Object and feature based modelling of attention in meeting and surveillance videos

Karlsson, Stefan

Object and feature based modelling of attention in meeting and surveillance videos

A thesis presented to the University of London

by

Stefan Karlsson

for the degree of

Master of Philosophy

in

Electronic Engineering

School of Electronic Engineering and Computer Science,
Queen Mary, University of London,
Mile End Road, London, E1 4NS

January, 2012

I confirm that the work presented in this thesis is my own and the work of other persons is appropriately acknowledged.

Sincerely yours,

Stefan Karlsson

Abstract

The aim of the thesis is to create and validate models of visual attention. To this extent, a novel unsupervised object detection and tracking framework has been developed by the author. It is demonstrated on people, faces and moving objects and the output is integrated in modelling of visual attention. The proposed approach integrates several types of modules in initialisation, target estimation and validation. Tracking is first used to introduce high-level features, by extending a popular model based on low-level features[1]. Two automatic models of visual attention are further implemented. One based on winner take it all and inhibition of return as the mechanisms of selection on a saliency model with high- and low-level features combined. Another which is based only on high-level object tracking results and statistic properties from the collected eye-traces, with the possibility of activating inhibition of return as an additional mechanism. The parameters of the tracking framework thoroughly investigated and its success demonstrated. Eye-tracking experiments show that high-level features are much better at explaining the allocation of attention by the subjects in the study. Low-level features alone do correlate significantly with real allocation of attention. However, in fact it lowers the correlation score when combined with high-level features in comparison to using high-level features alone. Further, findings in collected eye-traces are studied with qualitative method, mainly to discover directions in future research in the area. Similarities and dissimilarities between automatic models of attention and collected eye-traces are discussed

Acknowledgements

First I would like to thank my first supervisor Prof. Andrea Cavallaro for his support in my research as well as valuable comments on my report during writing. I also would like to thank my second supervisor Dr. Anastasios Tombros for providing important guidelines. Further, I would like to thank Dr. Emilio Maggio for providing me with a particle filtering framework for object tracking and Dr. Murtaza Taj for his framework in moving object and pedestrian detection, in combination, enabling me to develop the modules necessary to study visual attention. Also, special thanks to Assoc. Prof. Peter Ellmark, Dr Emilio Maggio, Dr. Ioannis Tziakos and Dr. Murtaza Taj who have proofread the text and given comments.

This thesis is dedicated to my mother Kerstin Karlsson and my father Bo Karlsson.

Contents

Associated Publications

The following publications have been produced in association with this thesis:

Journal Papers

[Karlsson et. al.(2008)] Karlsson, Taj, and Cavallaro, "Detection and tracking of humans and faces," EURASIP Journal on Image and Video Processing, 2008

List of Symbols

σ_p	Pyramid level	43
r	Red colour component	43
g	Green colour component	43
b	Blue colour component	43
R	Red feature	43
G	Green feature	43
B	Blue feature	43
Y	Yellow feature	43
I	Intensity	43
O	Orientation	43
F	Across scale difference	43
N	Non-linear normalisation operator	43
θ	Angle	43
l	Feature	44
L	Set of features	44
C	Conspicuity	44
S	Saliency	44
V	Optical flow	49
H	Entropy	49
P	Probability	49
R	Spectral residual	49
$\mathbf{f}$	Feature vector	51
p, q	Probability density function	57
$\mathbf{x}$	Position in state space	57

Glossary

Adaboost Short for "adaptive boosting", a machine learning algorithm that uses the output of several weak classifiers to make a final decision, 53

animacy Degree of being sentient or alive, 31

blob Set of connected pixels, 55

cold cognition Cognition driven by information, 30

endogenous Caused by factors inside the organism or system, 27

exogenous Caused by factors or an agent from outside the organism or system, 28

ghost Area falsely outputted as foreground due to a foreground object moving from this area, 54

hot cognition Cognition driven by affect or motivation, 30

inferior temporal cortex An area of the brain crucial for visual object recognition, 32

inhibition of return (IOR) A mechanism that temporarily inhibits reallocation to previously attended points, 3

lateral intraparietal area A part of the intraparietal sulcus located at the lateral surface of the parietal lobe, thought to be involved with saccade generation and working memory in guiding eye movements, 27

MPEG-1 The standard on which such products as Video CD and MP3 are based, 40

MPEG-2 The standard on which such products as Digital Television set top boxes and DVD are based, 146

MPEG-4 The standard for multimedia for the fixed and mobile web, 40

phenomenology Phenomenology takes the intuitive experience of phenomena (what presents itself to us in phenomenological reflection) as its starting point and tries to extract from it the essential features of experiences and the essence of what we experience., 19

pop-out An effect where a part of the stimuli stands out in comparison to its neighbourhood, 25

posterior parietal complex Receives somatosensory, proprioreceptive, and visual inputs and plays a role in voluntary movements, 32

primary visual cortex A brain area highly specialized for processing information about static and moving objects and for pattern recognition, 32

priming A process in which the processing of a target stimulus is aided or altered by the presentation of a previously presented stimulus, 29

retino-geniculo-striate A pathway to the primary visual cortex that conveys elemental information for visual perception[2], 32

saccade Fast eye-movement between fixation points, 21

saliency map A 2D array that encodes the relative attractiveness of each point to visual attention, 18

smooth pursuit Following a moving object with gaze, 46

subliminal below threshold for conscious perception, 29

superior colliculus A major component of the vertebrate midbrain, processing input from the eyes as well as other sensory systems, 27

TEO A part of extrastriate visual cortex associated with form and color vision, 24

the pulvinar of thalamus the most posterior region of the thalamus, 24

V1 See primary visual cortex, 32

V2 A extrastriate visual cortical area sending and receiving strong feedback connections to V1, 32

V4 One of the visual areas in the extrastriate visual cortex of the macaque monkey. The homologue in humans is disputed, 24

vergence The simultaneous movement of the eyes in opposite direction to obtain or maintain binocular vision, 30

visual cortex The part of the cerebral cortex responsible for processing visual information, 17

what stream A neural processing pathway that is involved with object identification, 31

where stream A neural pathway that processes spatial information, 31

Chapter 1

Introduction

1.1 Motivation

Visual attention is a mechanism by which the organism chooses particular points of interest in the surroundings[3]. A small focal area around an attended point is processed with extraordinary resources in comparison to other areas of visual input. About 50% of the primary visual cortex is devoted to processing input from the central 2% of the visual field[4]. It is actually only in this small area that the visual input is clear enough to make an accurate picture of the surroundings, which is surprising since humans often experience a clear 180° view. The explanation for this is that the brain actively fills in what is missing in the rest of the view. A clear example of this phenomenon is the blind spot where no information is received at all. But people are normally not at all aware of this gap in the receptive fields.

There are several reasons for a selective visual attention mechanism. First, obviously humans have a limited view of the surroundings, which makes body, head and eye movements necessary to gather enough data about the surrounding. Further, the mechanism of attention provides a way to serially process visual input[5]. The process of scene understanding is thus rapid with limited capacity and enables real-time operation despite these limitations of the brain. During evolution it has been important for animals to swiftly become aware of important events in the surroundings.

Many factors are influencing the way allocation of attention is done. In frogs, the eyes are largely comparable to humans eyes, but the processing is different. Low- and

perhaps mid-level vision is involved in localising small moving object for consumption of possible prey. The actual capturing of the pray is instinctual. In humans higher-level representations and processes are involved. There is a lot of evidence that cognitive state and personality affect the way people look[6, 7]. For example "task" is a high-level concept that humans engage in as opposed to being hungry in the case of a frog.

So, where do people look, and how can you take advantage of a visual attention model in computer vision. There are plenty of possible application areas of visual attention in multimedia processing, illustrating the importance of studies in the field. One is to utilise a visual attention mechanism for scene interpretation[8], to retrieve semantic information from video sequences. A saliency map can contribute to highlight important events in a CCTV camera capture. An interesting research question would be if the attention mechanism can help to make sense out of a video sequence. For example, particular series of fixations could possibly be a means to classify events. Let's say a thief is trying to steal a bag at an airport. Given the importance of features like faces and moving objects, a saliency map could easily encode the face and the moving bag as important areas to attend, and would probably produce a specific trace of attentional fixations. The output of such a system would be a classification of video events after comparison with a trained database.

This thesis investigates human visual attention and exploit this to model visual attention. Such modelling of visual attention for computer vision applications is something that has not been thoroughly studied, due to the complicated nature of visual attention, and given the impact of such a system it is worth doing. However, as I will mention in this thesis visual attention has been used for example in video compression applications[9], object detection[10] and object recognition[11]. My work covers first the state of the art on visual attention. Then experimental work is presented that has been aimed at trying to model the visual attention mechanism. First, a saliency map from the low-level features: colour contrast, intensity contrast, orientation, similarly to previous work[1, 12, 13], and high-level features as illustrated in Fig. 1.1, has been built. The second goal was to generate automatic eye-traces on dynamic media which is done by two different approaches described in Chapter 4. The first one involves shifts of attention with winner-take-it-all and inhibition-of-return (IOR) mechanism on low- and high-level feature maps, and the second

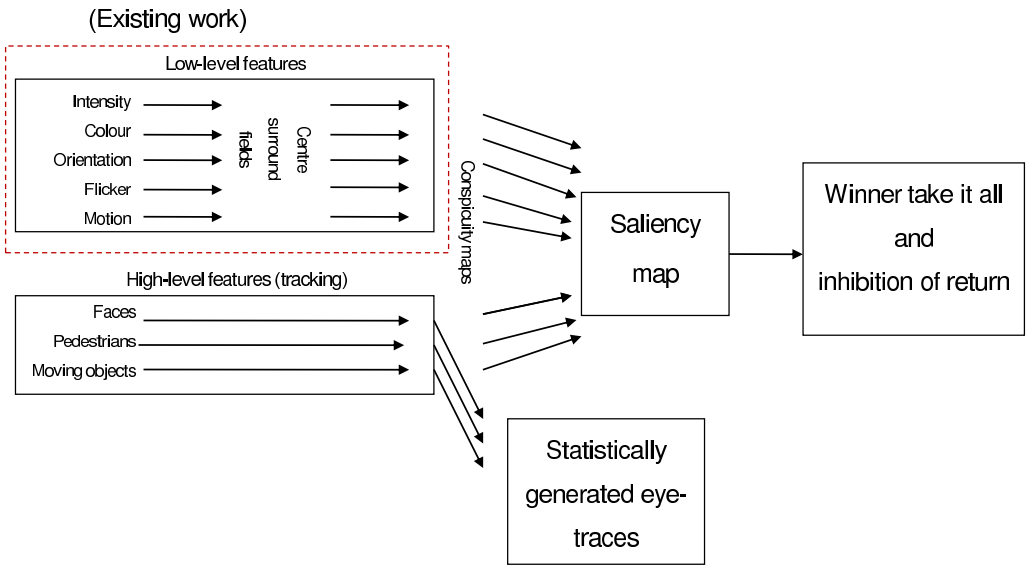


Figure 1.1: Low-level features: intensity, colour and contrast (extracted with existing methods[12]) and high-level features: faces, moving objects and pedestrians. These can be utilised to generate a combined saliency map and further automatic eye-traces given winner-take-it-all and inhibition-of-return. High-level features can also be fed directly into the module for statistical generation of eye movements.

one involves shifts of attention between high-level objects based on statistical properties of sampled eye-tracking data.

1.2 Main contributions

Main contributions of the work are as follows:

Face, pedestrian and moving object tracking: Since visual attention to high-level object was to be studied the aim was to develop an algorithm that extracts such objects automatically. A tracking system for faces, humans and four and five dimensional tracker for moving objects has been developed[14], where others in the MMV research group have contributed with particle filtering[15] and people/moving object detection[16]. The system is validated against state of the art methods, and its parameters are optimised for accuracy and precision.

Models of visual attention: Models of visual attention have been developed and validated to test theories of its functionality. First, correlation scores have been calculated between saliency based on combined low– and high–level features, and eye-

traces of subjects watching the same videos, enabling us to study which features attract attention the most, how they interact, as well as the spatial extend of object based attention. The eye-tracking data has been studied phenomenologically and quantified to build one model based on winner-take-it-all and inhibition of return as well as one model based on Gaussian statistics. In the first case comparisons are made between with and without including high-level features. In the second case comparisons are made with the original eye–traces, and a phenomenological analysis is presented.

1.3 Organisation of the thesis

This thesis consist of the following chapters:

- *Previous work*: State of the art in visual attention modelling as well as detection and tracking is covered.
- *Object detection and tracking*: The implemented system for object detection and tracking is described in detail.
- *High-level and low-level visual attention*: The generation of combined saliency map with low-level and high-level features, as well as eye-tracking measurements are described. Validation of the proposed model in Fig. 1.1 is presented. Qualitative descriptions of eye-traces are given. Two models of generating eye-traces are described where one uses statistical properties of eye-traces in terms of saccade [1] frequency and saccade speed.
- *Conclusions*: Achievements are summarised and conclusions drawn.

[1] A saccade is fast eye-movement between targets, see section 2.3.1 for a thorough description

Chapter 2

Previous work

2.1 Introduction

The most used description of visual attention is the analogy of a spotlight. By directing the eyes to different areas in the surrounding these are highlighted with respect to the rest. This is consistent with the fact that the centre of the eye is packed with photo-receptors and gives the picture that visual attention is used to successively scan the surroundings for interesting objects.

The study area of visual attention extends from neurology, psychophysics, cognitive psychology, cognitive science and computer vision. Understanding the attentional mechanism is challenging. As a result there is a lack of a big picture of visual attention, merely theories of parts of its functionality.

When it comes to visual attention within computer vision or multimedia processing, studies have mostly been concerned with images, and have been focused on low-level features. In images low-level features like intensity gradient, colour gradient and orientation have been used to predict visual attention. In videos motion and flicker tend to attract visual attention more[1]. These studies compare collected eye-traces, defined as the path visual attention follows over time, and compare to the model. Eye-traces are more similar between different persons on videos, which makes videos more feasible to create an automatic model.

Models of eye-movements are usually governed by inhibition of return. This produces something similar to human scan patterns on images but as will be shown is not sufficient

for videos. Here eye-movement only consist of saccades , i.e. fast eye-movements from one point to another, but also smooth-pursuit, i.e. a consistent tracking of one object over time.

In this chapter aspects of visual attention and theories about these are covered, as well as some background on work within the field of computer vision and state of the art in detection and tracking. A summary of papers read about modelling visual attention is presented in table 2.1.

Ref	Method	Application	Features	Purpose	Validation	Level	Interest
[17]	Eye tracking and summing of feature maps correlated	Stereo image pairs	Depth plus colour contrast	To test the importance of disparity in attention	Yes	Low	5
[18]	Cognitive architecture with attention and eye movements	A visual attention model (EMMA) as part of a cognitive architecture (ACT-R-PM) demonstrated on equation solving		To capture important aspects of behaviour	Yes	High	3
[10]	Motion pop out	Object detection and recognition	Colour sub-band features plus context based object presence (saliency map)	To test a sort of top-down process	Yes	High/low	8
[19]	Maximum entropy model (MEM) with image seq., audio and text	Baseball highlight detection	Colour, edge distribution and estimated camera motion + audio, text	To detect and classify important events in baseball	No	Low (high level text)	3
[20]	Bottom-up saliency calculation	Robot guidance	Features plus motivational bias	To test theory in robot	No	Low	4
[21]	SVM-classifier	Video classification	3D - Saliency volume, K-Means, feature vector, colour histograms, entropy, inertia, energy and homogeneity	To classify soccer, baseball, swimming, boxing and snooker	No	Low	8
[9]	Feature saliency	Compression	Centre-surround on colour contrast, temporal flicker, intensity contrast, four orientations and motion energies	To build saliency map and also foveation areas	Yes	Low	9
[13]	Centre surround to conspicuity maps	Validation of saliency map technique	Intensity, two chromatic features, four local orientations	To validate saliency maps against mean saccade maps and to build a testing framework	Yes	Low	8
[1]	Saccade/random vs. saliency	Model and data comparison	Centre-surround on colour contrast, temporal flicker, intensity contrast, four orientations and motion energies	To validate the bottom-up influence on allocation of visual attention	Yes	low	10

Table 2.1: Relevant papers on modelling of visual attention.

2.2 Visual attention

2.2.1 Attention as selection

An definition that further clarifies the workings of attention is given by[22]:

> This process of selecting and disregarding environmental stimuli for further processing is called attention.

The process of listening only to relevant voices in a crowd is captured and it fits well with the view of visual attention as a spotlight. Some stress the serial nature of visual attention. The theory is that there is a single processing resource with limited capacity available, which can change what to process serially. This has been demonstrated[23] with performance measures on dual tasks that show that when one task performance increases the other decreases, conforming to the hypotheses that a single resource is available. Similar results have been found[24] by measuring resource allocation in multiple object tracking. However, it has been claimed[25] that there are two independent belonging to each hemisphere and by pointing to fMRI(functional Magnetic Resonance Imaging) data from subjects fixating an area with task irrelevant sequence of digits, with relevant sequences of digits and letters to the left and right. The task was to report matching digits and letters to the left and right. The fMRI data indicate that corresponding retinoscopic areas to, left and right eye, are activated during task engagement. Evidence has been provided[26] that more resources are utilised in dual tasks. With a task of pointing towards locations bi-manually (with both left and right hands), simultaneously reporting letters either at the same position or the other. By comparing with a uni-manual condition they show that more resources are deployed in the bi-manual task.

Visual attention has been portrayed as a multilevel selection process[27], distinguishing four different levels from functional brain imaging research. The first one is located to an area called lateral geniculate nucleus, the second areas are V4 and TEO , the third areas are in frontal and parietal cortex, and the fourth in the pulvinar of thalamus. Such a position allows the possibility that serial and parallel processes cooperate, which I find credible.

Although attention is related to awareness, attention and visual awareness can be separated. With psychological/theoretical and neurological arguments, Lamme[28] first

notes that items, that are visually attended, results in either unconscious or concious experience, and also claim that conscious data can be attended or unattended at a later stage, resulting in phenomenal awareness as or access awareness respectively. Phenomenal experience is short-lived, vulnerable and not easily reported, whereas access awareness is more durable. Neural correlates are further supporting the theory.

2.2.2 Visual search

In a visual search paradigm a slightly different definition of attention has been given by[29] as

> attention is a set of strategies that attempts to reduce the computational cost of the search processes inherent in visual perception

Visual search involves several different levels of visual processing. Early in the visual cortex parallel activation is processed, resulting in for example the pop-out effect, where a part of the stimuli protrudes visually due to the visual attention mechanism. Further, covert attention can focus on parts of visual input and specific areas are foveated.

In this area of visual search, interesting evidence can be found for top–down modulation of low–level saliency. Eye-movements are guided to areas with orientation and spatial frequencies close to the target[30]. Evidence has been found that semantic priming can affect visual search for an odd-one-out target[31], thus proposedly higher–level processes interacts with low–level saliency calculation. Also, in the multiple target paradigm it has been shown that suppression of distractor targets is intelligent in the sense that only confusable distractors are suppressed[32]. Thus task based conspicuity involves top–down information.

A recent study on how visual working memory interacts with visual search supports that the influence is under strategic control[33]. The study investigate attentional responses to distractors identical to the content of an object visual working memory, and find that the covert attention instead directed away from the distractors in the experiment. Even in the mechanism of pop-out, top-down processing has been proposed[34], shown with studies of event related potentials (ERP) on Monkeys, where the FEFs are which are involved in top-down processing has activation first. They conclude that temporal cascade of selective activity is similar for both efficient and inefficient search tasks. In a study[35] visual

working memory content is varied, testing attentional selection, with results supporting the view that top-down processes is involved in pop-out.

Detecting objects in cluttered scenes is a difficult task. Work has integrated top-down and bottom-up frameworks for saliency calculation[36, 37], where top-down saliency is defined as the weighted sum of features that are salient for a particular object being looked for. Another method uses statistics of natural images, feature target resemblance and prior knowledge of where the target is likely to be[38].

Research in artificial intelligence has been directed towards the problem of detection of objects in cluttered scenes[39]. It is a difficult task that humans do well in a two stage model where the entire visual field is processed in parallel in the early visual system and serially as regions of interest selected by the attentional spotlight. It has been argued that this is a strategy to overcome the limitation of purely feed forward processing in the presence of clutter and crowding. A computer implementation[40] replicates phenomena such as pop-out, multiplicative modulation and change in contrast response, emerging naturally as a property of the network. Another psychological model takes cognitive factors into account[41].

For visual search, a model has been implemented[42] that is much faster that previous models (SWIFT[43], HMAX[44]) without loosing performance. It is used for top–down guided search. The algorithm learns 42 separate Gaussian distributions corresponding to each feature and is able to output the probability of an object being located at a particular position. Following the maximization of gain train of thought Bayesian models have been proposed as of how humans solve the visual search problem of deciding whether or not a target is present in a scene or not[45].

