tables in order to model the eye movement. Two different types of graphs were generated, which have some common properties as well. The connections between the vertices and the weight of the edges were computed from the transition probability tables. The vertices of both kinds of graphs are the generated SIFT points. All of them were connected to some of the closest neighbor vertices, and the number of the connections was a determined parameter that was constant in the same graph. The weights of the edges were determined by normalizing the values of the transition probability table to 1. In

the first approach, the vertices of the graphs were connected to the N closest neighbor vertices. In our model, the optimal value of N was between 10 and 15. This is a rather complex graph in the sense that the most likely edges have a larger weight, while the less probable edges have a smaller one. Another kind of graphs was constructed as well to simplify the graphs of the first approach (see the right images of Fig. 5.2). In this case, the number of connections between the vertices is lower. The simplification is based on statistical solutions where the connected vertices are assigned within a given radius. The edges with the maximal weights are located about the same place in the two kind of graphs.

5.3 Estimation of the most probable points - Regions of Interest

The goal of this step is to find the most attractive saliency regions of the human fixation pattern. In the graphs, an inverse approach was applied to determine the most possible attention points: the probability of arriving at the given point was determined by this statistical summation. The values of the possibility were then summed and normalized to 1. As a result, a statistical probability was calculated, where only the points above a certain probability were taken into account and were marked as interesting points. Points below a given threshold were filtered out.

The left side of Fig. 5.3 shows some estimation results. The most attractive points inferred by our model are illustrated by purple dots. However, these results can be improved by taking behavior during the perception phase into consideration.



Figure 5.3: Comparison of the most probable gaze points to the real human gaze path. Images on the left images show the estimated points based on the transition probability graphs with statistical summary, which were trained by evaluating some other videos. For verification, images on the right show a real human eye movement of the same images.

Chapter 6

Verification

In the pages below the human eye paths to transition probability-based graphs are compared. Therefore, some test experiments were carried out, and the volunteers' eye movement was transformed into fixation density maps. For the artificial saliency detectors, different image descriptors were applied. The results are compared in the following sections.

6.1 Test experiments by a desktop eye-tracker

In order to compare the results of our model to real human eye movement, test experiments were carried out. Ten volunteer subjects were taking part at this kind of experiment, they were asked for these experiments to verify our model. For these measurements, another eye-tracker device was used. The iView X (Fig. 6.1) is a non-invasive, video-based eye tracking system with monocular dark-pupil tracking. The sampling rate of this device is 240 Hz. Before recording the observers' eye movement, a 13-point calibration was used (Fig. 6.2).

Subjects were asked to sit down before a large monitor in a dark room, where the eye tracker was fixed on a table. Different images were shown to these volunteers on a large monitor, in 1280x1024 pixel resolution. The desktop eye-tracker was set on the desktop in the front of the monitor within a 50 cm distance. The subjects had to be motionless during the experiment. More precisely, their head was stabilized by a stand and they were not allowed to speak.



Figure 6.1: iView X fast, accurate, video-based eye tracker. Volunteers using this device were taking part at verification experiments. Some images were displayed in blocks in a large monitor, and the observers were asked to memorize the images. The recorded human eye movement was then compared to the estimated Regions-of-Interest inferred by our model.

According to some studies [12], different tasks can cause different behavior during human vision. For this reason, the task assigned to the subjects during the experiments must be well deliberated and communicated. Therefore, the images were displayed in blocks in random order and they had different tasks in each block. The given task was displayed before starting a new block.

In the most cases, the subjects had to memorize the displayed images because after some images a question appeared that they had to answer to step forward. The first questions usually were difficult, this way subjects were motivated to be fast and deep.