2.2.3 Covert and overt attention

The attention mechanism can be divided into covert and overt attention. It is possible that the eye is focused on one spot, whereas another point in the periphery is actually attended. Overt attention is where the eye is currently focusing, whereas covert attention is where the brain is focusing. The highest resolution is obtained in the overtly attended area. There is also psychophysical evidence that visual attention can actively enhance the resolution. In a study[46] performance was tested in covert attended and unattended condition with

stimuli designed to measure spatial acuity (e.g. a square with a small gap on one side). There is also evidence that contrast[47] is enhanced where exogenously attended areas are described as having more contrast when they in fact have the same as non-attended ones. It has been claimed[48] that covert attention to peripheral cues can even decrease acuity in unattended locations . Such differences between attended and unattended areas in the visual field points to early economization of resources.

The pre-motor theory of attention states that covert or spatial attention is equivalent to planning but not executing a saccade[49, 50, 51]. Evidence for this includes the coupling of covert attention and saccade preparation observations that neurons in sensorimotor structures such as frontal eye fields (FEF), that the trajectories of saccades can be influenced by the allocation of attention, and that electrical stimulation of FEF and superior colliculus can influence the allocation of attention. Further it has been found that neural activity in the lateral intraperietal area has been associated with attention to a location in visual space, and with the intention to make saccadic eye movements[52], a similarity also found in other areas[53, 54]. It has been claimed[55] that information about upcoming movement mediates this shift of attention.

Recent evidence suggest that covert and overt attention are more decoupled[56, 57]. As a result it is possible to covertly attend an area without executing a saccade. Neurophysiological studies with monkeys have shown that covert attention can be directed to a particular salient area for an extended period without causing eye movement and even inhibit saccades[58]. One explanation is that all considered fixation points are attended in parallel prior to saccade execution[59]. It has been argued[60] that this can only happen during endogenous covert shifts of attention. In has been claimed[61] that no covert attention is required for overt attention but only pre-attentive parallel processing.

Research indicates that that there is a common mechanism for covert and overt attention as far as the selection process goes[62]. It has also been claimed[62] that a functional difference between covert and overt attention exists in that attentional modulation is separate from saccade programming.

The interaction of overt and covert attention is particularly important for models concerned with visual search[12]. A model of this interaction is also necessary for the understanding of mechanisms like saccadic suppression, dynamic remapping of the saliency map

and inhibition of return, covert pre-selection of targets for overt saccades, and the on-line understanding of complex visual scenes. Evidence[63] supports that a parallel mechanism is involved in visual search not necessarily involving relocation of attention.

How exactly covert attention operates is quite unknown. It has been shown that attention can be split between two targets[64]. Some findings indicate that the divided spotlight is actually a rapid temporal switching[65]. Also that selective attention need not be mediated by spatial attention[66], since you can see a sensory element without mediation by spatial attention. Except for the spacial extent of attentional processing, space–time maps of both endogenous and exogenous visual attention has been outlined[67]. The theories of covert attention are disparate and further research needs to be done to draw definite conclusions.

2.2.4 Bottom-up and top-down processing

Two distinguishable processes are involved in attentional allocation, bottom-up (image-based) and top-down (task-based) processes. The difference between these two kinds of processes is the origin of action/activity or in this case eye-movement. In bottom-up processing the origin is stimuli. For example simple features like contrast, corners and crosses attract attention. In top-down processing the origin is high-level cognitive processes. A clear cut example is when people are asked for example "Where is the red car?", which leads people to direct the attention to a red car if present. It has been demonstrated[68] that bottom-up information is more important in unknown images whereas in specific kinds like web sites top–down is more important.

Studies indicate a fundamentally different visual and abstract information processing [69]. For example dissociation between automatic and controlled processes have been demonstrated[70] and enhanced methodology for this[71] has been developed. As further evidence for distinct neural mechanism in endogenous and exogenous attention is the dependence on shift time on distance between attentional points[72]. A computational model based on experimental results[73] has been developed that proposes that stimulus–driven allocation of attention exists early at appearance of attractor, but is later modulated by top–down signals. Similarly in research on visual search it has been shown that, although bottom–up processes initially control attention, top–down processes defined as accumulat-

ing scene knowledge quickly take over and dominate search[74].

For bottom-up processing there is neurological evidence that some sort of saliency map is calculated[75]. This saliency map encodes how important particular areas are and is hypothesised to be the primary guide for the attention mechanism. Several computational models have been implemented to simulate such a process[5, 12], where features like colour contrast, intensity contrast and orientation are used to build the saliency map. These maps have been validated with eye-tracking data[13, 5], which provide evidence that bottom-up processes contribute significantly to the selection of fixation points. Additional experiments indicate that selecting interesting objects in a scene is largely constrained by low-level visual features[76]. It has been claimed[77] that salience map models contributes significantly, but accounts only for a small amount of the variance in where people fixate, especially pointing to that scan-path sequences are not predicted by a saliency map. Research on stimuli manipulation confirms that visual saliency is a poor predictor of real observer scan-paths[1], and in fact objects are better candidates for fixation points[78]. Some research though indicate that top–down search strategies cannot override reflexive attentional capture[79].

The pure bottom-up and top-down processes are pretty simplistic as described above. Bottom-up processing requires only visual input for its functionality whereas top-down processing needs a cognitive model of the surroundings (i.e. where to look for the red car), and knowledge about the world (what is a red car?). An important question is how these processes interact. Evidence points to that these compete for focal attention[80] and one possibility to unite top-down and bottom-up is the Unitary Saliency Map Model where both types of processes feed into the same saliency map[81]. It has been proposed that integration of multiple sources, i.e. sensory, motor and cognitive variables, is done in the lateral intraparietal area (LIP)[82].

It has been proposed[83] that there are three main sources of guidance information available when watching a new image: low-level saliency, target template information and scene context. The saliency being purely bottom-up, a target template matches features and attention is directed according to match. Further, scene context indicates where to look for particular targets. Supporting the template target information theory subliminally primed targets attract attention[84] better than semantic priming. Furthermore, occluded object parts attract attention[85]. With respect to representation of visual categories some

evidence points towards an example based representation working as a template guiding search[86]. Evidence has been found on that contextual conspicuity and physical presence are governed by distinct neural mechanisms[87].

A study[88] indicate that purely top-down processes provides a much closer match to human behaviour than a mixture model using bottom-up information. In the study bottom-up saliency as well as a feature template match to a stored representation is used to predict eye–movements in visual search. In another study[89] involving a hand pointing task it has been found that initial saccades are directed towards saliency as defined by low–level features however the subsequent towards the target. Although attention in the everyday sense evokes conscious mediation of the stimuli, advanced aspects of attention are dissociable from awareness[90].

It has been demonstrated that saccades are influenced by visual working memory and these are thus controlled at least partly top–down[91]. Interesting studies have been done on how spatial working memory operates to control planned sequences of eye-movements, possibly clarifying aspects of automatic attention models[92, 93]. Another study has modelled the way implicit knowledge affect eye-movements[94].

Not only cold cognitions (information based) operate attention, but also hot ones like emotions. Attention move rapidly towards threat[95] consistent with theories of emotion. Further, reward and attentional systems are interdependent[96, 97]. Here reward-associated stimuli are preferentially processed over other valanced stimuli[98]. It has been shown that reward information is readily integrated with saliency in the sense that it affect target selection and exact landing position of a saccade, but this is a time-consuming process[99].

It has, furthermore, been shown that eye movements such as saccades and vergence are guided by the perceived stimulus and during fixation by the physical stimulus. Thus dissociation between perception and eye-movement during fixation has been demonstrated[100].

2.2.5 Attention to objects

Some have argued that attention to objects is orthogonal to saliency driven attention since the visual stimuli of objects, which is not salient, are still attracting attention[101]. Further, there seems to be both bottom-up and top-down object processing since recent results

suggest that object-based attentional capture guide both types of attentional orienting [102]. Some make claims that saliency mainly acts through objects in as it is the objects that attract attention[103].

To direct attention towards object they must be detected. Grouping is then an important step[104]. The ground-breaking work, by the Gestalt perception work done (e.g. [105]), postulate that there are a number of different principles of perceptual organisation that bind features in stimuli together, and continue to inspire researchers today. Some of the principles are closure (filling in missing parts), similarity, proximity, symmetry and common fate (two object share same motion endpoint). For example texton-based segregation can be used to find object boundaries[106], which is based on feature similarity. Another method is edge extraction and interpretation, which build on all of the principles except common fate. Possibly objects are also recognised and thus attended to because of its importance as object type or instance. A question is if object recognition is done in parallel, with all visible object attended at the same time, or serially, with one object being attended at a time. Recent research rejects both serial processing, and unrestricted parallel processing, as the best model of object recognition[107] and proposes a parallel model with restricted capacity.

It has been shown that higher level properties such as animacy and goal-directed behaviour improve higher–level classification of behaviour as opposed to random movements, a spatio–temporal property assigned to the particular object being looked for[108]. Thus not only spatial properties of objects are important in visual attention. Similarly object–tracking predictability matters for multiple object tracking in as so far as that objects moving in a predictive way are more easily tracked[109]. Finally, some evidence point to that attentional object tracking is carried out independently by each brain hemisphere[110].

2.2.6 The “where” and “what” streams

There are proposedly two separate neurological information processing streams in the visual attention mechanism, the where and what streams. The existence of these have been investigated in a behavioural study[111], where the internal object representation has been separated into a “motor” and a “sensory” memory. Each of these chains reflect an alternating sequence of elementary motor and sensory signals which are expected to arrive in

response to each action. These are used in subconscious behavioural recognition when the object is known. The matching of incoming sensory stimuli is compared with the expected from executed motor commands.

Also, neurological experiments have found two major low-level bottom-up streams in the visual system[112]. Information from the retino-geniculo-striate pathways enters the visual cortex through primary visual cortex V1 in the occipital lope and proceeds into two separated streams. The first one leads through extrastriate visual cortical areas V2, V4 to IT (the inferior temporal cortex), and is mainly concerned with object recognition. The second one leads through PP (posterior parietal complex) and is responsible for maintaining a spatial map of an object's location and the spatial relationships between object parts as well as the spatial allocation of attention. Neurological and behavioural findings represent two separate lines of research that have produced convergent results on visual attention and scene understanding. Thus this model has a sound scientific base and might be considered for computer implementations of visual attention.

2.2.7 Fixations on faces

Faces are of importance to humans and thus extraordinary resources are utilised to find, scan and remember faces[113]. When it comes to visual attention in the past it has been found that the human visual system can detect animals in a complex natural scene very fast (120–130ms before saccade), and new research has found that saccades towards faces are even faster[114].

On still images people tend to first look at a centre-point of the face and then successively scan features like the eyes, the nose and the mouth. Further, a study[115] show that involuntary attraction of visual attention to faces is stronger than other visual objects. Many studies have been on frontal faces but people look at profile faces too[116]. Recent findings[117] show that no preference for eyes exists for dynamic faces. Rather, attention is adjusted to dynamics, i.e. speech induces attention to the mouth area, face looking directly into camera induces attention to eyes, looking at another person to the other person and finally during face movement to the nose, which is attributed to optimal tracking of the face as a moving object.

2.2.8 Attention and scene understanding

To understand the more complicated aspects of the visual attention the processes of scene understanding, learning, expectations, competition and consciousness must be considered. Visual attention not only results in scene understanding but is an integrated part of it as has been shown[118] that entire stimulus or objects can be selected as a whole by the attention mechanism, including all its features. For example, the picture of a target is a good cue in a search task[119]. Some more elaborate theories are here described that try to describe the attentional system from a functional point of view, thereby linking bottom-up and top-down processes.

One of the earliest models is MORSEL (Multiple Object Recognition and attentional SELection[120, 121]). This model is applied to the recognition of words processed through a recognition hierarchy. Without attentional selection, the representations of several words in a scene would conflict and confuse that recognition hierarchy, yielding multiple superimposed representations at top level. The addition of a top-down attentional selection process allowed the model to disambiguate recognition by focusing on one word at a time.

Another theory[122], supported by recent experiments[123] states that the next fixation point is chosen as to maximize the gain of information about the object currently investigated. This means discrimination is permitted between current candidate object classes in a hierarchical internal tree of objects and object classes.

A model has been proposed[124] related to the where and what streams which relate to different neurological pathways in the brain. This theory states that particular scan paths are learned during the lifespan of a human for each particular object or scene to be recognized. In a particular scene a person chooses between scan paths that are particularly useful for the understanding of that scene and objects in that scene. The result of following a scan path in the "where" memory is compared to object appearance templates in the "what" memory, and thus object recognition and scene understanding is possible. The two different approaches that has lead to converging results gives broad scientific support to the theory of the "what" and "where" memory.

Another proposed model[112] combines bottom-up selection of object locations and object recognition. Here object locations are found in a bottom-up manner at a coarse scale. Candidate locations are scanned serially at progressively finer scales until object

recognition is completed. This model has been supported by psychophysical studies that show attentional enhancement of spatial resolution as mentioned earlier.

Finally, a theory that emphasize top-down processes has been proposed[125, 126]. According to this theory what we see is only vaguely related to what is received at a retinal level, which is supported by the fact that people experience a vivid perception of the full view, whereas only the attended part is actually clear. This is called the "scan path" theory and states that a cognitive model of the surroundings is the main basis of selection of focal points. Here the attentional system is more of an adjusting mechanism to be able to cope with the details of the environment that adjusts according to the task.

2.3 Eye movements

There are several different types of eye–movements. These are[127] *saccade*: fast voluntary jump-like movements, *vestibular–ocular reflex*: stabilizes visual image on retina as head moves, *nystagmus*: resetting of compensatory movements, *optokinetic nystagmus*: stabilizes gaze during low–frequency rotations at a constant rate, *smooth pursuit*: voluntary tracking of moving stimuli, *vergence*: coordinated movements of both eyes to account for divergence and *torsion*: coordinated rotation of eyes around optical axis, dependent on head tilt and eye elevation. In the following subsections fast eye movements (like saccades), smooth-pursuit, inhibition of return as well as application areas of eye-movement models are covered.

2.3.1 Fast eye movements

Shifts of attention are very rapid and are called saccades. These takes in the order of $100 - 300ms$ to plan and execute[128], and are the fastest movements produced by the human body with speed up to almost $1000°/s$. Most of the time the eyes are directed to points in the surrounding that are important for the current task a human is engaged in. Saccades often land at the middle point of targets[129]. There are several types of saccades: reflexive saccades, memory guided saccades, antisaccades and catch–up saccades[130]. Characteristically saccades do not follow motion of objects. In target selection, at least for saccadic eye movements the superiour colluculus in primates play a role[131]. It has been shown[132]

that two saccades can be programmed simultaneously which can lead to a very short inter–saccadic interval.

Recent findings indicate that several brain mechanisms can be involved to a varying degree[133], as opposed to previously assumed either indirect versus direct control. Saccades can be abrupted and continued to another location as a result of race[134]. Also, both common and differentiated activation can be traced in comparing saccades and vergent eye movements[135]. Further, some have found a difference between within object saccades and between object saccades[136]. Here it is claimed that they operate in different coordinate systems (retino–centric and oculo–centric).

A distinction is made[55] between reflexive and volitional saccades. The theory is that reflexive saccades are triggered by peripheral stimuli automatically and is often faster with latencies in the order of $180ms$. Volitional saccades are the effort of intention to locate towards the target, typically measured by a target that does not trigger reflexive saccades. Latencies for volitional saccades are in the order of $250ms$. Evidence has been found that reflexive and voluntary saccades are programmed in parallel[137]. However, apart from targets, it has been shown that linguistic cues can induce involuntary programming of eye movements[138]. This puts the proposed distinction between reflexive and voluntary programming of saccades in question in that higher–level concepts are involved in the involuntary programming. Contrary to the belief that the fastest saccades are reflexive both verbal and visual information in working memory has been shown affect these[139]. So the subject area might need more detailed philosophical analysis.

A sequence of saccades is often taken as to have an intrinsic order, called scan–path. Previous fixation point is an important predictor of subsequent saccade, including both target selection and fixation duration[140]. Recorded scan-paths on artificial displays with arranged targets have shown evidence for scanning strategies based on both directional orientation (raster–like) and global external contour[141]. A simple search strategy where the scene is scanned in a coarse-to-fine manner has been proposed[142]. Here it was found that mean saccade amplitude decreased and mean fixation length increased as a function of the ordinal saccade and fixation number. Although over the years it has been assumed that covert shifts of attention mediates serial scanning newer research points to that it could be limited capacity parallel processing[143].

Small fast movements are called microsaccades, which are not as investigated, but refer to shorter fixational changes. Competing motor plans generate microsaccades and saccades. Some[144] seem to argue that microsaccades are produced with great influence from noise (activation spread to nearby areas). Others have noted that microsaccades not only counteract perceptual filling in, but also maintain figure–ground separation[145]. Yet others[146] claim that there is no evidence that microsaccades serve as a necessary role in improving oculomotor control or in keeping the world visible.

2.3.2 Smooth pursuit

Smooth pursuit is primarily driven by visual motion[147]. A qualitative difference has been shown between smooth pursuit and fixation as simply steady smooth pursuit. Changes in visual feedback have little effect when subjects fixate a stationary target, but the same changes produce large oscillations in eye velocity when the subject tracks a moving target. It has also been found that object recognition performance is lower during smooth pursuit than fixation[148].

Past research has empathised the automatic character of smooth pursuit. For example the latencies for smooth pursuit is shorter (100-125ms) than for saccades (200-250ms). However more recent findings show that pathways, such as the basal ganglia, the superior colliculus, and the nuclei in the brain stem reticular formation, suggesting that smooth pursuit has a similar functional architecture to that of the saccadic system, being controlled more volitionary than previously believed. Humans also have the capability of predictive smooth pursuit, i.e. following an occluded target. Predictive smooth pursuit is driven by an internal representation of target motion that evolves with time[149]. However, it has been pointed to that[150] pursuit initiation is driven by retinal image motion signals, not yet processed for figure completion.

The spatial location and extent of visual attention during smooth pursuit has been tested by[151], who found that attention is biased just in front of the pursuit stimulus (about 1° ahead) extending an angle of about 6°, by measuring the response time to peripheral targets. It moves away from the pursuit stimulus as target velocity increases. Others claim[152] that there is no appreciable lead or lag, but showing that smooth pursuit of a translating string does not improve attention with a lead or lag. The different result

could possibly be explained by that higher-level processes must be involved to recognise characters, and in such case there would be an extra sensitivity to low-level events in front of the followed objects, but not resources to recognise characters. Another study[153] shows that visual short term memory is impaired for the position of peripheral objects in comparison to when fixating an object.

Older neurophysiological evidence points to difference between neuronal pathways for pursuit and saccades. However, more recent findings indicate that the neuronal pathways are not independent in some structures in the brain[154]. Based on the finding a new model where pursuit and saccades are coordinated is presented where it is proposed that covert attention is engaged to plan saccade to and pursuit of a new target.

2.3.3 Inhibition of return

Inhibitory mechanisms play an important role in cognitive processes. When confronted with an environment that contains hundreds of objects our thoughts and actions are directed to only a few of these. This volitional mechanism thus inhibits processing of irrelevant stimuli or objects.

Studies over the years have shown that there exists a low-level inhibitory mechanism, inhibition of return (or IOR)[155]. This is a bottom-up mechanism that simply prevents the human visual system to attend one point several times, and it is proposed that this provides an efficient low-level strategy to scan the environment. By inhibiting return to the same point it is ensured that several interest points in the environments are attended instead of only one. Here no volitional thought is necessary and the selection procedure is thus extremely fast. This mechanism has been used in a object recognition framework[11] where a visual attention algorithm is implemented as a way to highlight interesting objects or object groups. In a review it was found, from an evaluation of results obtained in research on visual search, that IOR lasts for at least $1000ms$ or about four previous inspected items[156].

It is debated whether the IOR effect is predominantly a perceptual or a motor process [155]. Some theorists argues that it is solely a perceptual process and that perceptual processing is inhibited at the previously attended location. Others believe that it is solely a motor process, and that motor responses toward the previously attended location are

suppressed and thus delayed. Yet others claim that it is a mixed process[157]. The IOR mechanism has often been studied in cued/target perceptual experiments, where the target is cued before presented. Measurements of the time it takes for a saccade to reach the target depends on the cue. For example if a cue is placed in the target position, the IOR effect ensures that a saccade toward the target is delayed (e.g. [155]). In that particular experiment the authors managed to contrast the two explanations in a single behavioural task. Their conclusions where that the IOR is predominantly a perceptual response. The effect could possibly be simply due to the data being maintained in visual working memory and thus not needing update[158]. This supports a view where motoric inhibition is not involved in inhibition of return. Other recent findings[159] also indicate a sensory component of inhibition of return, i.e sensory data is adjusted.