Between the images, the subjects could relax and twinkling for a few minutes. Then a small fixation cross appeared in the center of the screen, and then the next image appeared. The images were displayed for a certain time, depending

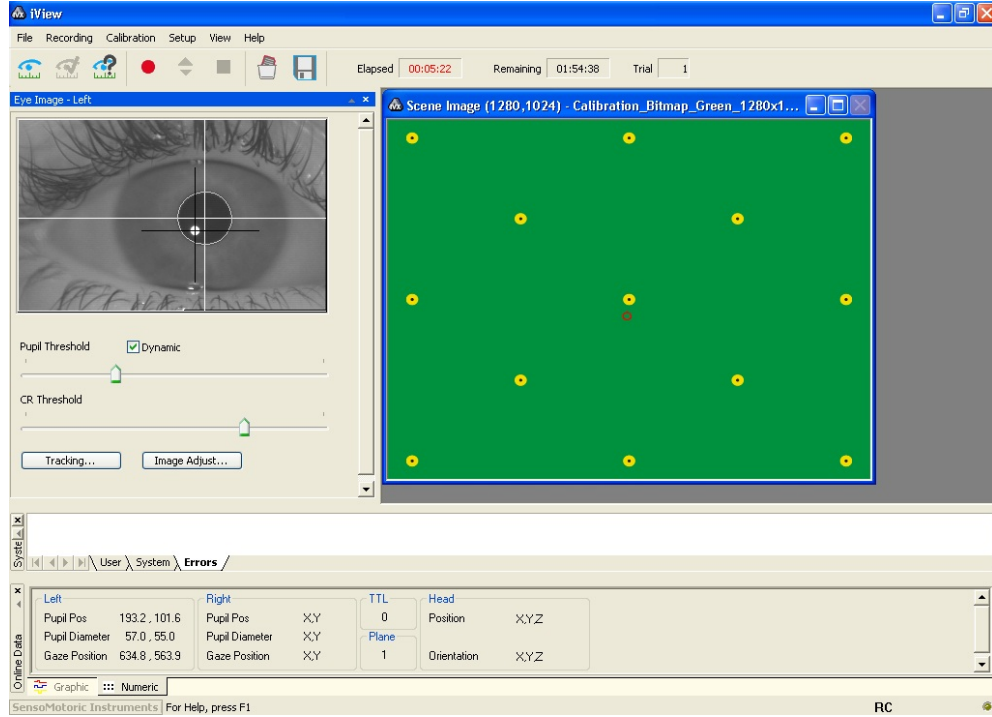


Figure 6.2: A screenshot about the software environment of the iView eye tracker. The observers' left eye is flashed by an infra LED, this way the pupil can be easily detected. The current observer's gaze can be continuously followed on the monitor to check the accuracy.

on the given task. This time could not been reduced by clicking.

Since during the training records the subjects were looking around, they had to look at some images for a certain time and patrol them with their eyes before they move to the next image feeling that they were ready.

These test images were derived from different sources: some of them were frames captured from the human videos recorded by the head-mounted eye-tracker, while other images were acquired from street image taken from public databases [25], [26], however, some images were taken around the university. The path of the eye was recorded by a desktop eye-tracker device [27] and annotated to the images for analysis. Between the consecutive images, a pause was given for relaxing.

When we repeated the tests with the same subject on the same images, the patrolled path was different in all cases. Therefore, the goal of this work is to give estimations for the most likely points of a path based on an averaged human data set, and not for the path itself. On natural images, this path could be different, while on images with a few numbers of salient regions, it was very similar and easily predictable.

6.2 Creating fixation density map

The results of the test experiments were transformed into a fixation density map. This map shows whether a point of the image is a part of the fixation region. A fixation map containing only the fixation regions was filtered by a Gaussian distribution function. The density value was calculated for all the pixels of the images by

$$s(x, y) = \frac{1}{N} \sum_{n=1}^N \frac{1}{2\pi\sigma_s^2} \exp\left(-\frac{(x - x_n)^2 + (y - y_n)^2}{2\sigma_s^2}\right) \quad (6.1)$$

where (x_n, y_n) is the location of n^{th} fixation point and σ_s^2 is the standard deviation of the Gaussian distribution.

6.3 Detection of the fixation regions on human experiments

The human fixation patterns from the test experiments were superimposed on each image. Images on the right of Fig. 5.3 show some examples for the path of the human subjects. I used a similar approach to [28], where identification of the fixation points is based on the velocity-difference between the fixation and saccade points. The wearable eye-tracker device is able to distinguish between the two types of points, differentiating them by a given velocity threshold that can be parameterized. The results generated by the desktop eye-tracker had



Figure 6.3: Input images with human experiment-based regions of interest and the human fixation density maps. After the test experiments, the fixation regions were calculated based on the different speed of eye movement. The labeled images were then converted to the fixation density map. On this map, every pixel is compared to the entire fixation points, following the formula explained in the text. As a result, this kind of two-color map is generated where the color intensity is proportional to local image saliency.

to be processed manually. The method was based on the distance between the neighboring points, assuming the temporal constancy of the sampling. However, I assumed that the movement of the eye consists of only fixation and saccade points, and we did not pay attention to the glissade periods like the software environment of the wearable eye-tracker. This way, groups of fixation points were collected and painted to the original images as well.

6.4 Filtering the estimated fixation patterns

Since filtering the estimated fixation patterns influences the results, we present this selection method in this section. The goal of filtering the estimated points is to improve the reliability of our model. Therefore, we statistically analyzed the biology-driven behavior of the users detecting statistical correlation.

In some cases, we should take into account the movement and the current position of the head. We can assume that the head usually turns into the direction of activity, objects and interesting landmarks. As a result, the most attractive points are usually in the center of the images. Although bottom-up properties of the peripheral regions can be similar to the points located at the center of the images, their importance is less for the subjects due to some top-down factors. In other words, the probability that a peripheral region is attractive for a human subject is less likely, the probability of arriving to that point by the gaze can be reduced. The extent of the distance-based filtering was determined by the analysis of the human test experiments. The images were divided into rings which had the same origo, while their radius was different. The spatial distribution of the



Figure 6.4: Comparison of the fixation patterns on images with and without filtering. The left image shows the generated output where estimated points are located at the periphery of the images as well. The right image is after filtering the less probably attractive fixation points according to HVS-based statistics.

detected fixation points was analyzed and behavior-tables were created including transitional probability tables about the movements from one fixation point to the next one. After generalization, the numerical results were integrated to our model, and probabilities of the peripheral edges were reduced according to their distance to the center. Fig. 6.4 shows the comparison of an original image to the filtered version.

6.5 Comparison of the estimated points to real measurements

To compare the prediction of my model, the spatial saliency map of the images were created. Similarly to the fixation density maps of the test experiments mentioned in Section 6.1, the spatial saliency map shows how attractive an image region of the image is. The subjective comparison of these maps can be found in Fig. 6.5, which contains the input images, fixation density maps, and spatial artificial saliency maps.

For the objective comparison of the artificially inferred points to other state-of-the-art methods ([16] and [17]) and to human test experiments, the symmetric Kullback-Leibler distance was used. The saliency maps can be considered as two dimensional probability density functions. In the unattractive regions (black parts of the saliency maps) the value of the function is actually zero, while in the other parts the function is 1. Let us denote the probability function of the human measurements by $P(x)$ and the probability function of the artificially measured points by $Q(x)$. The symmetric Kullback-Leibler distance is

$$D_{KL}(P, Q) = D_{KL}(P||Q) + D_{KL}(Q||P), \quad (6.2)$$

where

$$D_{KL}(P||Q) = \sum_i P(i) \log \left(\frac{P(i)}{Q(i)} \right) \quad (6.3)$$

and

$$D_{KL}(Q||P) = \sum_i Q(i) \log \left(\frac{Q(i)}{P(i)} \right). \quad (6.4)$$

If the two density function is the same, the KLD value is 0. If the division in the formula would be zero, for that i value the sum is zero as well.

The estimated points of each test experiment were transformed into fixation density maps, and these maps were summed into one map. The comparison of the inferred points and the real human path can be found on Fig. 5.3. The points determined by our model describe the most probable points of the path that a human subject patrolled with his or her eyes. The numerical results can be found on Table 6.1 and in Figure 6.6.