2.3.4 Automatic eye movements

Most models of automatic eye-movements utilise saliency maps, either with bottom-up components only or with top-down influences on bottom-up saliency[75, 160]. Considering the complicated mechanisms in cognition it is questionable if real task related top-down processing is replicated since higher level cognitive processes are traceable in saccadic patterns[161]. Considering the limited capacity of computers from a computational point of view, the bottom-up saliency with high-level influences is a reasonable approximation. However top-down information, including non-image based, needs to be integrated in the process of selecting focal points. Early such models include adding object recognition as input to an interest map[162].

An early active vision system[163] uses iconic scene descriptions to guide attention to targets. An example of an overt visual attention mechanism based on saliency dynamics has been presented[164]. Here a robot is to localize and position the number of relevant coloured objects in the environment, and this is done by continuous scanning of the scene incorporating objects as fixation point candidates, saliency, IOR and winner take it all as final selection. Objects are tracked by having an object memory that is projected onto the scene and matched with camera input. Pioneer fixation points are introduced periodically as a systematic sweep. If an object is considered relevant it is inserted as a new fixation point.

An application has been presented[11], that uses more of the components of the saliency model. In their work, a saliency model is used to detect and segment objects, for learning and recognition of events in a cluttered background. The saliency map is used to find regions of interest at different spatial locations in the input image. Thereafter the object boundaries are found by tracing back which feature contributed most to the saliency at that particular point and segmenting by this feature. The idea is that if the salient point is on a red book, then it is best to segment the book in the feature map for red. This model uses unsupervised learning and is able to learn and recognize objects and groups of objects (for example a pile of books).

Interacting with the internal world a social robot has been implemented where attention is directed on bottom-up saliency calculations modulated by motivational factors[165]. Considering search as an essential feature of attention, a clear cut example of a robot realisation of the search in 3D space for an object has been presented[166]. Finally, an interesting model on oculomotor dynamics during smooth pursuit involving 1D and 2D motion cues as input to a Bayesian model has been implemented[167].

2.3.5 Other applications of a model of eye movements

The goal is not only to model visual attention but also to find application areas of visual attention not only considering it as a robot module. The application areas are plenty, e.g. video compression, video shot classification and object detection. Other areas of computer vision like object recognition would benefit to be studied from a visual attention perspective.

One area of application is video compression. Since only 2° of our 140° view provides a clear input image, only that area around where people are actually looking needs to be transmitted with a high bitrate, whereas the rest only needs a fraction of the bandwidth. A model[168] implements a predictive encoding of video based on where the observer has been looking at in previous frames. The user is equipped with an eye-tracker and the eye-tracker information is continuously sent to the encoder. The encoder then predicts the eye-gaze at to be sent frames, and thus reduces bandwidth. In this system simple motion prediction is used with a Gaussian visual window. The result is significant perceptual gain with limited bandwidth. The importance of localising faces in a saliency model for video

compression is obvious in telecommunication applications.

Another predictive coding scheme has been presented[9]. Here an entire video sequence is encoded by predicting visual attention with a saliency model. Two different models are tested. One where the original saliency map directly is used to determine where people with a certain likelihood will look, and another where circular areas follow the most salient area with a spring based dynamical model. The less salient areas are blurred according to the predicted likelihood. The sequences are later compressed with MPEG-1 or MPEG-4, resulting in less information encoded for the blurred or less salient areas. Significant compression ratios down to 11% are obtained, however with different perceptual quality. Such a scheme has been used[169] to control quality/bitrate across a single frame in real–time computer animations. Another model has been implemented[170], where higher–level content is used to determine the way the features are added up in a saliency map.

Top-down bias, i.e. modified bottom-up components by top-down information, has been implemented in an object detection framework[10], where contextual information is used to modulate the saliency map. A statistical model is here manually trained enabling prediction of the locations of people in different images. The result is a probability density function (*pdf*) that determines the likelihood of a person in a particular location. Eye-tracking data validates that the addition of context as a top-down process in attention, here biasing the saliency map, produces better predictions of actual human saccades in images.

Another system that uses top-down processes as modulation of the saliency maps has been implemented[20]. Here bias coefficients are learned for each object that the algorithm tries to find. In this way the actual features that are presented are favoured in the calculation of the saliency map. This is successfully used in a robot navigation framework. A saliency model with a winner-take-it-all choice mechanism is used to navigate the robot toward the selected object. Experiments show that the robot attention modelling with the top-down effects included it is very successful. A fast implementation of visual attention based on feature saliency for humanoid robots has been presented[171].

A clear cut video classification system that uses a saliency-based model has been implemented[21]. Here a saliency calculation scheme is developed that works in the spatio-temporal domain. Interesting or salient events are extracted and used to train simple

SVM classifiers[172] to discriminate between soccer, swimming, basket, boxing and snooker videos. The model is able to discriminate between the different sports with very high accuracy. This indicates that saliency models has the potential to highlight characteristic spatial or temporal events in a video sequence, facilitating information extraction or classification.

Attention modelling can be used in 3D rendering. With the goal to render virtual environment, task related attentional factors have been studied with respect to interaction of notice of degradation[173], with perceptual quality in the sense that regions of interest are given more processing power. Further, top–down selective visual attention has also been used to improve SLAM[174] and bottom up visual attention has been used to segment active contours[175].

Visual attention can be used to extract semantics from media. For example image retrieval by semantics using region saliency has been done[176]. Saliency has also been used for video event detection and summarization[177]. Further, selective visual attention has been used in pattern recognition[178], here specifically handwritten numerals. Feature based attention has been used for moving object segmentation[179]. Spatio-temporal saliency has also been used in a background subtraction task showing good performance on difficult stimuli[180].

Regarding medical application, saccade trajectories deviations can reveal a lot about psychological processes[181]. Thus it might be possible to use saliency models in psychiatry, to distinguish persons with Attention Deficit Hyperactivity Disorder (ADHD), Fetal Alcohol Spectrum Disorder (FASD) and Parkinson's Disease (PD) from other persons by comparing correlation between salience and gaze[182].

2.4 Saliency

Most relevant research in the area is concerned with the bottom-up processing of features, which basically has been done by calculating saliency maps that encode the relative importance of different areas of visual stimuli. Others have included top-down information in the saliency map calculations, for example context[10]. However, it has been pointed out that saliency affect attention even when saliency is task irrelevant[183]. There is neurological evidence that guides the construction of such models, and most of this research has been fo-

cused on the processing of still images. The impact of different type of features varies[184], for example there is evidence that intensity contrast is a very important feature, and especially junctions formed by such contrasts (i.e corners and crosses). Others have shown that the relative weighting of these features varies in a context dependent matter on task and training[118]. Additionally, there seems to be little evidence for considerable interaction between different modalities (e.g. [185]), including motion processing[186]. Finally, feature contrast, as spatial difference of feature values, seems to be of most importance and not the features themselves, i.e. interesting areas are where something happens in the spatial plane.

Research have also been done of visual attention on video sequences[1, 187], but to a lesser extent. The most important finding here is that motion and temporal change are stronger predictors of human saccades than features like colours, intensity or orientation features, (i.e gradient along particular directions). It has been found that a majority of the saccades are directed toward a minority of the salient locations. This further filtering of saliency is necessary, either by recalculating or fusing with additional information. An example of such information could be new motion, i.e. objects that start to move, new objects appearing and the combination, since they attract attention better than motion and objects themselves. According to one study[188] new objects seem to attract attention the most. In dynamic scenes the relevance of the bottom-up generated saliency map looses validity[189]. Further, it has been shown that disparity information changes basic eye movement properties and that subjects tend to first fixate closer locations and later more distant[190], however this has not been extensively studied since most studies has been on 2D image displays. Finally, biological motion attracts attention[191] more than other motion.

Many model utilize saliency as the only input to attentional selection. Assuming that other processes do not contribute to a global saliency map there is a severe limitation. It has been shown that saliency driven attention only affects visual selection shortly from the onset of a visual scene[192, 193]. Algorithmic approaches have been used to speed up saliency calculations while preserving performance[194].

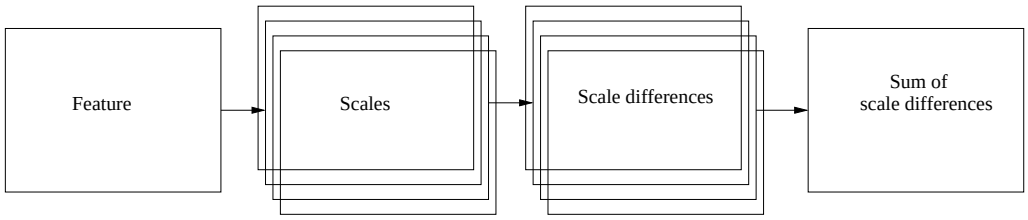


Figure 2.1: The calculation of centre surround differences as sum of scale differences in the model[11].

2.4.1 Computational models: Saliency maps

Several different types of computational models have been constructed to calculate saliency maps. In one of the most common types presented[5], features, e.g. colours, intensity and orientation are extracted. Centre-surround differences[5, 195] are calculated, to account for the fact that feature contrast is important and not the features themselves. These feature contrast maps are then normalized and linearly combined to create a final saliency map.

One such model[11], is a straightforward implementation that can stand as a typical example of saliency models. Here the input image I is sub sampled into a Gaussian pyramid, and each pyramid level σ_p is decomposed into channels for red(R), green(G), blue(B), yellow(Y), intensity(I) and local orientation(O_θ). If r, g and b are the red, green and blue colours of a image, normalized by the intensity(I), then $R = r - (g + b)/2$, $G = g - (r + b)/2$, $B = b - (r + g)/2$ and $Y = r + g - 2(|r - g| + b)$. Local orientations are obtained by applying steerable filters to the images in the intensity pyramid I. From here centre-surround maps are generated as illustrated in Fig. 2.1 with the following formulas:

$$F_{I,c,s} = N(|I(c) \ominus I(s)|), \tag{2.1}$$

$$F_{RG,c,s} = N(|R(c) - G(c)) \ominus (R(s) - G(s))|), \tag{2.2}$$

$$F_{BY,c,s} = N(|B(c) - Y(c)) \ominus (B(s) - Y(s))|), \tag{2.3}$$

$$F_{\theta,c,s} = N(|O_\theta(c) \ominus O_\theta(s))|), \tag{2.4}$$

where $\ominus$ denotes across-scale difference between two maps at the centre (c) and the surround (s) levels of the respective feature pyramids. $N(\cdot)$ is an iterative non-linear normalisation operator. These feature maps are further summed

$$F_l = N(\oplus_{c=2}^{4} \oplus_{s=c+3}^{c+4} F_{l,c,s}) \quad \forall l \in L_I \cup L_C \cup L_\theta \tag{2.5}$$

with $\oplus$ denoting summation that is done across scale and

$$L_I = \{I\}, \quad L_C = \{RG, BY\}, \quad L_\theta = \{0°, 45°, \quad 90°, \quad 135°\}. \tag{2.6}$$

The different colour and orientation channels are each summed and conspicuity maps are formed as

$$C_I = F_I, \quad C_C = N(\sum_{l \in L_C} F_l), \quad C_O = N(\sum_{l \in L_\theta} F_l) \tag{2.7}$$

Finally the saliency map is computed as a sum of the conspicuity maps

$$S = \frac{1}{3} \sum_{k \in \{I,C,O\}} C_k \tag{2.8}$$

A common way to model visual attention here is to use the winner-take it all model and make use of the Inhibition of Return (IOR) mechanism. This has for example been implemented[196] by temporarily attenuating a circular area around the most salient location. After that the next most salient location will be attended, and so on. In still images this only creates an effect which will enable focus on a number of points. In video sequences, however, this can produce a more advanced model where different important areas are attended as the sequence continues.

Several issues here are worth thorough investigation. The first one is to find out, which low-level features contribute most to a saliency map. Colour, intensity, orientations are used in several different computational models. Second, it would be of interest to find out which high-level cues are important. There is considerable evidence that the brain has a more or less separate system to process faces, and that attention is focused at faces to a considerable extent[197]. Therefore faces are beneficial to include in saliency calculation. Further, motion is also important. Would a change detector do, or is tracking

of moving objects necessary? Moreover, relevant objects like people should be tracked, since reasonably these are frequently attended. Finally, there is the issue of combination of conspicuity maps. In all articles I have read a linear combination is used. The question is first what the coefficients in a linear combinational model should be? Second, is there something better than a linear combinational model? Eye-tracking data should be used to optimize the combinational model. It has been shown that using only first order terms instead of including second order terms is significantly better[198]. Attempts have already been made to optimize the combination of features as well as receptive field sizes[199]. Finally, is the linear combination of features feasible as the only way to fuse different sort of information to determine allocation of attention. Given an intelligent human behind the choices of fixation points, although most decision visual attention orienting probably is unconscious, and not mediated by rational thought, but still high-level processes, it would seem not. For example a combination of features do or do not portray an object a judgement that is most likely not proportional to the content of feature contrast in the neighbourhood.

2.4.2 Saliency with top–level influence

There are limitations to bottom-up saliency models of visual attention as for example these predict fixational patterns poorly in relevant contexts like social scenes[200]. It has been shown that when top down–target information is available bottom-up information is discarded[201]. The addition of top-down information, i.e. prior knowledge, expectations and contextual guidance, has been investigated[202]. Top–down search can be done based on low–level features[203]. A number of different ways to integrate top–down information with bottom-up saliency has been tested.

An early model using non–linear relaxation was developed to integrate bottom–up and top–down cues[204]. In another model inhibitory top–down processes influence the bottom–up saliency[205]. Further, an attentional system where top–down task and context biases bottom-up saliency has been developed[206]. A modern model incorporates feature–based attention as an active top–down inference process where top-down activated features are enhanced[207]. Another model (SUN, Saliency Using Natural statistics) calculates top-down saliency based on natural statistics[208]. Here an object based representation is

looked for to generate the top-down term in the saliency calculations.

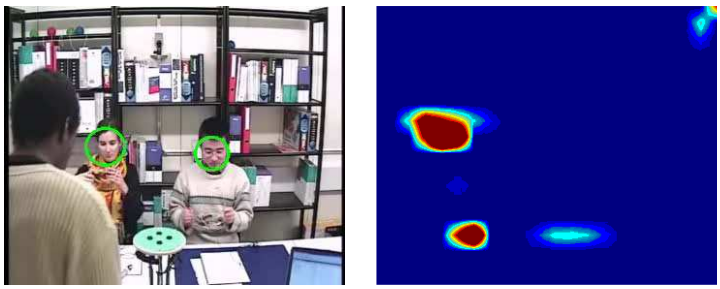
Another way to solve the problem with integration is to incorporate switching between top–down and bottom–up processes[209]. This is here a solution to the problem of top-down biased bottom-up saliency not providing enough discrimination to localize a target. Attempts to involve a visual working memory module in visual attention have been made[210]. A stochastic model based on saliency[1] uses previous eye-movements as further input. Studies have shown that scene context guide attention in that detection of scene-constrained targets are done faster and with fewer eye movements[211]. More initial saccades are directed towards target-consistent scene regions and more time is spent scanning those areas.

2.4.3 Validation of saliency models with eye-tracking data

Saliency models with high and low–level features have been validated with eye-tracking data, with slightly different techniques. In most studies some sort of experimental validation is applied[1, 12, 17, 13]. Quantitative measurements are needed not only to test a certain technique, but also to adjust the model and its mathematical parameters. One could use standard measurements like mean distance between predicted saccade locations and real saccade locations[10], but this is not appropriately testing the full value of a saliency map.

Another approach is to use a correlational approach[13]. Here the stimuli are still images and a Gaussian smoothed mean human attention map (see Fig. 2.2 for a saliency map and Fig. 2.3 for a Gaussian smoothed mean map) is calculated from saccade locations. A correlation score is calculated between the human attention map and the final saliency map. This technique has the advantage of being intuitive since one can easily compare the output of the human map and the saliency map, and a correlational score is easily interpreted. The correlational score works fine for still images since on every frame each individual is directing their attention to several points. By taking the mean of several people a human attention map is obtained which is similar in character to the saliency map itself.

In another method[1] fixations to random fixations are compared. Here calibrated eye movement data is segmented into saccades, eye-blinks, and fixation/smooth pursuit



(a) (b)

Figure 2.2: Example of saliency map (b) calculated from a video frame (a).

(a) (b)

Figure 2.3: Example of a Gaussian smoothed mean human attention map (b) calculated from eye-traces on a video frame (a).

periods. The following samples were taken into account for each video frame:

- S_h: Saliency at human eye position, computed as the maximum over a circular aperture of diameter 5.6° (9 pixels in the saliency map) of the model's dynamical saliency map sampled at the moment a saccade starts and around the location of the future endpoint of that saccade.
- S_r: Saliency at a random location, computed in exactly the same manner as S_h except that a random endpoint within the image (with uniform probability) is considered rather than a given human saccade endpoint.
- S_{max}: Maximum saliency over the entire frame, computed as the maximum over the spatial extent of the entire dynamical saliency map, at the same instant as the other measurements were taken.

The validity of the saliency model can then be tested with a one-way ANOVA (ANalysis Of Variance) as the difference between $S(h)/S_{max}$ and S_r/S_{max}. In the same study the correlation between saliency and saccade duration was also calculated. Interesting to note, no such correlation was found in the study.

An issue with most research concerned with eye–trajectories is that gaze patterns from free viewing of natural dynamic scenes differs from those obtained with still images or professionally cut material[212], and results can not be generalised to allocation of attention real life.

2.5 Low–level features

Common features used for generation of saliency maps are depth, colour contrast, audio and entropy. A summary of features used in different papers is presented in table. 2.2. Classical features are colour contrast and orientation.

Features	Explanation	Papers
Intensity	$(r+g+b)/3$	[213],[214]
Intensity contrast	centre-surround filtered $I=(r+g+b)/3$	[9],[1],[214],[215],[216]
Colour contrast	centre-surround filtered $R=\hat{r}-(\hat{g}+\hat{b})/2, G=\hat{g}-(\hat{r}+\hat{b})/2, B=\hat{b}-(\hat{r}+\hat{g})/2, Y=(\hat{r}+\hat{g})/2-\lvert\hat{r}-\hat{g}\rvert/2-\hat{b}$ where $\hat{r},\hat{g},\hat{b}$ are rgb normalised with I	[17] ,[9],[5],[216]
Chromatic features	Hue R - G and B - Y (Y = R + G / 2)	[213]
Orientation (Gabor)	Cosine gratings in 2D-Gaussian envelope	[10],[9],[213], [214],[1],[215]
Skin tone	Model from[217]	[170]
Motion	Change detection as deviation from model moving/non-moving region Gaussian modelling	[170]
Oriented motion energies	Shifted Gabors by one pixel	[9],[1]
Optical flow	V from $\nabla I(x,y,t)V(x,y,t)+\frac{\partial I(x,y,t)}{\partial t}=0$	[215]
Temporal flicker	Absolute difference In - In-1	[9],[1]
Entropy	$H(x)=-\sum_{i=1}^{n}p(x_i)log_b p(x_i)$	[21]
Inertia	$\sum_{i=0}\sum_{j=0}\lvert i-j\rvert^2 P(i,j)$	[21]
Energy	$\sum_{i=0}\sum_{j=0}P(i,j)$	[21]
Homogeneity	$\sum_{i=0}\sum_{j=0}\frac{P(i,j)}{1-\lvert i-j\rvert}$	[21]
Self information	$-\log(p(X))$	[218]
Spectral residual	$R(f)=L(f)-A(f)$	[219]
Self resemblance	likeliness of pixel to its surroundings	[220]
Bayesian surprise	$\int_{\mathcal{M}}P(M\lvert D)log\frac{P(M\lvert D)}{P(M)}dM$	[221]
Texture	directional high pass filters	[170]
Object context		[10]
Depth	Disparity contrast,Depth, 3D-Curvature	[17],[213]
Audio		[19]

Table 2.2: A table of features used in different papers that models visual attention.

2.5.1 Colour and intensity features

Colour and intensity are biologically supported features, for example the pop-out effect have been reported for colour and intensity[222]. These features have been suggested to model visual attention[75]. The features are readily available in the image as RGB channels or simple derivatives and can be rapidly processed[222]. They have also been extensively tested[1]. What is interesting is feature contrast and is here measured with centre-surround fields, thus producing of pop-out effect.

Further work has included utilizing chromaticity and intensity[216] in a coherent psycho-visual space which improves performance in comparison to previous results[5]. A spectral residual method has been developed[219] which is a fast way of computing saliency. Also, attempts have been made to add skin (and motion) detection to a saliency map[223].

2.5.2 Orientation feature

One of the most commonly used biologically motivated processing unit in computer models of visual processing is the Gabor filters. The Gabor filter is composed of the product of a cosine grating and a 2D Gaussian envelope[222], and is applied as a filter at a number of different orientations, say 12, to sample orientation as intensity contrast along a number or directions. These are plausible as filters since they are localised orientation responses of variable scale, responses that have been found in neurons in physiological measurements (e.g. [224]). The detection of a contour for example can be construed as the detection of connected orientation elements. Pop-out has been demonstrated for orientation[222]

Not only real contours attract attention but also illusory contours. A typical example of illusory contours is the missing parts of the Kanizsa triangle[105] as presented in Fig. 2.4. Here people tend to actually see a triangle in front of a couple of discs, and these illusory contours might even attract attention. Thus a model of short-range suppressive cross-orientation and cooperative long range interaction effect given orientation measures[225] and cognitive completion[8] could be studied in a saliency model.