However, some other points were detected by the algorithm, mostly at the periphery of the images. Most of them can be successfully filtered out (Fig. 6.4). These points can be typical for other volunteers; therefore they can be likely to have been visited by one of our subjects as well. Our algorithm generates points which have a high correlation to generalized data generated by several users.

	Our model vs. human	Itti vs. human	GBVS vs. human	Random graphs vs. human	Uniform distribution vs. human
# 1	1.524	1.222	1.099	1.806	1.863
# 2	1.162	1.138	1.169	1.412	1.641
# 3	0.251	0.958	0.897	0.915	1.524
# 4	0.634	1.134	0.968	1.310	1.465
# 5	0.344	1.749	1.657	2.092	2.391

Table 6.1: The numerical results. The first column contains the KLD distance between our prediction and the averaged human experiments. The second and the third columns show the similar results of two state-of-the-art methods. The averaged fixation density map was compared to the uniform distribution as a reference: the accuracy of the model reduces if the transition probability table is a equal distribution table (fourth column). The last column is a reference where the fixation density maps have uniform distribution.

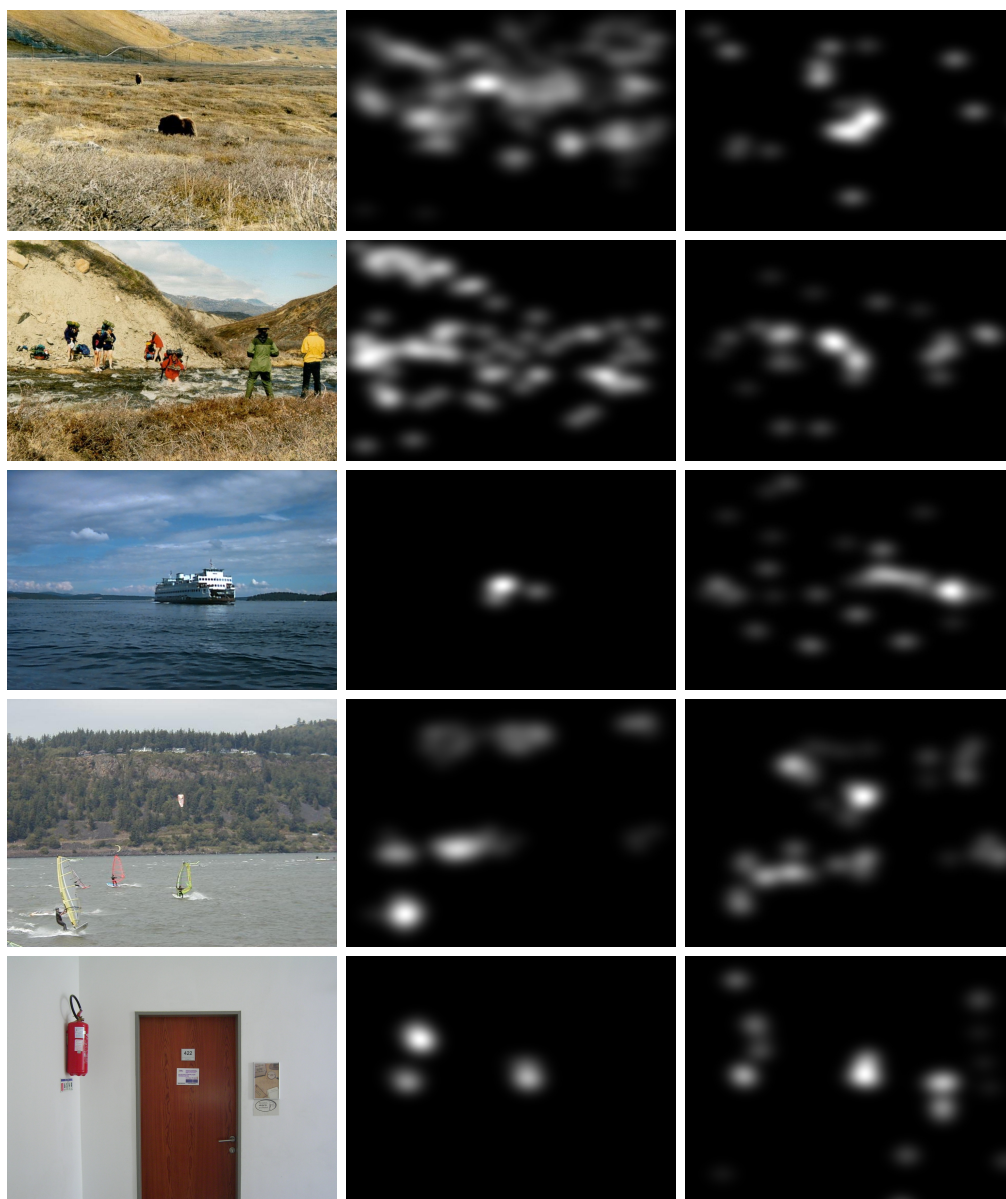


Figure 6.5: Comparison of the artificial fixation patterns on Images #1-5 to the real human salient regions. The input images are in the first column, the second column contains the artificial saliency maps generated by my model. The color intensity of white shows how a local region was labeled as salient by applying my approach. Human fixation density maps can be found in the third column, these maps were constructed by summing the fixation density maps of all observers. Besides these maps show the salient regions of the images, they are needed for the objective comparison of the human and artificial regions of interest by calculating the Kullback-Leibler distance.

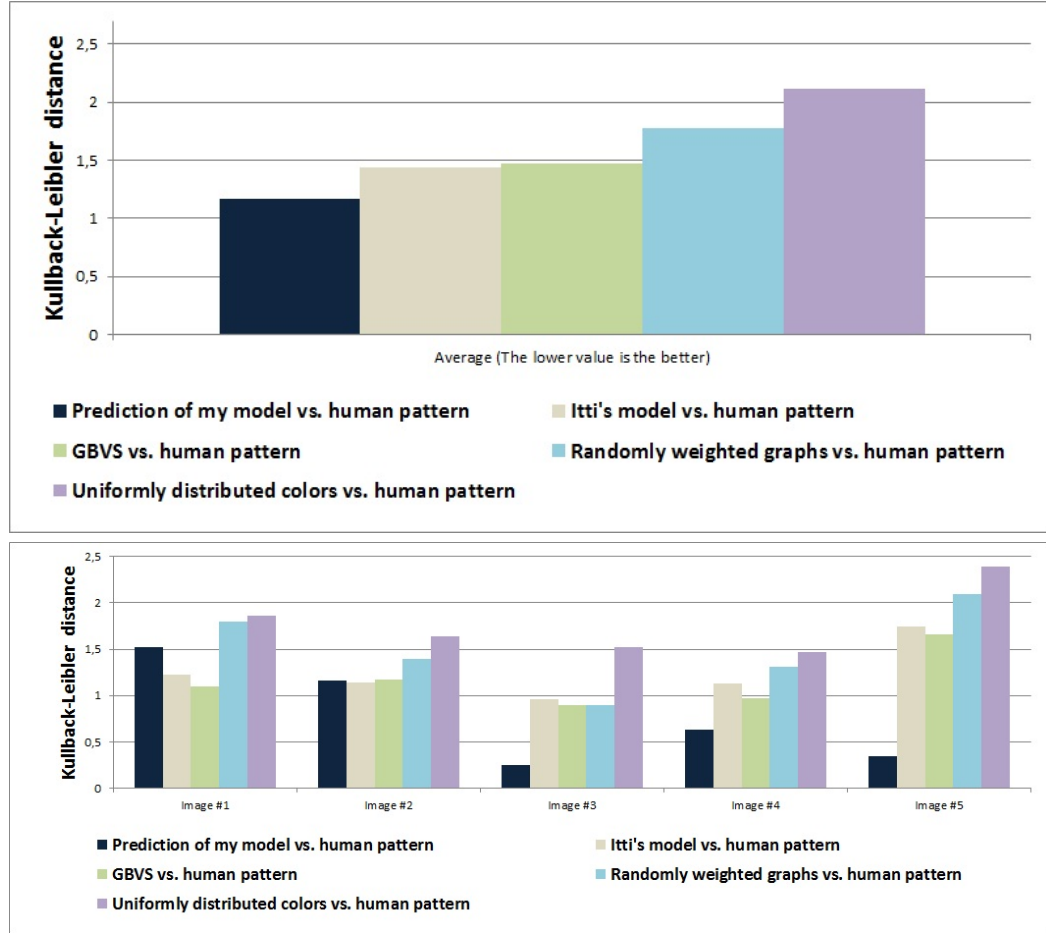


Figure 6.6: Diagrams about the results of the predictions. Our data set contains 50 images. In both cases, the vertical axis is the Kullback-Leibler distance between the saliency maps of the algorithmic predictions and the human fixation density patterns. The lower this value is, the more accurate the prediction is. The first column contains the KLD between our model and the human patterns. Our results are compared to two state-of-the-art methods: [16] (second column) and [17] (third column). The fourth column shows the distance, if the transition probability tables have uniform distribution. This means that biological data does significantly influence the results. The last column shows a reference where the fixation density maps have uniform distribution (gray maps). On the second diagram, the horizontal axis shows some images taken from the test experiments.

Chapter 7

Applications

This chapter mentions some possible application areas where gaze estimations can help science or the everyday life. Since the parameters of my model can be easily modified, it can be used for different purposes.

7.1 Human-Computer Interaction

From a scientific approach, Human-Computer Interaction (HCI) is a discipline concerned with the evaluation and implementation of interactive computing systems for human use and with the study of major phenomena surrounding them. The goal of improving HCI is to make our work with the computer more efficient and comfortable. The focus is on interaction between one or more humans and one or more computational machines. In the recent years, computational power has increased in a large extent which opens new dimensions in computing. This performance creates the possibility for embedding bio-inspired features into computational models. Due to the fast convergence of the biology and informatics, the role of biology is on increase; therefore, computers can analyze the behavior of humans. Moreover, personal behavioral patterns and biological models can be faster computed than ever.

HCI is an interdisciplinary area, it is essential in computer science or in psychology where the application of theories of cognitive processes and the empirical analysis of user behavior is studied and modeled. Because human-computer interaction studies a human and a machine in communication, it draws from supporting knowledge on both the machine and the human side. On the machine side, techniques in computer graphics, operating systems, programming languages, and development environments are relevant. On the human side, communication theory, graphic and industrial design disciplines, linguistics, social sciences, cognitive psychology, and human performance are relevant. Consequently, the subject of my work is one of the key areas of interest for HCI.