2.5.3 Motion and dynamic features

Motion and dynamic events have received little attention since the bulk of studies are focused on still images. However, motion and dynamic events seem to attract attention

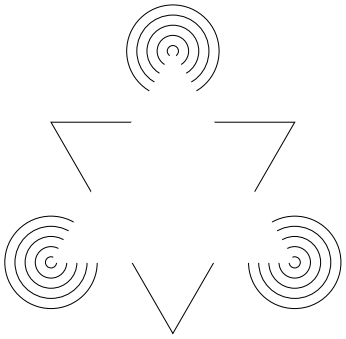


Figure 2.4: A version of the Kanizsa triangle.

more than static features, thus it is utmost important to extract and use such information in a model of visual attention[1]. Temporal flicker (i.e. onset and offset of light) and orientational motion energies[226] are fast to extract and simple features to use in a model of visual attention.

Although, some research by Abrams et. al. [227, 228] indicated that the onset of motion attract attention not the motion itself, a study[229] show that onset of motion is not necessary for motion to capture attention. Abrams et. al. replied[230] that the onset of motion does capture attention but in addition motion itself possibly attract attention. Further, some[231] have found that change detection is guided by bottom-up saliency as such, thus making it a secondary feature. This is demonstrated with experiments that show that low-level saliency, as defined by the Itti et. al. model[5](hereafter only called the Itti et. al. model), predict performance in a change detection task.

Temporal saliency has been calculated by hierarchical block matching motion estimation [232] and a method[220] using a self-resemblance measure has been proposed, which is a united model for static and space-time saliency detection, based on calculating the posterior probability given model and feature including values in a surrounding region.

2.5.4 Statistically based features

Common statistical measures have been used. The features of entropy, inertia and energy are supported by statistical analysis of the input signal (visual stimuli), but lack both empirical as well as biological support. Another model uses a precise mathematical formulation based on maximisation of information gain[218], where Shannon's self-information is calculated as $-log(p(\mathbf{f}))$ where $\mathbf{f}$ is a local feature vector derived from independent com-

ponent analysis (ICA) performed on a large sample of small RGB patches in the image, improving performance in comparison to the Itti et. al. Discriminant saliency has been calculated[233], where discriminant information is salient. These are defined as the points that best separate the class of interest from all others. By computing the salient feature of each texture class on a training database, enabling calculation of saliency on test images. Unfortunately, comparison to the Itti et. al. model is lacking, but results show improvement with respect to using the Harris saliency detector (corner detector) and Scale saliency detector (entropy measures, see [234]).

A massively parallel method has been presented[214], where a graph based calculation method improves results in comparison to what was produced by Itti et. al. [5]. Further, a space-time saliency model has been implemented[215], thus introducing dynamic events. Utilizing surprise to detect events has been successfully applied[235] and has been used[236] to calculate saliency.

With regards to the theory of information gain some research has shown that this is not driving attentional selection but rather the features currently being processed[237]. Further, it has been shown that statistical regularities guide the deployment of visual attention without semantic scene recognition[238].

Object based saliency calculation has also been obtained from grouping[239]. Here it is claimed that the Gestalt grouping principles form the necessary bridge between space based and object based attention.

2.6 Object detection

2.6.1 Classifier based detection

Object detection and specifically detection of faces, humans and moving objects is a difficult task where the detector must be reliable during change in illumination, pose and occlusion. Detection must often also be done across scale and is thus potentially a time consuming process. Heuristics must thus ensure speed-up in processing. Detection of human and faces is done on a still image basis and detection of moving objects on videos. These are thus two quite different problems.

Faces have been the most important feature to extract in the experimental work behind

this thesis, and these algorithms extract either local or holistic features, to distinguish faces from the background[240]. Different representations have been used in face detection, for example pixel–based[241, 242]. Only frontal faces have been detected[241] with a neural network and one of the problems is to select non-face training examples. This is solved by using the classifier during training to select non-face training examples, with typically around 8000 non-face ones used in a final training run. Their method detects 90.5% of the test set faces. Here a window is moved over the image in progressively down–scaled image and the area is classified as a face or not. To speed up classification a larger more general network is used of size 30×30 at intervals of 10 pixel. This network detects smaller faces than the window, and the detection is verified and more exactly localised with a 20×20 network in each position. The networks rely most heavily on the eyes, then the nose and then the mouth, and outperform both clustering and support vector machine methods..

Instead of full face feature representations, in a study[243], parts are represented as transformations of wavelet coefficients, with properties in space, frequency and orientation. Classification is done in stages where each stage can classify an area as not containing a face, and detect frontal as well as profile faces. Each part is detected as well as the configuration of the parts to classify a patch as a positive sample. A true positive rate of $60 - 85\%$ is obtained, which can definitely be improved with other methods, but there is a speed-up as compared to the neural-network approach.

Local edge features[244] have also been studied. Arrangements of oriented edge fragments are found from training examples. The information considered are comparisons of intensity differences and thus we have invariance to linear transformations on the grayscale and no histogram equalization needs to be done, being a speed–up in comparison to other methods. Discriminating arrangements of elementary edge test are done, by testing for specific edge arrangements, in a cascade to find faces in the image. A problem is that the smaller the scale the smaller is the probability of finding an arrangement of oriented edge fragments.

Finally, features with rectangular elements have been studied in an Adaboost trained detector[245] which detects faces in real-time and performs well even under partial occlusion. It is fast since the rectangular elements can be computed on the integral image. The cascaded weak classifiers are each representing a particular feature combination. In the

used implementation, opencv [1], the first one represents the eye and nose regions. Availability, speed and accuracy have been the reasons to choose this method in this thesis work for the detection of faces and pedestrians. It was at the point of selection of this algorithm not tested how well it performs in pedestrian detection and has been the work of a college[16] to test this in association with the work presented in this thesis.

2.6.2 Motion based detection

In video surveillance, foreground objects and their behaviours are the main objects of interest. Several techniques have been used to detect and analyse them. These techniques range from pixel-level change detection to higher level semantic object detection. The former are usually unsupervised methods whereas later ones requires some supervision.

Motion based object detection aims to identify the pixels that belong to a moving object between two frames. The first problem is to find pixels in one frame that are significantly different from the previous ones[246]. Change can come from appearance or disappearance of objects, motion of objects relative to background, shape changes of objects as well as changes in brightness or colour, and the second problem is to filter out unimportance change with respect to important change.

Techniques are background subtraction, temporal differencing and optical flow[247]. There are several problems that arise during pixel segmentation: bootstrapping (need to initialize), foreground aperture (inner part not segmented), ghosts (where background is suddenly visible doe to object leaving area), stopped objects, illumination changes (and shadows), camouflage (pixel features in background and foreground to similar), clutter in motion (movements of background pixels) and camera motion.

Background subtraction in its most simple form consists of taking a background reference frame and subtract the background from each frame. An improvement is to update the background each frame by a fraction. More recent approaches involve statistically modelling each pixel or groups of pixels, where outliers are classified as foreground pixels.

Temporal differencing is done by instead calculating the difference between the current frame and the previous frame. The advantages of using temporal difference is that it is indifferent to illumination changes, requires no bootstrapping and handles stopped objects

[1] Open Computer Visual Library url: http://sourceforge.net/projects/opencvlibrary/

well. A problem is that it generates ghost objects, since uncovered background becoming visible will be detected as moving pixels. Hybrid algorithms (e.g. [248]) that take advantage of both background subtraction and frame differencing have been developed.

Optical flow is an algorithm where feature points between frames are matched, calculating a flow of motion for each pixel. In such a way it works well even under camera motion. Moving objects can be segmented by looking at blobs having coherent motion, but suffers from limitations due to noise and background motion. Optical flow is computationally demanding and fails if constant brightness of objects and velocity smoothness conditions are violated[249], but can provide information that gives more detail on the movement of the target, for example distinguishing between rotation and translation.

The choice in this thesis work is background subtraction, which provides a fast algorithm that avoids the problem of ghosts. A reference frame is always available. Limitations of using simple reference frame differencing can be circumvented with statistical modelling. The output is viable both as simple evidence of a pixel being part of the foreground and as detections using blob grouping. As will be demonstrated shadows can also be removed with post-processing.

2.7 Object tracking

Tracking estimates the state of an object through frames in a video sequence. The problem consists in localizing and describing the same objects in successive frames. The state space can consist of a different number of parameters, for example position, width, height, contour, and pose. A summary of papers presenting tracking applications where detector input is utilised is presented in table 2.3.

Surveillance is the tracking of object for the purpose of detection, monitoring and identification. In surveillance the tracking of objects can be used to find and track faces and moving objects in outdoor or indoor scenes. In indoor office scenarios an object tracker in combination with a face recognition module can be used for identification and monitoring of people. Further, the tracking of the face of a lecturer can be used for automatic redirection of cameras. In multimedia database retrieval systems the knowledge of whereabouts of faces and moving objects can be used to build an indexing system for efficient retrieval of multimedia[250, 251, 252].

Several difficulties must be addressed in tracking. One of the major problems is automatic initialization, usually through a detection process. Present detectors fails to reliably detect objects of interest in every frame, and produce many false detections, which necessitates additional modules to avoid initializing false tracks. Another problem is partial or full occlusion. Partial occlusion is difficult since tracking should continue although the entire object is not fully visible. This is a problem of matching between model and data, since only partial data is available. Full occlusion or even disappearance of the target from the screen is different. Here tracking of the target should be terminated, and the same track should continue in case of reappearance of the target. In the visual attention system rescue saccades have been postulated for recovering occlusion[253].

2.7.1 Particle filtering

Tracking using particle filtering has been used extensively in the literature, for example in person tracking[254], tracking of active contours[255] and multiple object tracking[256]. It is a recursive estimator belonging to the probabilistic Bayesian family of estimators. Albeit slower than Kalman filtering[257] it can represent non-Gaussian and multi-modal distributions. Particle filtering estimates states from previous observations and uses an object model to calculate a posterior probability. Often a colour histogram model is used to model the object. A limitation of using a colour histogram object model is, that even if it is updated per frame basis, it cannot follow abrupt changes in illumination, or changes in which part of the object that is visible.

Particle filtering is a sequential Monte Carlo method and works with a discretization of the state space. Particles represent a *pdf* in a certain interesting amount of the state space. This *pdf* is successively updated for each frame with a motion model and an object model. The position in the current frame can be estimated from the previous position and velocity by spreading particles to an area around the estimation and then comparing an object model to what is found in image areas, as defined by each particle.

Let us represent the target state as $\mathbf{x}_t = [x, y, w, h]$, where t is the time, (x, y) is the centre of an ellipse approximating the shape of a target, and (w, h) the width and height

of the ellipse. The posterior *pdf* of object location in state space is

$$p(\mathbf{x}_{0:t}|\mathbf{z}_{1:t}) \approx \sum_{n=1}^{N_s} w_t^n \delta(\mathbf{x}_{0:t} - \mathbf{x}_{0:t}^n); \sum_{n=1}^{N_s} w_t^n = 1, \tag{2.9}$$

which is a sum of dirac-functions centred around particles $\mathbf{x}_{0:k}$ with weights w_t^n, and the number of particles is N_s.

Particles are spread according a zero order motion model around the previous particle position in state space. In particle filtering the update of the *pdf* consists of recalculating the weights with

$$w_t^n \propto w_{t-1}^n \frac{p(\mathbf{z}_t|\mathbf{x}_t^n)p(\mathbf{x}_t^n|\mathbf{x}_{t-1}^n)}{q(\mathbf{x}_t^n|\mathbf{x}_{t-1}^n, \mathbf{z}_t)}, \tag{2.10}$$

where $\mathbf{z}_t$ is the measurement and $\mathbf{x}_t^n$ the state of the nth particle in frame t. Thus $p(\mathbf{z}_t|\mathbf{x}_t^n)$ is the probability of the measurement or *likelihood*, given state $\mathbf{x}_t^n$, $p(\mathbf{x}_t^n|\mathbf{x}_t^{n-1})$ the state transition probability and $q(\mathbf{x}_t^n|\mathbf{x}_{t-1}^n, \mathbf{z}_t)$ the proposal distribution (see [258]). The current system uses a sampling importance re-sampling filter[258], which means that $w_{t-1}^n = 1/N \forall n$ and Eq. 2.10 simplifies to

$$w_t^n \propto \frac{p(\mathbf{z}_t|\mathbf{x}_t^n)p(\mathbf{x}_t^n|\mathbf{x}_{t-1}^n)}{q(\mathbf{x}_t^n|\mathbf{x}_{t-1}^n, \mathbf{z}_t)}. \tag{2.11}$$

The presented tracker uses a colour histogram as object model[259, 260]. Note that there are two types of *object models*: one from the classification and one for tracking. Tracking object model $\mathcal{M}$ is initialized with detection and updated online (see next section).

The likelihood is calculated as

$$p(\mathbf{z}_t|\mathbf{x}_t^n) \propto \frac{1}{\sqrt{2\pi}\sigma_l} e^{\frac{d_J(\mathbf{z}_t, \mathbf{x}_t^n)^2}{2\sigma_l^2}}, \tag{2.12}$$

with normalisation obtained later since we require $\sum_{n=1}^{N_s} w_t^n = 1$. $d_J(\mathbf{z}_t, \mathbf{x}_t^n)$ is the colour distance between the histogram associated to a particle and the colour model as measured by Jeffrey divergence[261],

$$d_J(\phi^{\mathcal{M}}, \phi^p) = \sum_{r,g,b} (\varphi_{r,g,b}^{\mathcal{M}} log(\frac{\varphi_{r,g,b}^{\mathcal{M}}}{\varphi_{r,g,b}^{\mu}}) + \varphi_{r,g,b}^{p} log(\frac{\varphi_{r,g,b}^{p}}{\varphi_{r,g,b}^{\mu}})), \tag{2.13}$$

where $\phi^{\mathcal{M}} = [\varphi_{1,1,1}^{\mathcal{M}}, ..., \varphi_{R,G,B}^{\mathcal{M}}]$ and $\phi^p = [\varphi_{1,1,1}^{p}, ..., \varphi_{R,G,B}^{p}]$ are the two histograms and $\phi_{r,g,b}^{\mu}$ is the mean of the histogram elements. The histogram has $10 \times 10 \times 10$ uniformly

quantized bins in the RGB space.

2.7.2 Integration of object detection with particle filtering

The incorporation of recent high-level observations with particle filtering improves performance [262]. It has for example been used in a hockey player tracking system[259]. Here the observation is the output of an Adaboost detector. The integration is done by adding a term into the distribution which consists of a Gaussian distribution around the detection in state space.

Instead of using the transition prior only, the *proposal distribution* will include current detections. First, some particles are spread according to the zero-order motion model, whereas the rest are spread around the classification results. This is incorporated in Eq. 2.10[259] with:

$$q(\mathbf{x}_t^n|\mathbf{x}_{t-1}^n, \mathbf{z}_t) = \alpha_c q_d(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{z}_t) + (1 - \alpha_c)p(\mathbf{x}_t, \mathbf{x}_{t-1}), \tag{2.14}$$

where α_c is the fraction of particles spread around the detection in state space, c is f for face or p for people, $q_d(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{z}_t)$ a Gaussian around the detection and $p(\mathbf{x}_t|\mathbf{x}_{t-1})$ a Gaussian according to the zero-order motion model. Without an associated detection however $\alpha_c = 0$ and Eq. 2.14 reduces to

$$q(\mathbf{x}_t^n|\mathbf{x}_{t-1}^n, \mathbf{z}_t) = p(\mathbf{x}_t|\mathbf{x}_{t-1}), \tag{2.15}$$

and Eq. 2.11 reduces to

$$w_t^n \propto p(\mathbf{z}_t|\mathbf{x}_t^n). \tag{2.16}$$

A limitation of this method is that manual initialization is necessary, a problem dealt with in this thesis.

2.7.3 Variation to detector integration

An alternative method is to use a contour extraction algorithm instead of a detector[260]. Then you can include the contour as a part of a combined object model. This facilitates a much more robust estimation of state space than the colour model itself.

The contour is incorporated in the likelihood so that Eq. 2.12 becomes

$$p(\mathbf{z}_t|\mathbf{x}_t^n) = \alpha_s p_g + (1-\alpha_s)p_c, \tag{2.17}$$

where p_c is the likelihood that the candidate location colour belongs to the object This according to the Bhattacharyya coefficient on histogram distributions fed into a Gaussian instead of the Jeffrey divergence measure (see [260] for details). p_g is the likelihood that the contour belongs to the object. For this the maximum gradient

$$gr(x_i, y_i) = \max_{(y_R, y_R) \in L_R} gr(x_n, y_n) \tag{2.18}$$

is calculated by traversing in the local search along the normal direction for each contour point. The gradient is calculated with

$$gr_x(x_n, y_n) = I(x_n - 2, y_n) + 2I(x_n - 1, y_n) - 2I(x_n + 1, y_n) - I(x_n + 2, y_n), \tag{2.19}$$

$$gr_y(x_n, y_n) = I(x_n, y_n - 2) + 2I(x_n, y_n - 1) - 2I(x_n, y_n + 1) - I(x_n, y_n + 2) \tag{2.20}$$

and

$$gr(x_n, y_n) = \sqrt{gr_x(x_n, y_n)^2 + gr_y(x_n, y_n)^2}. \tag{2.21}$$

A normalised average

$$\Psi_{gr}(s) = \frac{1}{N_s} \sum_{i=1}^{N_s} gr(x_i, gr_y i), \tag{2.22}$$

given state s and number of contour pixels of the state ellipse N_s, is calculated and fed into a Gaussian with

$$p_g = \frac{1}{\sqrt{2\pi}\sigma_\Psi} e^{-\frac{(1/\Psi)^2}{2\sigma_\Psi^2}}. \tag{2.23}$$

2.7.4 Other methods

Other modern tracking applications also rely on combined face detection and prediction from the previous frame. A stochastical model has been implemented[263] to track faces, where faces are detected in a coarse-to-fine network producing a hierarchical trace of face detections. This is used in a trained probabilistic framework to determine face positions. For each frame not only the detections are considered, but also face states close to the

previous frame.

Here the probability of a trace is calculated as

$$P(\mathbf{X}^* = \mathbf{x}^*) = \prod_{\eta \in T^*} P_\eta(x_\eta) \tag{2.24}$$

where T is the trace and the conditional probability given an observation θ

$$P(\mathbf{X}^*|\theta) \propto \frac{P(\mathbf{X}^*(i(\theta))|\theta)}{P(\mathbf{X}^*(i(\theta))|B)}, \tag{2.25}$$

where $P(\mathbf{X}^*(i(\theta))|B)$ is the marginal probability given background model B. P_η are learned and the state is estimated by the MAP estimator given current measurement and previous observations where

$$\hat{\theta}_t = arg \max_{\theta_t \in \Theta} P(\mathbf{X}_t^*|\theta_t) P(\theta_t|\theta_{t-1}). \tag{2.26}$$

This is very similar to particle filtering integrated with a detector in the sense that current detection data is integrated into the tracking process, and in fact Eq. 2.25 could be used as the likelihood in a particle filtering framework. Instead a Markov model is used depending only on the previous frame, which does carry less information forward. The advantage of this method is the speed-up in computation from using a coarse to fine hierarchy of detectors, only requiring fine calculations on ambiguous areas. Some other methodologies (see Table 2.3) involve Kalman filter, neural networks and PDAF (Probabilistic Data Association Filter).

Ref	Target	Detector	Tracker	Where info from detector is integrated
[260]	Face	Contour extraction	Particle filter	Contour incorporated into the object model in a particle filter.
[263]	Face	Coarse to fine	Markov model	Positions close to the previous position are considered as well as detection positions.
[264]	Face	Neural Network	Motion segmentation	Positions close to estimation by motion segmentation are considered by the neural network.
[265]	Moving objects	Change detector	Entropy model	The detector output is used directly and only matched with the entropy model to create tracks.
[257]	Moving objects	Histogram matching and pixel-wise likelihood	Kalman filter	The Kalman filter object model is updated each frame.
[259]	Hockey player	Adaboost	Particle filter	Particles assigned to the detection area.
[266]	Dim targets	Bayes detector	Maximum a posteriori (MAP)	The detector and MAP probabilities are combined.
[267]	any target	Bayes Detection	Probabilistic data association filter	Detection data inputted to PDAF.

Table 2.3: Summary of papers presenting trackers where a detector is integrated in the tracking process.

2.8 Summary

To model visual attention is a challenging task, since so many processes are involved, and the few ones known are not well characterized. Current research explores models that imitate aspects of human visual attention. Considerable research has been done on bottom-up visual attention modelling, but one of the most important questions to answer is how top-down processing comes into the equation, e.g. task and context. Questions remaining to be answered are what are the processes to decode the visual information stream, what are the representations and what are the driving forces and ultimately goals of visual attention.

How visual input is decoded into in the end cognitive descriptions of the outer world is intriguing. The intermediate representation used to interpret the world, e.g. $2\frac{1}{2}$ sketch of Marr's theory of perception[268] has not yet been described. It is the mid-level processes and representation connecting low-level and high-level mechanisms that are most unknown. Obviously, features like colour, intensity and orientation are elements that are processed. However, the elements are most likely grouped[105] to form objects of higher abstraction order, such as thoughts, that can be conveyed in spoken language and manipulated in a conventional logical or rational manner and in turn later be fed back to fuse as information guiding visual attention. Top-down processes provide task, context and object templates to be matched with the incoming data, bottom-up at least features and similar low-level information. Yet, how much bottom-up processing is done before fusion with top-down information is an open question.

Visual attention from low-level features alone is without the higher-level cognitions like goals and without psychic energies, or spirit, that set the organism into motion. For example, hunger might lead to fixations on food, and might form the goal of obtaining a meal. In the context of a kitchen the person is most likely fixating on utensils, to grab them, and prepare the dinner. In this hypothetical scenario the subject is an acting agent and visual attention is a part of interacting with the outer world. Low–level saliency could simply be sort of a guide to these activities, but cannot account reasonably for the entire allocation of visual attention. Thus higher level concepts are most likely necessary to incorporate as factors influencing attention.

Adding high-level features to a saliency map will provide us with relevant experimental

data. Attention to objects is first related to the general interpretative mechanism of the brain. Seeing objects is an ongoing process in a seemingly task-less environment. In other scenarios like surveillance and talking to others it is an inherent part of the task.

Low-level spatial features like colour contrast, intensity contrast, orientation as well as the dynamic feature flickering have a well established biological base. For example colour features invoke excitation in specific areas in the brain and contribute to a pop-up effect, especially interesting from the perspective of visual attention. Given the low-level features of Itti's validated model[1] to calculate low-level saliency, it is easy to expand with high-level features as a matter of further investigating the multifaceted aspects of visual attention. Valuable insight can be made in how saliency (if a biological fact) is fused with higher-level information, via some sort of global saliency map that includes top-down information or via different direction modules that distribute motor commands. For this it might be necessary to study the intricate structure of visual attention in terms of different processes that might give rise to different patterns in eye-traces, perhaps temporally defined[269]. Trivially, e.g. fixations and saccades can be extracted from eye-tracking data with the help of automated classification[270]. In the combined method IOR should be possible to measure. The integration of low–level and high–level information can also be studied by attempting to generate eye-traces automatically given different models.

In the area of tracking, particle filtering is a reliable established technique that not only can be used in itself but also can be improved with additional information like detections of varying types and structural entities like contour. If the start of the track can be established with detections and the tracking with the integrated system, it should be possible to build a good tracker. The Adaboost-trained detector based on features with rectangular elements is a widely used detector that is especially tested for faces. The cascade of classifiers and the use of intensity image speeds up calculations, first since features with rectangular elements are simply calculated from the input image, and second since most areas are rejected early, concentrating computation on ambiguous areas similarly to the coarse to fine method[263]. Change detection is a good choice for a classifier that detects for example vehicles and other moving objects in surveillance scenarios, that will be the visual input to the developed algorithms in later chapters.

To sum up the limitations of current described models are that high-level features

are not tracked, and are based on the saliency, winner-take-it-all and inhibition-of-return paradigm, which do generate a scan-path. In this theses these limitations are addressed, first by adding the output of tracking modules into a saliency model of attention. Second a new model is developed, that breaks with the saliency, winner-take-it-all and inhibition-of-return paradigm, and is based on attention to objects and statistics of real eye-traces, and qualitative descriptions of traces point to some of the mechanisms needed in a visual attention model. Finally, some visual attention concepts are analysed, which leads to theoretical clarifications as well as suggestions for future experimental work.

Chapter 3

Object detection and tracking

3.1 Introduction

In this chapter a multi-object tracking system is presented that uses an object detection algorithm and integrates its output into an object tracking module based on particle filtering. The original idea behind this is to use the detector to detect objects of interest (faces, people or moving objects), in frames where the detector works, and to use the particle filter to track the objects in between detections. In this work this idea has been extended to enable more elaborate interaction between modules. Initialization and termination is done automatically. The particle filter[15] and pedestrian and moving object detection modules[16] have been developed within the MMV lab and the extension to the tracking framework used in this work has been published[14].

Tracking of objects in videos offers many challenges. Objects change shape and appearance and a good tracker needs to be able to manage initialisation (object appearance), termination (object disappearance) and reinitialisation (object reappearance), after temporary occlusion events. There is a limitation to the shape and appearance changes particle filtering using a colour histogram model can do alone and the integration with detections enable better tracking, by adjusting the state with information from the detection process, as well as updating the object model.

The detection of faces and people is done by a cascaded Adaboost algorithm and the detection of moving objects by a change detector. In this face and pedestrian detection use four dimensions ($width$, $height$, x, and y) to support object tracking in four dimension.

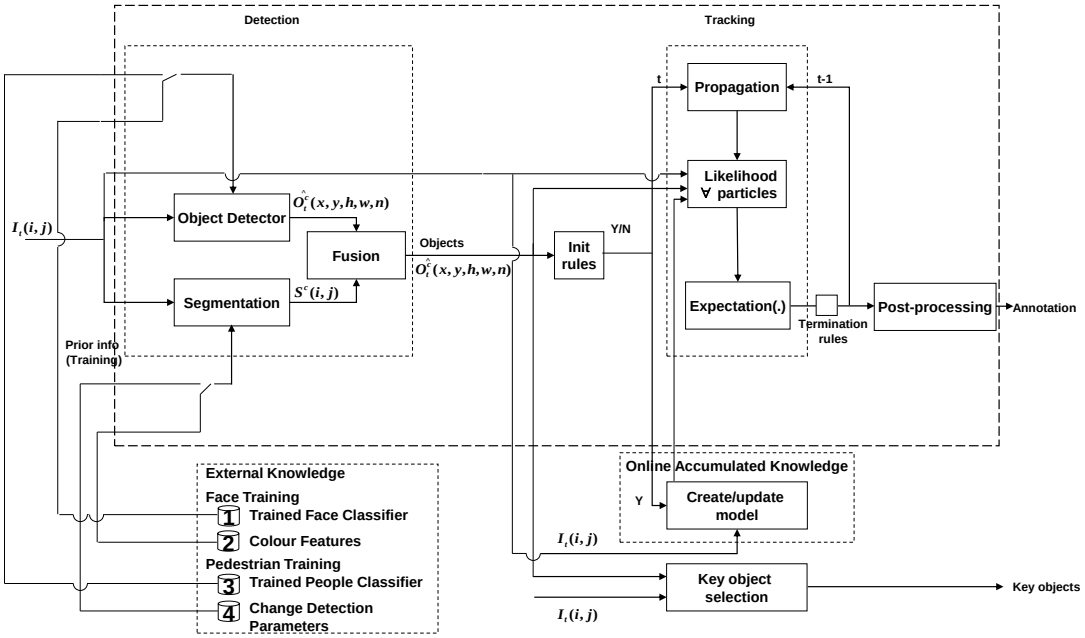


Figure 3.1: A flow chart of the face and people tracking system. To the left the image is inputted to segmentation and detection. The information is feed into a particle filter (tracking) algorithm. Initialisation and terminations rules determine the beginning and end of tracks.

When it comes to using change detection, grouped into blobs, four or five dimensions can be used. Detections are not only utilised to overcome the limitation of particle filtering alone, but in addition to initialize and terminate tracks.

Two different systems have been developed. The first one tracks faces and people, relying directly on detectors for the particular tracked object (see Fig. 3.1). The basic buildings block is fusion of detection and segmentation data to the left and its integration with particle filtering to the right (propagation, likelihood and expectation). Further, the initialisation rules and termination rules are important separate functions in the chart. The second one is a moving object tracker, at this point able to discriminate people from vehicles, set up to track in four dimensional or five dimensional mode (see Fig. 3.2). The major difference is how detection is done.

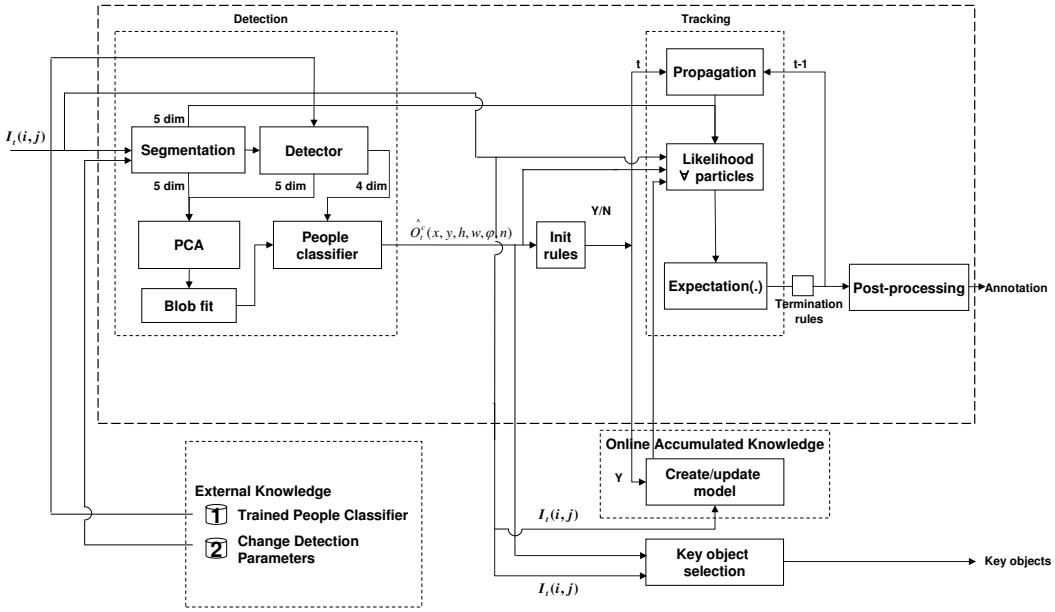


Figure 3.2: A flow chart of the four and five dimensional moving object tracking system. The information is feed into a particle filter (tracking) algorithm. Initialisation and terminations rules determine the beginning and end of tracks.

3.2 Detection

3.2.1 Adaboost face and people detection

Prior knowledge about object category is incorporated by training an object classifier. In particular, an Adaboost trained, rectangular element based, feature classifier [258, 271, 272] is used to detect faces and people. I will not go through the details here (see [271, 272]), but only explain why it is fast. This is since there is a restriction to utilize only rectangular areas in the features, which enables calculation on the integral image $\mathcal{I}(x,y)$, defined as

$$\mathcal{I}(x,y) = \sum_{i=1}^{x}\sum_{j=1}^{y} I(i,j), \tag{3.1}$$

where $I(i,j)$ represents the original image intensity. Since the features are differences between sums of all pixels within particular sub-windows, using Eq.(3.1), the sum of all pixels within a sub-rectangle $\mathcal{R}$ can be calculated with only four lookups

$$\sum_{(x,y)\in\mathcal{R}} I(x,y) = \mathcal{I}(x_2,y_2) - \mathcal{I}(x_1,y_2) - \mathcal{I}(x_2,y_1) + \mathcal{I}(x_1,y_1), \tag{3.2}$$

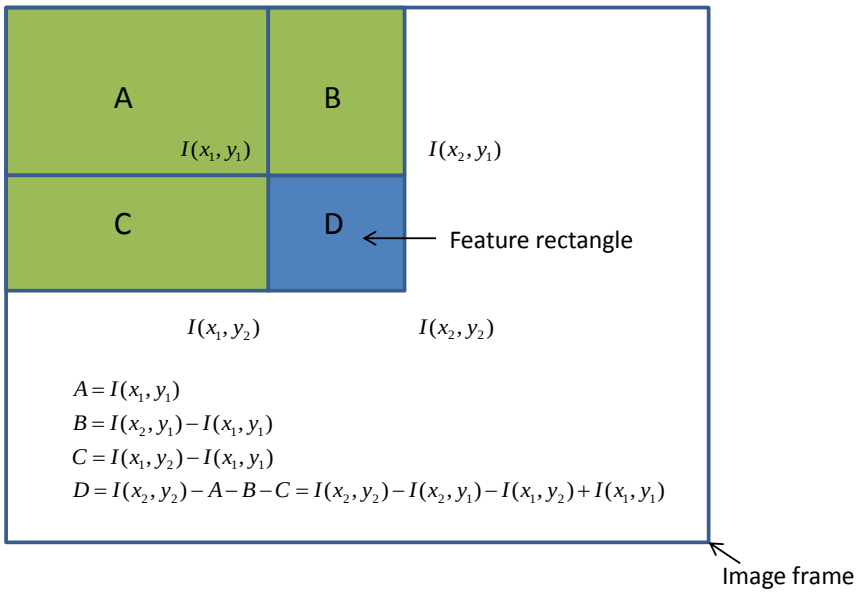


Figure 3.3: Illustration of how a feature rectangular element is calculated from the integral image.

where (x_1, y_1) and (x_2, y_2) are the top-left and bottom-right corners (see Fig. 3.3 for an explanation). Another integral image rotated by 45° is also calculated since rotated features are used as well.

For *faces*, a trained classifier[273] for frontal, left and right profile faces have been used. For *people*, training has been performed using 13 features [273, 272]. The method of training is Adaboost[274]. In the process the number of training samples is $n_t = n_t^+ + n_t^- = 4285$ with $n_t^+ = 2543$ positive training samples, selected from CLEAR[275] sequences, of resolution 10×24 and $n_t^- = 1742$ negative samples of different resolutions to train the classifier. Since there is one weak classifier for each distinct feature combination, effectively there are $2543 \times 13 = 33059$ weak classifiers for people classification. The Adaboost training selects and orders the best classifiers for fast classification (see [274] for details). Example output is presented in Fig. 3.6a.

The result of object classification is $\hat{O}_t^c(x, y, h, w, n)$, where c is the object class, $n = 1, ..., N_c$ is the number of the object of a certain category c in a frame at time t, (x, y) is the centre of the detection and (w, h) is the width and height of the detection. The tentative detection needs to be confirmed by low-level segmentation, described in the next section.

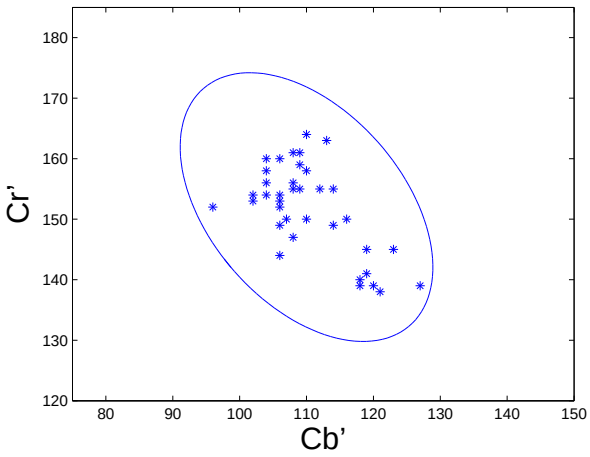


Figure 3.4: Skin colour ends up in a connected area in $C'_bC'_r$ space. Centres of ellipses defining skin colours in 38 sequences from the CLEAR evaluation video sequences on face tracking are marked with asterisks. The larger ellipse includes all these specific ellipses and is used to model skin chromaticity in the general case.

3.2.2 Skin chromaticity segmentation

Skin chromaticity segmentation is based on a non-linear transformation of the YC_bC_r colour space[217], which results in a new two-dimensional ad-hoc chromaticity plane $C'_bC'_r$. For grey pixels chromaticity is degenerate, and thus pixels with

$$0.975 < \frac{R}{B}, \frac{G}{R} < 1.025, \tag{3.3}$$

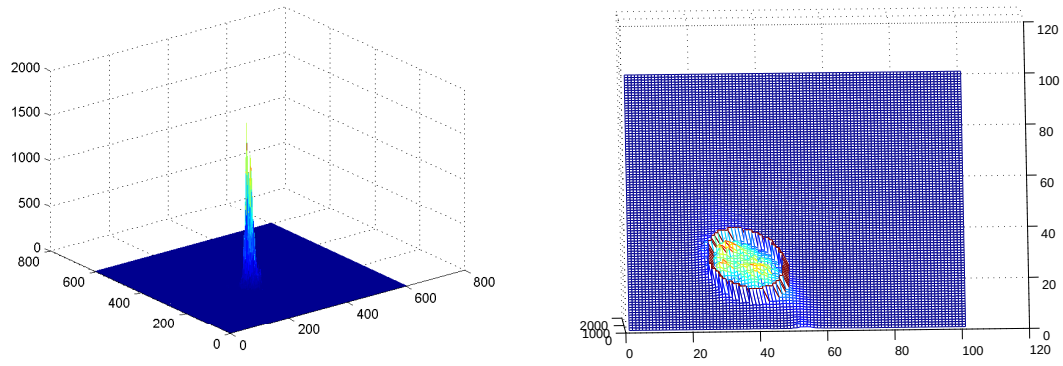
in the RGB colour space are discarded. To distinguish skin pixels in the $C'_bC'_r$ plane an ellipse encircling skin chromaticity is defined as

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1, \tag{3.4}$$

with

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} C'_b - c_x \\ C'_r - c_y \end{bmatrix}. \tag{3.5}$$

For the experimental results in this paper skin chromaticity was sampled from segments of the CLEAR [275] evaluation video sequences, determining the values $c_x = 110$, $c_y = 152$, $a = 25$, $b = 15$ and $\theta = 2.53$ which are similar to the replicated model[217] (see Fig. 3.4).



(a) (b)

Figure 3.5: Sampling of face pixels in $C_b'C_r'$-space generate a peak in (a). By encircling this peak (b) skin colour can be separated from background colour.

The transformation is defined in the following set of equations:

$$C_\tau' = \begin{cases} (C_\tau(Y) - \bar{C}_\tau(Y)) \cdot \frac{W_{C_\tau}}{W_{C_\tau}(Y)} + \bar{C}_\tau(K_h) & \text{if } Y < K_l \text{ or } K_h < Y \\ C_\tau(Y) & \text{if } Y \in [K_l, K_h] \end{cases}, \tag{3.6}$$

$$W_{C_\tau}(Y) = \begin{cases} WL_{c_\tau} + \frac{(Y - Y_{min}) \cdot (W_{C_\tau} - WL_{C_\tau})}{K_l - Y_{min}} & \text{if } Y < K_l \\ WH_{C_\tau} + \frac{(Y_{min} - Y) \cdot (W_{C_\tau} - WH_{C_\tau})}{Y_{min} - K_h} & \text{if } K_h < Y \end{cases}, \tag{3.7}$$

$$\bar{C}_b(Y) = \begin{cases} 108 + \frac{(K_l - Y) \cdot (118 - 108)}{K_l - Y_{min}} & \text{if } Y < K_l \\ 108 + \frac{(Y - K_h) \cdot (118 - 108)}{Y_{max} - K_h} & \text{if } K_h < Y \end{cases}, \tag{3.8}$$

and

$$\bar{C}_r(Y) = \begin{cases} 154 - \frac{(K_l - Y) \cdot (154 - 144)}{K_l - Y_{min}} & \text{if } Y < K_l \\ 154 + \frac{(Y - K_h) \cdot (154 - 132)}{Y_{max} - K_h} & \text{if } K_h < Y \end{cases}, \tag{3.9}$$

where C_τ stands for either C_r or C_b, $W_{C_b} = 46.97$, $WL_{C_b} = 23$, $WH_{c_b} = 14$, $W_{C_r} = 38.76$, $WL_{C_r} = 20$, $WH_{C_r} = 10$, $K_1 = 125$, $K_h = 188$, $Y_{min} = 16$ and $Y_{max} = 235$.

Skin colour has also been sampled from some web-camera sequences (see Fig. 3.5). Here the centre of the ellipse is outside of the model in Fig. 3.4.

3.2.3 Motion segmentation

Motion segmentation in people tracking is used to support the detections (see section 3.2.4). It is also used as detector of moving objects, either using simply blob bounding boxes in the case of four dimensional moving object tracking, or fitted ellipses in the case of five dimensional moving object tracking.

In the presented system, foreground segmentation is performed using a statistical colour based change detector [276], to detect changes with respect to constructed reference background. The result of the segmentation is heavily affected by noise introduced in the acquisition process. To overcome the effect of noise, a procedure was used, which is based on the hypothesis that the additive noise affecting each image of the sequence respects a Gaussian distribution with mean μ_n and standard deviation σ_n. The σ_n value of Gaussian, in each sequence is selected by performing the histogram analysis of image difference (in areas without moving objects) in RGB colour space. From the data sampled in these histograms the standard deviation σ_n is estimated for each sequence. Any isolated noise is further removed using the morphological operators erosion and dilation.