7.2 Medical applications

The model introduced in this paper can be applied for medical images as well. Since specialists and laic observers may have different eye movement patterns, first of all some training experiments should be carried out with the help of volunteers who are divided into two groups: laic subjects and professional radiologists. In these experiments, medical X-ray images about male and female chests are displayed and the observers' eye movement is recorded by the previously mentioned iView desktop eye tracker. The goal of this experiment is to find different eye movement patterns between the two groups and train the model with them. As a result, the averaged fixation density maps can be calculated which are expected to be different for laic observers and radiologists. After the training experiments, the image descriptors can be collected on the medical images. The parameters of the detection had to be optimized, otherwise the number of points were less. After optimizing, the transition probability tables can be calculated and the graphs can be constructed.

As a pre-experiment, the fixation density maps have been created for some laic observers and radiologists, and the model has been applied for both cases. As it is expected, the model predicted different salient points in the both cases. It goes without saying, this estimations are more accurate if the model is applied on the radiologists' training set.

After combining the model with another image processing tools, the results of the estimations can be used for detecting extraordinary lesions automatically, this way the work of radiologist can be helped and improved. If someone had to analyze lots of medical images in a short time, this algorithm can be a pre-filter which attempts to detect the lesions, this way the work can be faster and some mistakes can be fended off.

7.3 Marketing purposes

Analyzing human behavior during watching advertisements or during web browsing is an important field of research. Advertisement industry is interested in targeted advertisements, in most cases it is essential where passers-by gaze is directed during watching a shop window or in web pages. For this purpose can my model applied, having an appropriate training set.