3.2.4 Evidence fusion

Segmentation results are used to remove false positive detections. The detection denoted $\hat{O}_t^c(x_d, y_d, w_d, h_d, n)$, with (x_d, y_d, w_d, h_d) being the centre (x_d, y_d), width, height and detection number in frame, and is accepted if

$$\frac{|\hat{O}_t^c(x_d, y_d, h_d, w_d, n) \cap S^c(i,j)|}{|\hat{O}_t^c(x_d, y_d, w_d, h_d, n)|} > \lambda_c, \tag{3.10}$$

where $|.|$ is the cardinality of a set, λ_c is the overlap ratio and $S^c(i,j)$ segmentation result for each pixel (i,j). The results of *colour segmentation* support the final decision by requiring that a face must contain at least 10% skin pixels ($\lambda_f = 0.1$). The results of *motion segmentation* support the final decision by requiring that a person detection from Adaboost classification must contain at least 20% change pixels ($\lambda_p = 0.2$). The reason for low thresholds is that detections often contain background as well as hair in case of faces. The λ values have been validated experimentally. Further, both segmentation algorithms produce a low percentage of segmented pixels for true object in rare cases, e.g. due to

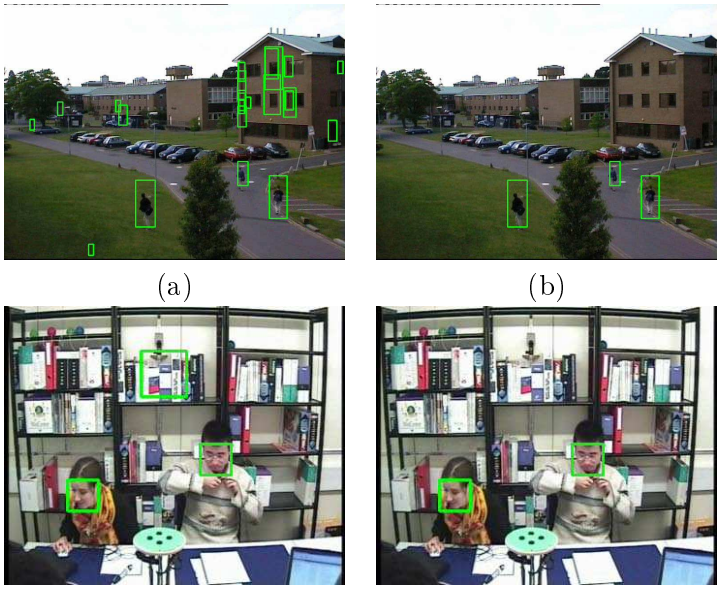


Figure 3.6: Examples of removal of false positives using segmentation. (a) People detection using Adaboost. (b) Removed false people detections. (c) Face detection using Adaboost. (d) Removed false detections.

poor illumination conditions for face and mixture of object with background for change detection. Detection with and without fusion of segmentation is displayed in Fig. 3.6. The results on tracks are illustrated in Fig. 3.7.

3.2.5 Fitting an ellipse to motion segments

Previously it has been assumed that detected objects (i.e. blobs) have no orientation $\theta = 0$, i.e. they have been marked with an axis-aligned ellipse. For five dimensional tracking an ellipse fit metric has been developed in this work which estimates objects physical location better. The fit is established by first performing PCA on the blob of each moving object. This identifies the major and minor axis and provides an estimate of radii, which is used as initial state in a maximization algorithm, where the fit is iteratively maximised. This restriction will ensure that parts of the blob like minor shadows are not included in the detection.

The blob fit is based on the following measure

$$F_b = \frac{N_b^3}{N_e^2 A_b}, \tag{3.11}$$

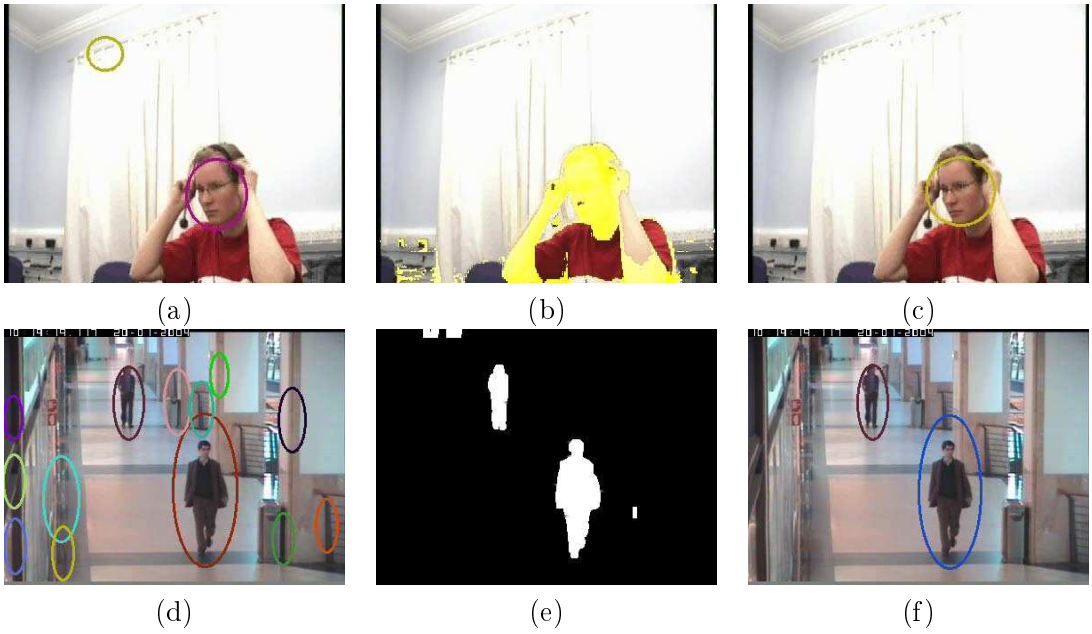


Figure 3.7: (a) Face tracks (instantaneous) generated without fusion of colour segmentation results. (b) Colour segmentation results. (c) Face tracks generated with fusion of colour segmentation results. (d) People tracks generated without fusion of motion segmentation results. (e) Change segmentation results. (f) People tracks generated with fusion of change segmentation results.

where N_b is the number of blob pixels within the ellipse, N_e the pixel area of the ellipse and A_b is the pixel area of blob. For N_b the blob pixels are only counted if the pixel is within the blob bounding box, ensuring that blob pixels are not counted from objects close by. The formula is based on that we want to maximise the number of blob pixels in the ellipse and at the same time minimise the pixel area of the ellipse. The former tends to expand an good ellipse and the latter contracts it and the opposing forces will stabilise a good intuitive fit of an ellipse to a blob.

The maximisation algorithm works by iterating in each dimension in the range $[-10\ 10]$ from the previous location in steps of 1. This equates to finding

$$argmax_d(F_b(p+\delta)); \quad \delta \in [-10, 10], \tag{3.12}$$

where d refers to the dimension (i.e. x, y, w, h or θ) and p the previous best value. For each dimension the best ellipse is moved forward, and this is continued until no better fit is found within the range for each dimension. To use the range $[-10\ 10]$ ensures that the algorithm does not just find the first local optimum. The algorithm lacks resolution

(a) (b)

Figure 3.8: Illustration of the fit of ellipses to blobs with a fit measure and a maximisation algorithm. (a) The original image. (b) Fitted ellipses to each object blob as outputted by the moving object detection process. The four dimensional tracker uses the bounding boxes as detection input, whereas the five dimensional code uses the ellipses.

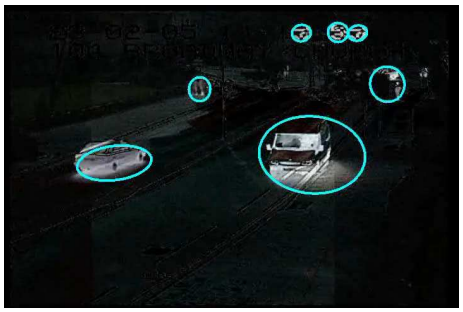
invariance, but is sufficient for the range of resolutions used on this thesis. Fig. 3.8 displays the result of this, and as can be seen the orientation of the bus in (a) is reflected in the estimated ellipse in (b). Basically, the four dimensional moving object tracking result in tracks that are of the size of the bounding boxes of each object, thus covering areas of the input image that does not contain moving objects.

A possibility is to use the difference image instead of the blob image to do the fitting of the ellipse. In such case equation 3.11 becomes

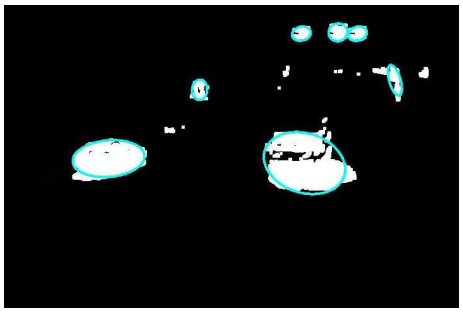
$$F_d = \frac{\sum_e I_d(i,j)^3}{N_e^2 A_b}, \tag{3.13}$$

where $I_d(i,j)$ are the pixels of the difference image within the ellipse e. Similarly only pixels within the bounding box are counted. One problem using the difference image is that it is not certain that pixels with higher difference values are more important than other ones in determining fit. In some cases half of the object has a colour which has a lower difference than the other half, resulting on only half of the object being tracked. In other cases the shadow of an object has higher difference intensities than the object itself, resulting in mostly the shadow of the object being tracked. This effect is shown in Fig. 3.9, where the maximisation by equation 3.11 produces better estimation of the object state (Fig. 3.9c) for the car to the left, than equation 3.13 (Fig. 3.9b).

(a)



(b)



(c)

Figure 3.9: Illustration of better performance of the fit measure in equation 3.11 than of equation 3.13. (a) The original image. (b) Fitted ellipses to each object by equation 3.13. (c) Fitted ellipses to each object by equation 3.11. The leftmost ellipse in (b) encircles only parts of the entire car due to low difference values on the top of the car, whereas the ellipse in (c) encircles the entire car.

3.3 Tracking

3.3.1 Integration of object detection with particle filtering

Instead of using the same *object model* over time, the colour histogram model $\mathcal{M}$ is updated based on the successive detections. This update allows to continue tracking the object during pose as well as illumination changes. Let's say one side of an object which previously has been turned away from the camera appears from one frame and on, and that side contains pixels colours which are not present in previous frames. By updating the colour histogram model, this new side can be a part of the object model in successive frames. Without this update, the colour histogram model will always ensure exclusion of this new side. More frequently there are small colour changes in face tracks, due to orientation changes, as well as relatively larger changes in people tracking, since a part of the background is often present in the model. The histogram is updated according to

$$\varphi^{\mathcal{M}}_{(r,g,b)}(t) = \beta\varphi^{d}_{(r,g,b)}(t) + (1-\beta)\varphi^{\mathcal{M}}_{(r,g,b)}(t-1) \qquad \forall r = 1,...,R; g = 1,...,G; b = 1,...,B \tag{3.14}$$

where r, g and b are the indexes in respective histogram φ and $\beta = 0.25$ the fraction of update, and (r, g, b) indices in the histograms. To update the object model $\mathcal{M}$ online helps helps improving the robustness of the tracking algorithm even if object appearance changes drastically during the sequence (due to illumination, size or orientation changes).

In tracking, updating the colour model might cause drift if background pixels start to become a part of the model $\mathcal{M}$. Also, the colour model in people tracking often contain some background colour and this can make the track stick on the background if the person moves to an area with a differently coloured background. Since the histogram is updated only when there is an associated detection, this does not happen though, and instead it prevents drift in people tracking. Further, a modification to 2.12 is done for five dimensional moving object tracking by setting

$$p(\mathbf{z}_t|\mathbf{x}^n_t) \propto \frac{1}{\sqrt{2\pi\sigma_l}} e^{\frac{d_J(\mathbf{z}_t,\mathbf{x}^n_t)^2}{2\sigma^2_l}} + F_b \tag{3.15}$$

This ensures that states that fit a blob with the correct orientation are promoted by giving a higher value if many blob pixels are within the ellipse.

Theoretically, setting $\alpha = 1.0$ and $\beta = 1.0$ (Eq. 2.14 and 3.14) when there is an association, would make the track follow the detections almost completely. However, this would not create smooth tracks, since the detectors are not 100% reliable in terms of position and size.

For the integration between particles and detections to take place, an *association* must be established between existing tracks with states at time t $\mathbf{x}_t = (x_t, y_t, w_t, h_t)$ and current detections $O_t^c(x_d, y_d, h_d, w_d, n)$. The association is done using a gated nearest neighbour filter. The proximity conditions are

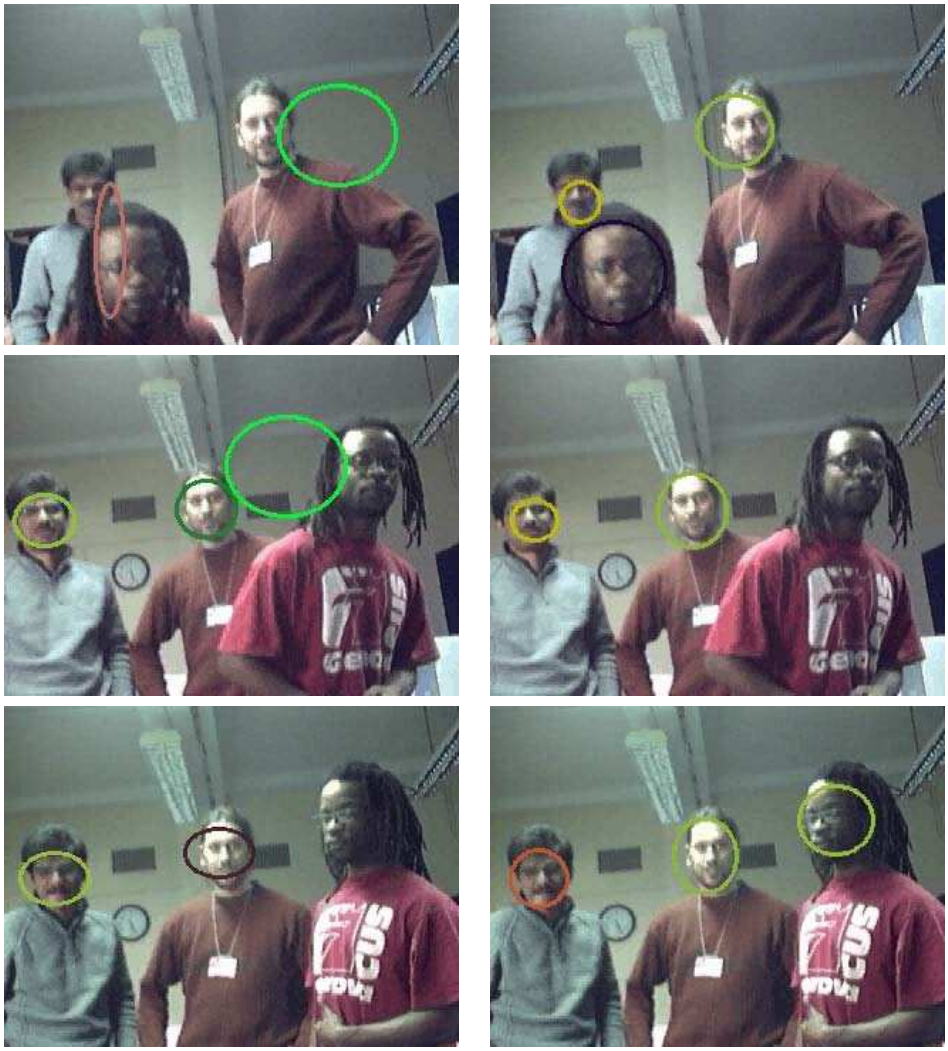
$$\begin{cases} |x_d - x_{tr}| < \delta_c(w_{tr} + \eta_c h_{tr}) \\ |y_d - y_{tr}| < \delta_c(\eta_c w_{tr} + h_{tr}) \\ (1-\gamma)w_{tr} < w_d < (1+\gamma)w_{tr} \\ (1-\gamma)h_{tr} < h_d < (1+\gamma)h_{tr} \end{cases} \tag{3.16}$$

where c is either f for face or p for human, $\eta_f = 1$ and $\eta_p = 0$, $\delta_p = 0.5$ and $\delta_f = 0.25$, x_d and x_{tr} are the horizontal centres of the detection and the track ellipse, and w_{tr} and h_{tr} are the width and height of the track ellipse. Further, for the width w and height h; where w_d and w_d are the width of the detection and the track ellipse respectively; $\gamma_f = 0.5$, $\gamma_p = 0.25$, and $\delta = 0.5$ were determined experimentally. If the proximity conditions are not satisfied, a new candidate track is initialized.

The result of integrating detections with particle filtering is illustrated in Fig. 3.10 and 3.11. Fig. 3.10 shows sample tracking result using the detector for track initialisation only. In several occasions the tracker loses the target and the ellipse visualising the target result does not overlap with the faces. In 3.11 there is no update of the colour histogram model, and one of the state estimation of the target is unsatisfying.

3.3.2 Track management

To account for initialisation and termination of tracks a number of rules are implemented. A detection in a new area is considered a candidate object appearance event(see table 3.1). Tracking is started but the track is in *sleeping mode*, i.e. it is not producing any output. Switching of tracks from *sleeping* to *active* mode is controlled by the successive detections. A certain number of detections are needed in successive frames to activate a track. The



(a) (b)

Figure 3.10: Illustration of limitation of particle filtering using the detector for initialisation only. (a) Tracks lost without integration of the detection with the particle filtering state estimation. (b) Better results with integration.

(a) (b)

Figure 3.11: Illustration of limitation of particle filtering without update of the the colour model. Without update of the colour histogram model state estimation is unsatisfying (a) however with update (b) state is accurately estimated.

Nbr.	Name	Description
1.	*Initialisation rule*	N_i detections, where N_i depends on frequency, initialises a track.
2.	*Termination rules*	
a)	Lack of detections	25 frames without detections terminates a track.
b)	Segmentation	A track is terminated when it is not supported by segmentation results.
c)	Overlap removal	A score is kept for each track based on length of track and frequency of detections. When two tracks overlap the one with lowest score is removed.
d)	Jeffrey divergence	When the Jeffrey distance between model and current track is to large the track is terminated.
e)	Size	To small or to large faces are discarded based on the mean and standard deviation of initial face detections.
f)	Face ratio	Face tracks are terminated if $\frac{w}{h} > 1.5$.
g)	Border objects	When the detection of a moving object is touching the border of the frame it is discarded. No new tracks are initialized in the borders.

Table 3.1: Proposed rules to use in tracking.

required number N_i is given by

$$N_i = \min\left(\frac{3}{2 - \frac{1}{f}} f, 9\right), \tag{3.17}$$

where f is the frequency of detections and $f = 9/20$ the lowest allowed frequency, a limit validated by qualitative evaluation of tracks. If there is not a sufficient number of successive detections the track is discarded.

For the moving object tracker, with detections based on change detection blobs, another rule has been necessary. When for example a car moves quickly into the camera view, first a too small track tends to initialise, since only a part of the object is visible. The track which fails to follow the car later on due to large inconsistencies in colour histogram model, and also in state space. Therefore a track is not initialised when the detector bounding box is touching the border of the screen.

3.3.3 Track termination

The most important rule used for termination is first the use of the segmentation described previously (section 3.2.2-3.2.3) for face and people tracking. A track is terminated if the low-level segmentation results do not provide enough evidence for the presence of an object i.e.

$$\frac{|\hat{T}_t^c(x_d, y_d, h_d, w_d, n) \cap S^c(i,j)|}{|\hat{T}_t^c(x_d, y_d, h_d, w_d, n)|} < \lambda_c, \tag{3.18}$$

where c is f for face or p for people, and $\hat{T}_t^c(x_d, y_d, h_d, w_d, n)$ is the nth target estimation at time t.

A person track is terminated if it contains $< \lambda_p$ change pixels ($\lambda_p = 20\%$) according to Eq. 3.18. A face track is terminated if it contains $< \lambda_f$ skin pixels ($\lambda_f = 10\%$). The effect of using this rule is illustrated in Fig. 3.12, which shows that false tracks on the background are successfully removed.

In people tracking, tracks are sometimes initiated on other moving objects like vehicles. Motion segmentation does provide support for such tracks, since there will be detected change in their occupied region. There needs to be another way to terminate such tracks. This is done by terminating tracks if there are $N_t = 25$ successive frames without an associated detection.

For face tracking, additional termination rules has been implemented. This since it sometimes happens that a track drifts away from a face to the background. The Jeffrey divergence measure (Eq. 2.13) is used to calculate the difference d between the current target and the colour histogram model. A cut-off distance of $d = 0.15$ has been found appropriate. There are however cases where tracks are over segmented, yet this phenomenon can be easily corrected with a post-processing step, as described later in section 3.3.4. The result of applying the histogram based rule is illustrated in Fig. 3.13.

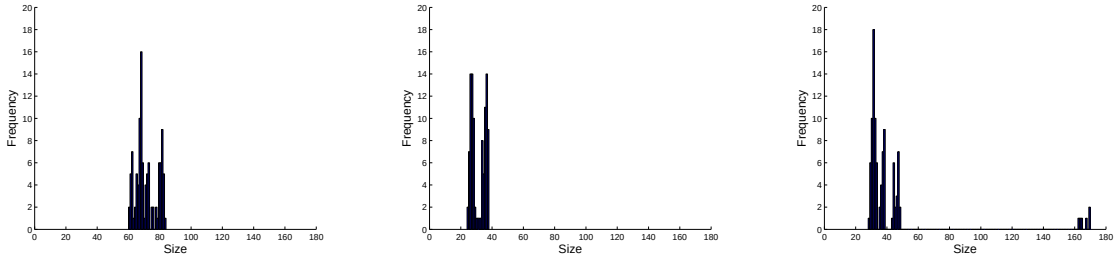
A second rule is based on sampling occurring sizes of faces. The average face size μ_{fs}, where face size is mean of width and height, and the standard deviation σ_{fs} are estimated from the first 150 tracked face states (see Fig. 3.14). Then tracks where the size of the face track state deviate more than $3\sigma_{fs}$ are discarded. The application of this rule is illustrated in Fig. 3.15, where one clearly false track has been successfully removed. Similarly, this rule can also incorrectly segment (i.e. cut) tracks short when the

(a) (b)

Figure 3.12: Illustration of using colour to distinguish between faces and non–faces for frames 50, 150 and 250 from a sequence in the CLEAR [275] evaluation dataset. In (a) colour is not used. In (b) three false positives have been removed. Here also the right most track has changed colour in the consecutive frames, which is due to fragmentation of tracks caused by the colour information.

(a) (b)

Figure 3.13: Illustration of termination by model (i.e. histogram) distance. In (a) a track has degenerated and does not follow the correct object any more. In (b) the track has been removed by measuring the Jeffrey divergence between the colour histogram model and current track.



(a) (b) (c)

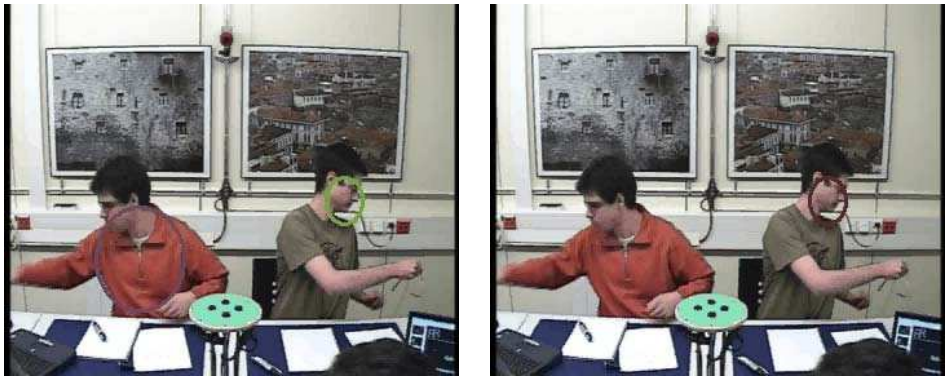
Figure 3.14: Sizes of the first 300 faces are sampled. Different distributions are obtained from example video sequences 1, 26 and 27 in the VACE face tracking dataset, as displayed in (a), (b) and (c) respectively. This is later used to remove faces which differ more that 3σ from the mean. In (c) outliers are sampled too the right, but these are still successfully removed by the model.

three standard deviations assumption does not hold. But this over-segmentation is also addressed with post-processing described in section 3.3.4.

Finally, it is very uncommon that real faces have a width/height ratio higher than 1.5. Therefore such tracks are removed as well. Unfortunately the width/height ration does not always hold when we track faces in profile pose.

3.3.4 Track verification, post processing and external knowledge

Detection can be generated in sub-parts of the tracked object, and to cope with this *track verification* is used, removing overlapping tracks. For example the face detector might find that the ear looks reasonably like a face, while the entire face is also detected. Also, the



(a) (b)

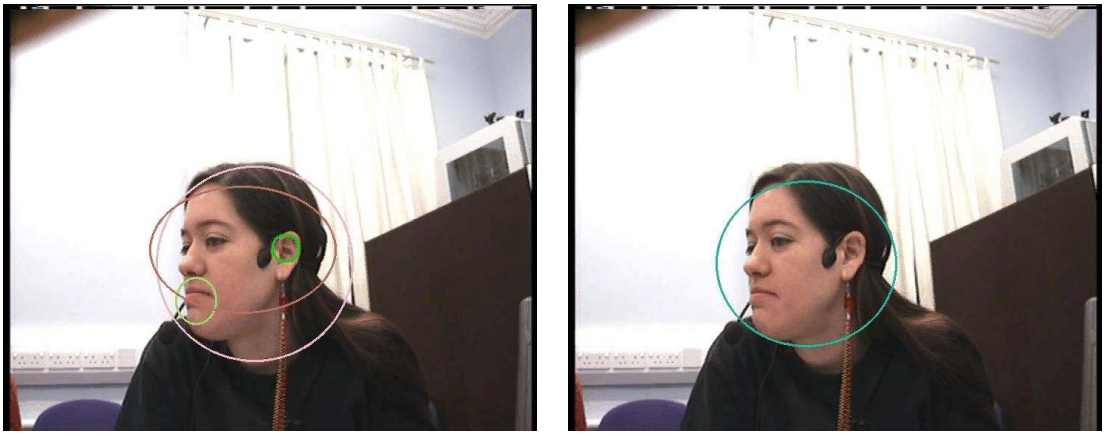
Figure 3.15: In many scenarios faces have sizes in a limited range. This can be utilized to remove the false track on the body in (a). This has been done in (b) by measuring the mean and standard deviation of faces in previous frames.

combination of two profile detectors and one frontal generate a lot of overlapping detections on faces. Since, longer tracks are more likely to be true tracks, as well as tracks with a high frequency of detection a probability of being a track being true can be estimated. If two tracks overlap the one with lowest probability to be a true positive is removed. For this purpose a score is calculated as:

$$s_t^n = (0.6N_f)/50 + 0.4fr_d, \tag{3.19}$$

where s_t^n is the score for track n at time t, N_f is the number of frames tracked up to 50 and fr_d is the frequency of detection. The different weights on N_f (0.6) and fr_d (0.4) favour tracks with a long history before new ones with a high frequency, and are only heuristically motivated. The effect of the use of the track verification score to remove overlapping tracks is illustrated in Fig. 3.16.

The tracks are post–processed to fix two of the problems generated by the face tracker. First, as mentioned some of the termination rules cause segmentation of tracks, and this can easily be fixed by rejoining tracks. This is done by finding pairs of tracks, where one track starts within 35 frames after the other one ends, where a limit of 35 was judged appropriate to exclude other objects moving to the end position of another track. If the



(a) (b)

Figure 3.16: Illustration of removal of overlapping tracks. In a) no removal of overlapping objects is done and several false tracks are generated on a face. In b) however only the correct track is kept.

following proximity conditions are satisfied:

$$\begin{cases} |x_1 - x_2| < (w_1 + w_2)/4 \\ |y_1 - y_2| < (h_1 + h_2)/4 \\ 0.5w_2 < w_1 < 1.5w_2 \\ 0.5h_2 < h_1 < 1.5h_2 \end{cases}, \tag{3.20}$$

where (x_1, y_1, w_1, h_1) and (x_2, y_2, w_2, h_2) are the ending and starting track states, the two tracks are joined and the gap is linearly interpolated. Second, the integration of particle filtering with detection data reduces the temporal smoothing aspect of the particle filter. Therefore, a triangular kernel of width 15 is convolved with the track, to remove high frequency components:

$$x^i_{conv}(t) = \frac{1}{64} \sum_{j=-7}^{7} (8 - |j|) x^i(t + j) \tag{3.21}$$

where x^i_{conv} is the filtered i'th dimension of track state at time t. Finally, very short tracks are likely to be clutter and therefore tracks shorter than 15 frames are removed. The improvement is difficult to show in images but output videos will have more temporally stable state space estimation.

External knowledge is the input to the trackers derived from training as well as knowl-

edge like size ratio. For face tracking external knowledges is provided by features of the face detector, as well as the parameters of the ellipse used for the colour segmentation. For the people tracker there are the features of the people detector, as well as parameters of the motion segmentation. For the moving object tracker there are the features used to classify people versus non people, bound on sizes and size ratios for people and vehicles, used in combination with the Adaboost detector, as well as parameters of motion segmentation.

3.4 Four trackers

Four different trackers have been developed using the same basic structure: a face tracker, a human tracker, a four dimensional moving object tracker and a five dimensional moving object tracker. The output of all the system is annotation (xml) in terms of trajectories. In addition to that the trackers have the capability to extract object examples, and in the case of faces these are classified into frontal, left and right profile, in the case of moving object, into human and vehicles.

3.4.1 Face and human tracker

Face and human tracking use the same framework as illustrated in Fig. 3.1. The difference is mainly in track management. In face tracking rules 2b-f, of table 3.1, are used, whereas the human tracker uses rules 2a-c. The people detector produces quite a lot of false negatives, thus the specific initialization rule in table 3.1 is not used. Instead a track is initiated after only one detection. The face and people trackers use the result of Adaboost detection, supported by low-level segmentation for the integration with particle filtering.

The resulting tracker is able to reliably track objects under different illumination conditions and different poses, can handle occlusions, object appearing from the side or any other position in the frame and object disappearing from the screen. An example frame with a face track is presented in Fig. 3.17a. The output of the system is an ellipse that encircles the object area. The system works with multiple faces (see Fig. 3.17b) and handles partial occlusion and different illumination conditions (see Fig. 3.18). The people tracker has problems with estimating the size of people (see Fig. 3.24) and with detecting humans on a dark background (see Fig. 3.25).

(a) (b)

Figure 3.17: Example tracks generated by the face tracker for (a) one face and (b) three faces.

(a) (b)

Figure 3.18: Example tracks generated by the face tracker under (a) partial occlusion and (b) low illumination.

(a) (b)

Figure 3.19: (a) Vehicle tracking between frames 3300 and 3400 (b) Vehicle and Pedestrian tracking between frame 3600 and 3530.

3.4.2 Four dimensional object tracking

The moving object tracker is similar to the people and face tracker, especially in the overall input and output of the modules. The rules used for termination are 2a, 2c and 2g. The output of vehicle and pedestrian detection and tracking is presented in Fig. 3.19. Fig. 3.19b also shows the effect of the illumination change, due to vehicle head lights, on the segmentation results of the change detector.

Instead of using Adaboost detectors it uses bounding boxes of blobs in the motion segmentation results. Motion segmentation generates a lot of small spurious detection, and to remove such noise only detections with blob area $A_b > 200$ are considered. Further, the detections are classified into people and vehicle by a number of conditions (see Fig. 3.20) and the Ababoost people detector is used as a component of this. The first conditions is that there is a person detection within the bounding box, that

$$\begin{cases} A_b < A_{vmin} \\ \frac{3 \cdot h_d}{3.8} > w_d \end{cases}, \tag{3.22}$$

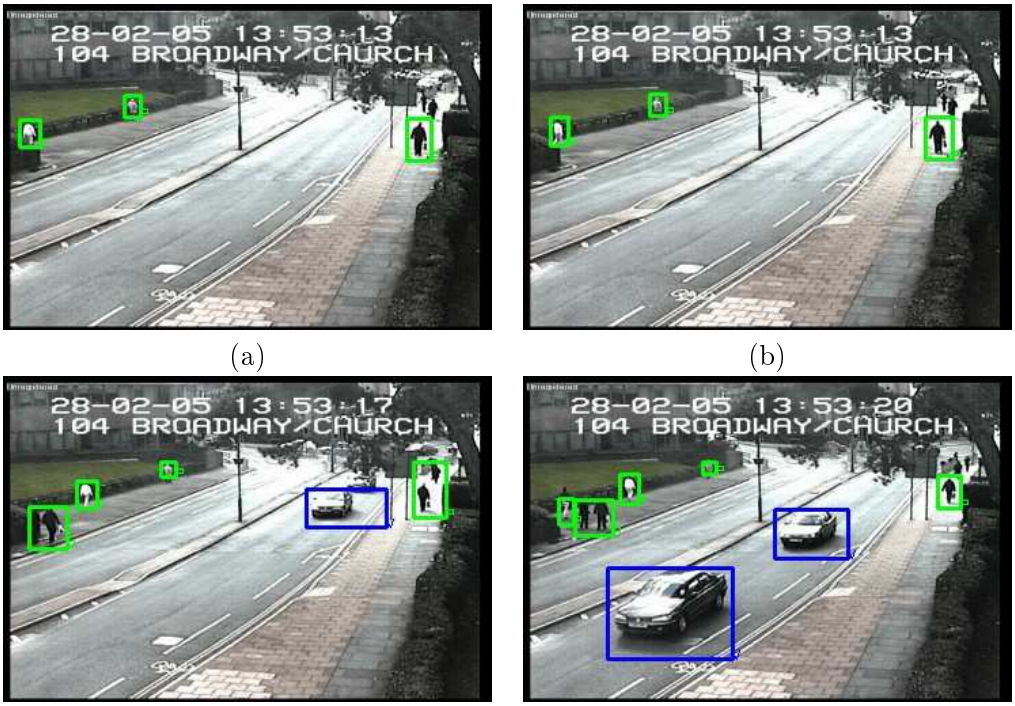


Figure 3.20: Example moving object detections for frame (a) 28, (b) 44, (c) 135 and (d) 202 on a selected sequence. Green indicates people and blue vehicle.

where $A_{vmin} = 200 pixels$ is the minimum area of a vehicle and (w_d, h_d) is the width and height of the detection. Further it is required that

$$\begin{cases} A_b < A_{pmax} \\ w_d < w_{im}/4 \\ h_d < h_{im}/2.5 \end{cases}, \tag{3.23}$$

where A_{pmax} is the maximum area of a person, and (w_{im}, h_{im}) the width and height of the video frame. Finally the moving object tracker can optionally treat border objects in a specific way (illustrated in Fig. 3.22). In this case the detection output is directly sent to the tracker, short-circuiting the particle filter module. The reason is simply that the particle filter fails to follow border objects as explained earlier.

3.4.3 Five dimensional object tracking

In addition to the four dimensional vehicle tracker described above also a five dimensional tracker has been developed. There were several reasons to develop this tracker. First, the

colour model does not work properly in situation where the difference in colour between object and background is not clearly defined, when a part of the object has the same colour as the background or the opposite. The fit measurement of equation 3.11 provides a way of keeping track of an object independent of colour, thus providing a model that better discriminates between object and background. Further, it can remove shadows of pedestrians and accurately track vehicles that are aligned neither to the horizontal nor the vertical direction (see Fig. 3.8).

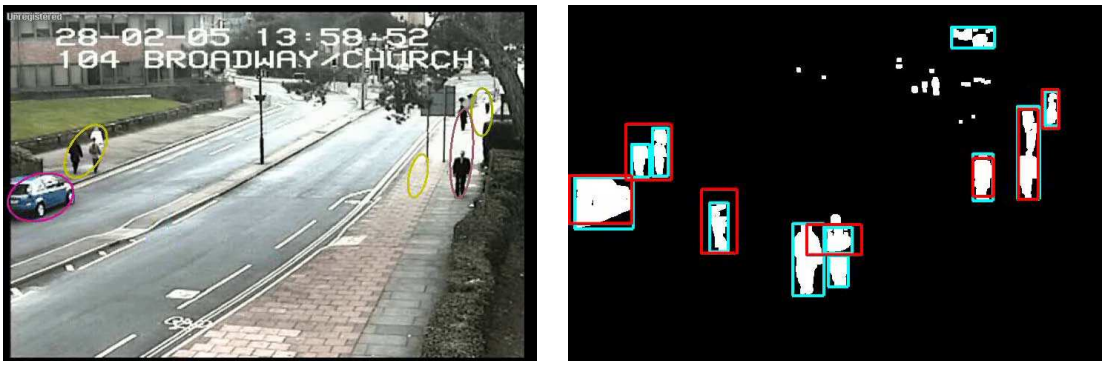
For the five dimensional tracker fitted ellipses according to equation 3.11 are used to initialise tracks, and inputted as detections to the particle filter (see equation 3.14). A further change to the algorithm is the association of a track to a detection. For this, bounding boxes of track ellipses are calculated and compared with the bounding boxes of detections. The rule is that a track bounding box must be contained within the detection bounding box and a frame of 40 pixels in all directions (see Fig. 3.21):

$$\begin{cases} x_{tr} - w_{tr}/2 > x_d - w_d/2 - 40 \\ x_{tr} + w_{tr}/2 < x_d + w_d/2 + 40 \\ y_{tr} - h_{tr}/2 < y_d - h_d/2 - 40 \\ y_{tr} + h_{tr}/2 < y_d + h_d/2 + 40 \end{cases} \tag{3.24}$$

The reason for this is that for example the ellipse around the pedestrian with a shadow in Fig. 3.8b is much smaller than the detection bounding box, and the conditions in equation 3.16 become inappropriate.

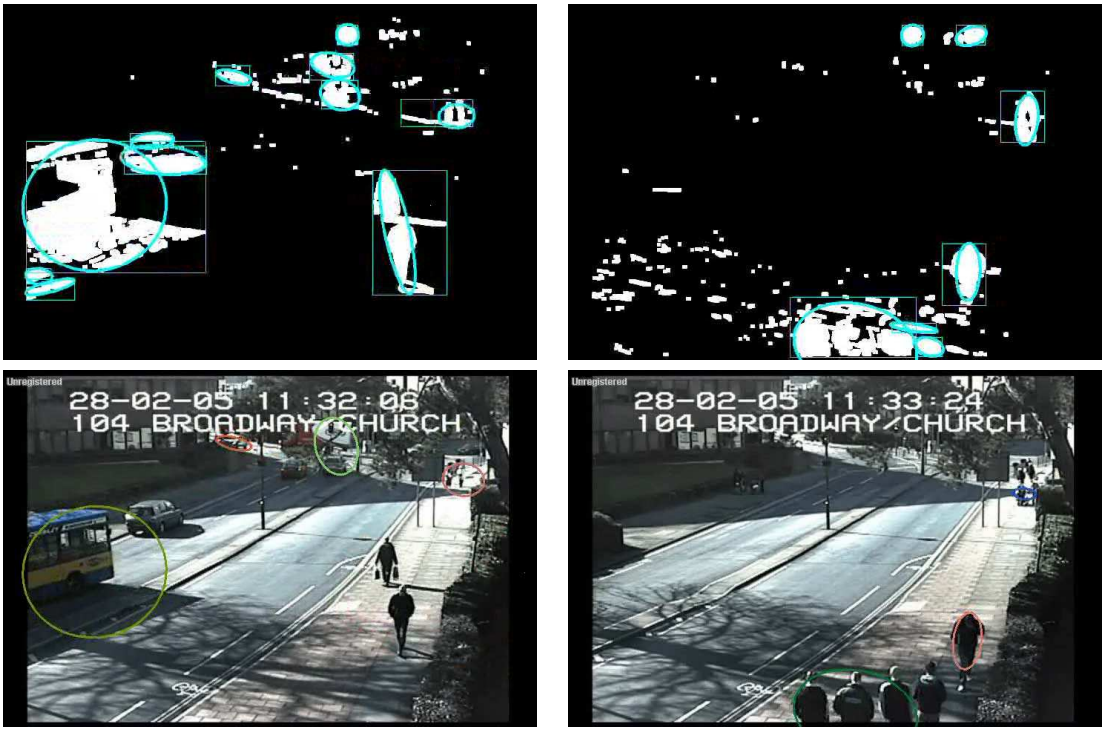
Finally, border objects are treated specifically. The detector ellipse output is directly inputted to the tracker and no particle filtering is done as illustrated in Fig. 3.22. In total only termination rule 2a according to table 3.1 is used.

The output of the five dimensional tracker is illustrated in Fig. 3.23. The main limitation of the algorithm is the output of the change detector, for example ghost objects and noise causing fragmentation of objects.



(a) (b)

Figure 3.21: For the five dimensional change tracker an association between a track and a detection is established if a track ellipse bounding box (in red) is within a detection bounding box (in cyan) with a frame of 40 pixels. (a) The tracks in a frame of sequence PVTRA101a04. (b) The change detector output, bounding boxes plus bounding boxes of tracks.



(a) (b)

Figure 3.22: Border objects are treated in the five dimensional moving object tracker by temporarily not using the particle filter, but directly using the fitted ellipse from detections as estimation of state space. This is the case for the bus in (a) and the group of pedestrians in (b) with change detector output and fitted ellipses above.