7.4 Eye movement analysis and fatigue detection

Eye trackers can detect the current eye movement of pilots in passenger aircrafts. For this purpose, different training sets are needed: senior pilots and beginners eye movement should be recorded under different circumstances (e.g. tiredness, freshness). Eye movement patterns can be analyzed by comparing the real eye movement to the artificially estimated regions based on the different training sets. The current patterns are typical for a training group; therefore, the mental status of the pilots can be estimated. Since fatigue of the pilots is responsible for the most accidents, passengers live can be saved in the future by applying eye tracking technology combined with eye movement estimation. Moreover, the quality of training of new recruits can be increased as well.

Summary

In this paper, I have presented an algorithm for inferring the eye movement on images and even on real scene situations for which HVS-based statistical data were used discovered in our experiments and in recent state-of-the-art studies. My proposed method can give a good estimation of visual saliency without object detection. It is based on low-level features (artificial saliency points) which are connected in a graph based on the transitional probabilities among feature points measured in human experiments. Moreover, the proposed method may give a good estimation of possible eye paths through the image as well; artificial saliency points where the graph of possible eye paths is dense may assign region of interest areas. In the model, I combined two approaches. During the video records, some kind of top-down features were recorded and used as a training set. During the processing of the captured frames, bottom-up methodology was applied taken from image processing algorithms. The paper describes an HVS-based statistic as well in order to improve the model of the human eye movement and to estimate visual pathways on images and scenes. Statistical data about the correlation of the eye and assumptions about head movement were also integrated and combined to a training-based graph-construction algorithm which is a new approach for modeling human vision and it seems to be promising. The reliability of our model is getting optimized by extending the human data set and by developing a filtering model. As a step forward, the amount of built-in biological, perceptual and head movement models will be increased in the future. I will also investigate how to use basic object components for localizing fixation patterns instead of simple feature

points. Moreover, our method should be objectively compared to more other methods and evaluated in different ways as well.

Acknowledgements

I would like to thank Dr. Tamás Szirányi, my supervisor, for suggesting this interesting topic and for his forward-looking questions and advice. Moreover, I would like to say thank the help of Dr. Zoltán Vidnyánszky for sharing his knowledge about visual perception and his suggestions on psychophysics.

I greatly acknowledge the help of the volunteer subjects who were taking part at the training experiments and evaluation tests.

This work was supported by the Computer and Automation Research Institute, Hungarian Academy of Sciences (SZTAKI) and by the Hungarian Scientific Research Fund.

References

- [1] S. Szalai, T. Sziranyi, and Z. Vidnyanszky, “Tracking the saliency features in images based on human observation statistics,” *Lecture Notes in Computer Science*, vol. 7252, pp. 219–233, 2012. (document)
- [2] S. Szalai, T. Sziranyi, and Z. Vidnyanszky, “Estimation for the eye movement using the spanning tree of transition probability-based graphs,” *2nd International Conference on Cognitive Infocommunications (CogInfoCom)*, pp. 1–5, 2011. (document)
- [3] T. Avraham and M. Lindenbaum, “Esaliency (extended saliency): Meaningful attention using stochastic image modeling,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, pp. 693–708, 2010. 1, 2.8, 3.2, 3.3
- [4] B. M. ter Haar Romeny, *Front-end vision and multi-scale image analysis*. Springer, 2003. 2.1, 2.4, 2.5
- [5] G. Osterberg, “Topography of the layer of rods and cones in the human retina,” *Acta Ophthalmologica*, vol. 6, pp. 1–103, 1935. 2.2
- [6] G. C. DeAngelis, I. Ohzawa, and R. D. Freeman, “Receptive field dynamics in the central visual pathways,” *Trends Neurosci.*, vol. 18, pp. 451–458, 1995. 2.3
- [7] J. J. Koenderink, “The structure of images,” *Biological Cybernetics*, vol. 50, pp. 363–370, 1984. 2.4

-
- [8] J. G. Daugman, “Two-dimensional spectral analysis of cortical receptive fields profile,” *Vision Research*, vol. 20, pp. 847–856, 1980. 2.4
 - [9] L. Itti and C. Koch, “Computational modelling of visual attention.,” *Nature reviews. Neuroscience*, vol. 2, pp. 194–203, Mar. 2001. 2.7, 2.8, 3.1, 3.3
 - [10] A. Lazar, *Modeling Visual Attention*. PhD thesis, Peter Pazmany Catholic University, Budapest, Hungary, 2008. 2.7, 3.2
 - [11] P.-H. H. Tseng, R. Carmi, I. G. Cameron, D. P. Munoz, and L. Itti, “Quantifying center bias of observers in free viewing of dynamic natural scenes.,” *Journal of vision*, vol. 9, no. 7, 2009. 2.8
 - [12] T. Betz, T. C. Kietzmann, N. Wilming, and P. Konig, “Investigating task-dependent top-down effects on overt visual attention,” *Journal of vision*, vol. 10, no. 3, 2010. 2.8, 6.1
 - [13] T. Liu, Z. Yuan, J. Sun, J. Wang, N. Zheng, X. Tang, and H.-Y. Shum, “Learning to detect a salient object,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, pp. 353–367, 2011. 2.8, 3.2, 3.3
 - [14] E. M. Banko, V. Gal, J. Kortvelyes, G. Kovacs, and Z. Vidnyanszky, “Dissociating the effect of noise on sensory processing and overall decision difficulty,” *Journal of Neuroscience*, vol. 31, no. 7, pp. 2663–2674, 2011. 2.8
 - [15] C. Koch and S. Ullman, “Shifts in selective visual attention: Towards the underlying neural circuitry,” *Human neurobiology*, vol. 4, pp. 219–227, 1985. 3.1
 - [16] L. Itti, C. Koch, and E. Niebur, “A model of saliency-based visual attention for rapid scene analysis,” *IEEE Trans Pattern Analysis and Machine Intelligence* 20, vol. 11, pp. 1254–1259, 1998. 3.2, 6.5, 6.6
 - [17] J. Harel, C. Koch, and P. Perona, “Graph-based visual saliency,” in *Advances in Neural Information Processing Systems 19*, pp. 545–552, MIT Press, 2007. 3.2, 6.5, 6.6

- [18] D. G. Lowe, “Distinctive Image Features from Scale-Invariant Keypoints,” *International Journal of Computer Vision*, vol. 60, pp. 91–110, 2004. 3.3, 5.1
- [19] *GigE-60*, <http://www.arringtonresearch.com/LaptopEyeTracker1.html>. 4.1
- [20] *ViewPoint Software*, <http://www.arringtonresearch.com/viewpoint.html>. 4.2
- [21] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, “Surf: Speeded up robust features,” *Computer Vision and Image Understanding*, vol. 110, pp. 346–359, 2008. 5.1
- [22] *Intel OpenCV Library*, <http://opencv.willowgarage.com/wiki/>. 5.1
- [23] G. Csurka, C. Dance, L. X. Fan, J. Willamowski, and C. Bray, “Visual categorization with bags of keypoints,” *Proc. ECCV Int. WS Statistical Learning in Computer Vision*, 2004. 5.1
- [24] J. Sivic and A. Zisserman, “Video google: a text retrieval approach to object matching in videos,” in *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, pp. 1470–1477 vol.2, oct. 2003. 5.1
- [25] *Street Scene Database*, <http://cbcl.mit.edu/software-datasets/streetscenes>, 2007. 6.1
- [26] *Ground Truth Database*, <http://cs.washington.edu/research/imagedatabase>, 2009. 6.1
- [27] *SMI iView X Hi-Speed eye-tracker*, <http://www.smivision.com/en/gaze-and-eye-tracking-systems/products/iview-x-hi-speed.html>. 6.1
- [28] M. Nystrom and K. Holmqvist, “An adaptive algorithm for fixation, saccade, and glissade detection in eyetracking data,” *Behavior Research Methods*, vol. 42, pp. 188–204, 2010. 6.3