Figure 3.23: Example output from the five dimensional tracker processing sequence PVTRA101a04. The tracks covering non existing objects are actually ghost objects, that have not been removed by the change detector.

3.5 Results

3.5.1 Performance measures

To quantitatively evaluate the performance of the proposed system, two groups of measures where used. The first group of evaluation measures were for the annotation relating existence of object and detection. These are precision P and recall R and are defined as:

$$\begin{cases} P = \frac{TP}{TP+FP} \\ R = \frac{TP}{TP+FN}, \end{cases} \tag{3.25}$$

where TP is the number of *true positives* (true detections), FP is the number of *false positives* (false detections) and FN is the number of *false negatives* (missed detections).

The second group of performance measures is for the evaluation of the tracking itself (typical tracking metrics) that evaluate the system precision and accuracy as defined by the VACE evaluation standard[275]. The measures are Multiple Object Tracking Accuracy (MOTA) and Multiple Object Tracking Precision (MOTP) as well as DICE ($d_{\mathcal{D}}$, from Lee Raymond Dice[277]) and DIST ($d_{\mathcal{D}ist}$), weighted distance). MOTP and MOTA are defined as follows:

$$MOTP = \frac{\sum_{n=1}^{N_{fn}} \sum_{t=1}^{N_{fr}} \left[\frac{\mid G_n^{(t)} \cap D_n^{(t)} \mid}{\mid G_n^{(t)} \cup D_n^{(t)} \mid} \right]}{\sum_{u=1}^{N_{fr}} N_{fn}^u}, \tag{3.26}$$

where $G_n^{(t)}$ is ground truth and $D_n^{(t)}$ is detection, and

$$MOTA = 1 - \frac{\sum_{n=1}^{N_{fr}} (cs_{fn}(fn_n) + cs_{fp}(fp_n) + log_e(id_{sw}))}{\sum_{i=1}^{N_{fr}} N_G^n}, \tag{3.27}$$

where N_{fn} is the number of mapped objects over the entire track, N_{fn}^u to the number of mapped objects in the uth frame, N^{fr} is the number of frames, $cs_{fn}(fn_t)$ and $cs_{fp}(fp_t)$ are the cost functions for the missed detections and false positives, and id_{sw} is the number of false identity switches for each object during the sequence.

The measure $d_{\mathcal{D}}$ is similar to MOTP, but has been used for some evaluations, since it

was decided to use this in the publication related to this thesis. It is defined as

$$d_{\mathcal{D}} = 1 - \frac{\sum_{n=1}^{N_{fn}} \sum_{t=1}^{N_{fr}} \left[\frac{2 \mid G_n^{(t)} \cap D_t^{(t)} \mid}{\mid G_{n_{gt}}^{(t)} + D_t^{(t)} \mid} \right]}{\sum_{u=1}^{N_{fr}} N_{fn}^{u}}. \tag{3.28}$$

$d_{\mathcal{D}ist}$ is the distance between track ellipse and ground truth centres normalized by the width w_g and height h_g of the ground truth and is defined as

$$d_{\mathcal{D}ist} = \frac{\sum_{n=1}^{N_{fn}} \sum_{t=1}^{N_{fr}} \sqrt{(\frac{x_d - x_g}{w_g})^2 + (\frac{y_d - y_g}{h_g})^2}}{\sum_{u=1}^{N_{fr}} N_{fn}^{u}}. \tag{3.29}$$

3.5.2 Experimental results

The evaluation of the face and people trackers consists of quantitative measurements, graphs and illustrations. Based on the experimental results, we first demonstrate that the integration of particle filtering with a detector improve state estimation of targets. To this end we simulate ideal detections reading them from the ground truth instead of using the output of the Adaboost trained classifiers. This was done to isolate the detection part from tracking part. Further, the full system has been tested against regular particle filtering and the nearest neighbour algorithm. The nearest neighbour filter simply connects detections that are close in state space and time. Also initialization and termination the ground-truth instead of using track management has been evaluated. Tracks generated under these different conditions are displayed and discussed.

The system has been tested on standard datasets; i.e. the CLEAR [1] dataset for face detection and tracking task and four face sequences of the AMI corpus [2] for a surveillance task as well as one sequence from the PETS 2001 dataset [3] for people tracking. These are static single camera scenarios of people in meeting rooms for face tracking and people and vehicles on roads for the people tracking task. The dataset has both indoor and outdoor scenarios with varying illumination conditions. The details of sequences used for quantitative evaluation is given in Table 4.1.

[1] See http://www.clear-evaluation.org/ for information about the competition and data used.
[2] The AMI Corpus Publicly available at http://www.idiap.ch/amicorpus.
[3] The PETS 2001 corpus is publicly available at http://www.cvg.cs.rdg.ac.uk/cgi-bin/PETSMETRICS/page.cgi?dataset.

For the evaluation a variance of $\sigma = 1.8$ was used and a kernel size of $k = 3$ in the statistical change detector. The particle filtering algorithm is run with 150 particles per object and a transition factor of 12 pixels per frames. For the likelihood (Eq. 2.12) $\alpha_l = 0.068$.

First, several of the parameters of the face tracker and people tracker have been tested against four selected sequences each. Here no post-processing of the tracks have been done, which might give non-accurate indications if differences are small. The MOTP and MOTA scores are denoted P and A in Table 3.2.

For face $\alpha = 1.0$ (in Eq. 2.14) gives highest MOTP scores, however MOTA scores seems to be reduced, with the conclusion that $\alpha = 0.9$ is the best choice of the tested values. For people the value $\alpha = 0.5$ produce good results in general, however the scores are very similar. When it comes to the β parameter $\beta = 0.25$ seems to be a reasonable value for both face and people tracking. A γ (see Eq. 3.16) value of 0.5 seems to produce best results in general for faces, whereas it is more difficult to say for people tracking. Basically, a low value gives better MOTP, but a high one gives better MOTA. The reason for this is that the people detector often produce output which differ in size quite substantially from the actual person being tracked, which lowers accuracy. Higher accuracy is due to the fact that the detection supports the track in terms of existence. This is congruent with better results for lower values of α as well.

Removal by Jeffrey divergence (rule 2d) improves accuracy, however, it lowers precision for face tracking, and in general lowers scores for people tracking. The conclusion is that this rule is useful for faces in the sense that it removes false positives, and the loss of precision if very low, most likely due to segmentation of tracks. It is not useful for people tracking though, and this is probably due to frequent interference of background in the colour histogram model, due to not precise boundaries in the people detections.

Specific rules have been tested only on the respective trackers. First, to use an initialization buffer for people is not useful according to results. The use of the size model (rule 2e) for termination of face tracks lowers MOTP in three out of four cases, possibly due to segmentation of tracks. MOTA is markedly higher in sequence 9 (0.37 vs. 0.5) and a slightly higher in sequence 12 with termination by size (0.52 vs. 0.55), but it is slightly lower in sequence 10 and 11 (0.91 vs. 0.90, and 0.38 vs. 0.48,). The latter is because se-

(a) (b)

Figure 3.24: Example frames of successful people tracking. The people detector fails to estimate the size of the person in many cases, which results in too large target ellipses.

Figure 3.25: Because of limitations in the people detector, in some sequences many frames lack tracks on a majority of present people. The wall behind the people to the left prevents the detector from working.

quence 10 and 11 contain only one face, and the size model constructed from that has a very low standard deviation. Thus, when the head turns and size changes, the model is no longer appropriate. This could possibly be fixed by using a different threshold than 3σ when only one face is present.

Finally, the improvement of using the colour segmentation was substantial. Accuracy scores were markedly improved when the colour model was better adjusted for the video data. The results with the title "ellipse" uses the ellipse described in section 3.2.2. The results with the title "circle" uses the same centre, but has both $a = 25$ and $b = 25$ in Eq. 3.4.

To test the integration between detection and particle filtering, experiments were conducted where detections were taken from the ground truth instead of the Adaboost classifiers, thus removing problems arising from imperfect detections. Here tracks are initialized and terminated by the ground truth only and the results have been obtained with detec-

		People							
		5		**6**		**7a**		**8a**	
		P	A	P	A	P	A	P	A
α	0.0	0.58	0.15	0.52	0.34	**0.49**	0.35	0.54	0.32
	0.25	**0.61**	0.16	0.53	0.34	0.49	0.35	0.55	0.30
	0.5	0.60	**0.16**	**0.53**	**0.35**	0.48	0.35	0.54	**0.33**
	0.75	0.59	0.16	0.52	0.34	0.48	**0.36**	**0.55**	0.32
β	0.0	0.54	0.18	0.45	0.38	**0.50**	0.36	0.49	0.37
	0.1	0.57	0.19	0.49	0.43	0.49	0.35	**0.54**	0.37
	0.25	0.61	**0.20**	0.52	**0.44**	0.48	0.35	0.51	**0.38**
	0.5	**0.63**	0.19	**0.55**	0.40	0.46	**0.36**	0.52	0.35
γ	0.1	**0.65**	0.10	**0.55**	0.29	**0.52**	0.27	**0.58**	0.24
	0.25	0.60	0.16	0.53	0.35	0.48	0.35	0.54	0.33
	0.5	0.55	**0.16**	0.50	0.35	0.45	0.35	0.50	0.34
	0.75	0.52	0.16	0.46	**0.37**	0.43	**0.36**	0.48	**0.35**
λ_p	0.1	0.60	0.14	0.50	0.32	0.47	0.33	0.51	0.30
	0.2	0.60	0.16	0.53	0.35	0.48	**0.35**	0.54	**0.33**
	0.3	**0.61**	**0.17**	**0.55**	**0.35**	**0.51**	0.35	**0.54**	0.30
Jeffrey	Y	0.59	0.15	**0.56**	0.34	0.46	0.32	0.54	0.31
	N	**0.60**	**0.16**	0.53	**0.35**	**0.48**	**0.35**	**0.54**	**0.33**
Buffer	Y	0.59	0.15	**0.58**	0.30	**0.48**	0.28	0.53	0.26
	N	**0.60**	**0.16**	0.53	**0.35**	0.48	**0.35**	**0.54**	**0.33**
		Faces							
		9		**10**		**11**		**12**	
		P	A	P	A	P	A	P	A
α	0.5	0.27	0.30	0.21	0.95	0.22	**0.09**	0.15	0.55
	0.9	0.29	**0.48**	0.18	**0.97**	0.22	-0.07	0.11	**0.55**
	1.0	**0.31**	0.37	**0.23**	0.81	**0.23**	0.07	**0.17**	0.50
β	0.1	0.27	0.20	**0.20**	**0.97**	0.21	-0.13	0.06	**0.56**
	0.25	**0.29**	**0.48**	0.18	**0.97**	0.22	-0.07	0.11	0.55
	0.5	0.29	0.45	0.09	0.89	**0.23**	**0.05**	**0.16**	0.55
γ	0.25	0.27	0.37	**0.21**	0.90	0.22	-0.45	0.10	**0.56**
	0.5	**0.29**	**0.48**	0.18	0.97	**0.22**	**-0.07**	0.11	0.55
	0.75	0.23	-0.17	0.09	**0.99**	0.16	-0.22	**0.24**	0.35
Size	Y	0.29	**0.50**	**0.19**	0.90	0.22	0.38	0.13	**0.55**
	N	**0.29**	0.37	0.18	**0.91**	**0.22**	**0.40**	**0.15**	0.52
Jeffrey	Y	0.26	**0.51**	0.17	**0.92**	0.22	**0.48**	**0.15**	**0.58**
	N	**0.29**	0.37	**0.18**	0.91	**0.22**	0.40	0.15	0.52
Colour	Ellipse	0.28	**0.66**	**0.19**	0.95	0.22	**0.74**	**0.16**	**0.59**
	Circle	**0.29**	0.48	0.18	**0.97**	**0.22**	-0.07	0.11	0.55

Table 3.2: Comparison MOTA (A) and MOTP(P) scores using different parameter settings for face and people tracking.

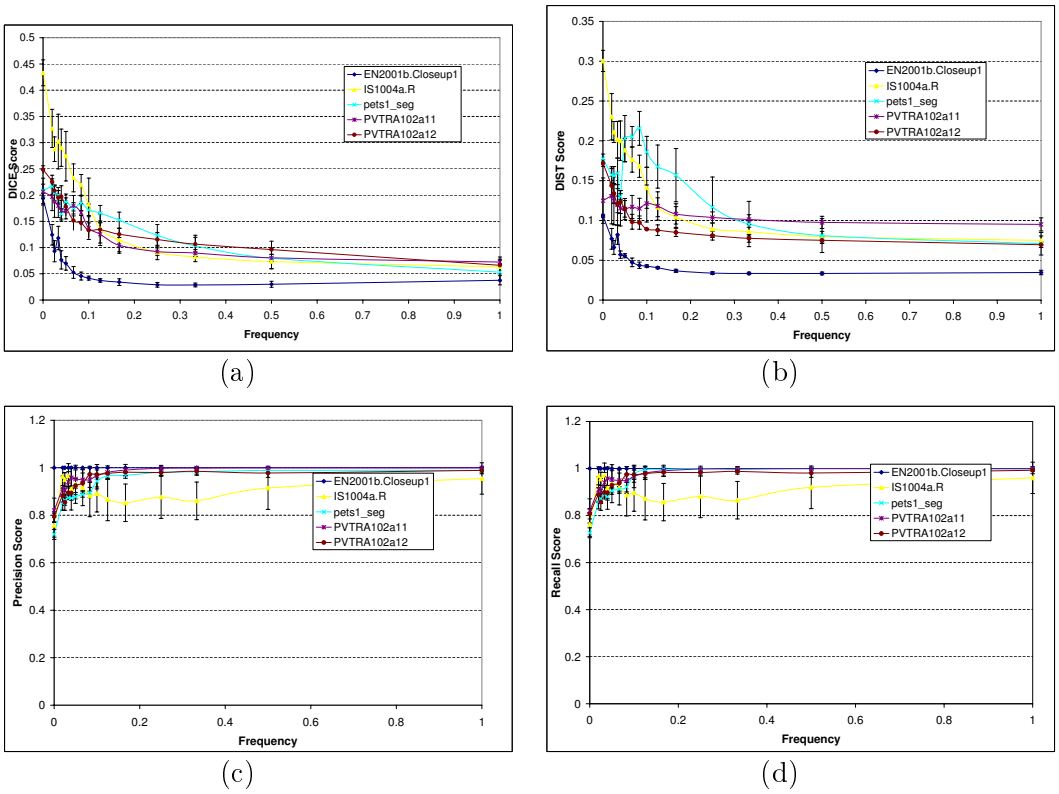


Figure 3.26: Performance improves in general as frequency of detections integrated into particle filtering increases. Scores (a) $d_{\mathcal{D}}$ (b) $d_{\mathcal{D}ist}$(c) $\bar{P}$(d) $\bar{R}$ have been calculated. This is not completely consistent thought, especially recall and precision scores for sequence S4 (IS1004a.R), as well as $d_{\mathcal{D}}$ and $d_{\mathcal{D}ist}$ scores for sequence S13 (pets1_seg).

tions taken from the ground truth at different frequencies. The parameter values used are $\alpha = 0.9$, $\beta = 0.35$ and $\gamma = 0.5$ both face and human tracking. Fig. 3.26 shows clearly that $d_{\mathcal{D}}$ and $d_{\mathcal{D}ist}$ are lower for higher detection frequencies. The left most value in all graphs indicate only particle filtering, whereas the right most value $f_d = 1$ indicates detections every frame. Further, the $\bar{R}$ and $\bar{P}$ improve with frequency. Exceptions are $\bar{R}$ and $\bar{P}$ for sequence S4 as well as $d_{\mathcal{D}}$ and $d_{\mathcal{D}ist}$ for sequence S13.

To test the integration in the real system a series of experiments were conducted. Here the parameter values differs for faces and people. For faces $\alpha_f = 0.9$, $\beta_f = 0.35$ and $\gamma_f = 0.5$ and for people $\alpha_p = 0.5$, $\beta_p = 0.1$ and $\gamma_p = 0.5$, values that have been found appropriate after extensive testing. The sequences are the same as the ones used for testing *ideal integration*. The result is displayed in Table 3.3 and the conditions were first divided

into **NOGT**, **GTIT** and **GTIO**, which stands for that no ground truth has been used and ground truth for initialization and termination as well as ground truth for initialization only. The **NOGT**, **GTIT** were divided into integration with detection (Int.) and no integration with detection (PF). The evaluation scores used are $\bar{\mathcal{D}}_R$, $d_{\mathcal{D}ist}$, $\bar{R}$ and $\bar{P}$, and a mean of 8 runs has been calculated with standard deviation within parentheses.

First, the comparison of integration vs. particle filtering alone in the **NOGT** condition for faces show that DICE and DIST scores are lower for three out of four face sequences indicating *better* correspondence between track ellipses and ground truth. Further, recall and precision are higher for the same sequences. The reason the results were better without integration for the first face sequence is that this sequence is very simple, with only one face in a not too changing posture and therefore particle filtering alone works fine for the 500 frames tracked. Similar results are obtained in the **GTIT** condition. For people tracking the scores were better with integration in the **NOGT** condition for two of three sequences, whereas they are worse in all three cases in the **GTIT** condition. Example tracks generated from the two conditions are displayed in Fig. 3.27(a-b). Better estimation is obtained with integration as can be seen in Fig. 3.28.

Further, comparison of the **NOGT** condition vs. the **GTIT** condition shows in general better performance in the **NOGT** condition. This is most likely due to lost tracks in the **GTIT** condition. Using ground-truth for initialization only (**GTIO**), gives in general lowest DICE and DIST scores as well as higher precision scores, however recall scores are lower due to a huge amount of false negatives. The result of track management was segmented tracks, but this leads to an improvement due to refresh (i.e. terminated and reinitialised tracks) as can be seen in Fig. 3.27(c-d) and Fig. 3.29.

Finally, it was also tested what happens if you change the particle filter to a simple nearest neighbour filter, where a detection determines the next state position if proximity condition of Eq. 3.16 are met. No particle filtering is done. Comparing the **NN** condition with **NOGT Int.** shows using particle filtering is better than using the nearest neighbour algorithm for 5 out of 7 sequences.

Since increased frequency of detection does not always improve accuracy a further experiment was conducted. Further, Fig. 3.27b indicate that particle filtering might work better in the short run. Therefore experiments where conducted where a maximum fre-

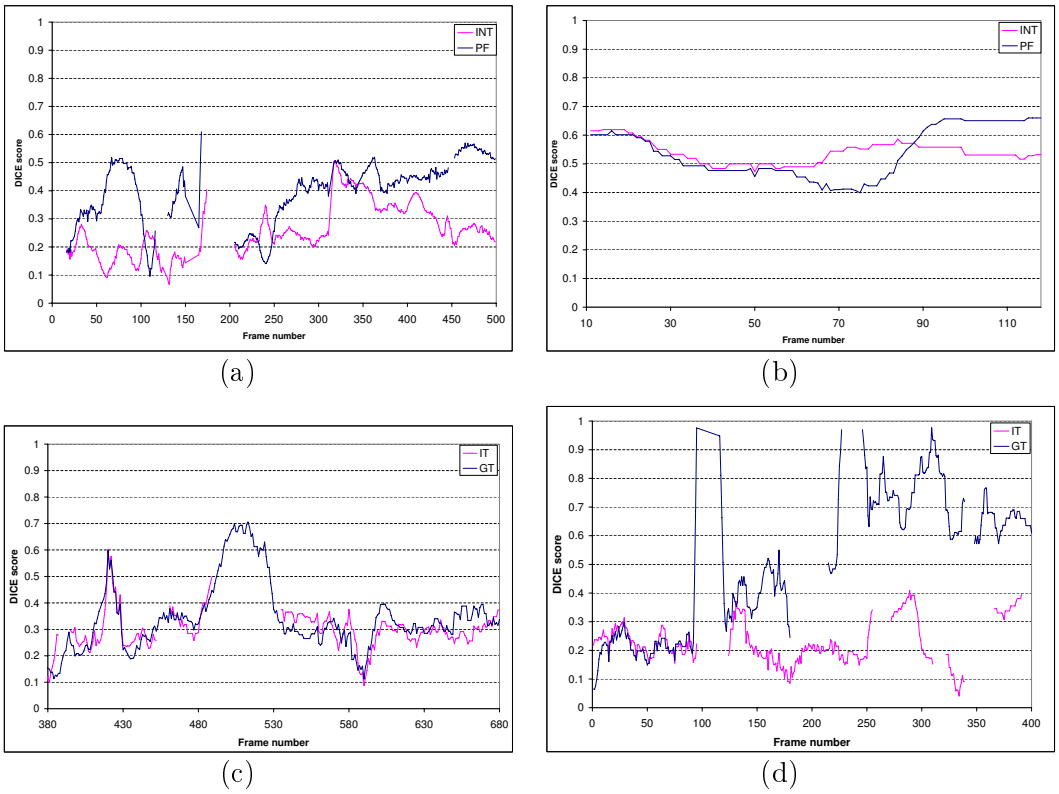
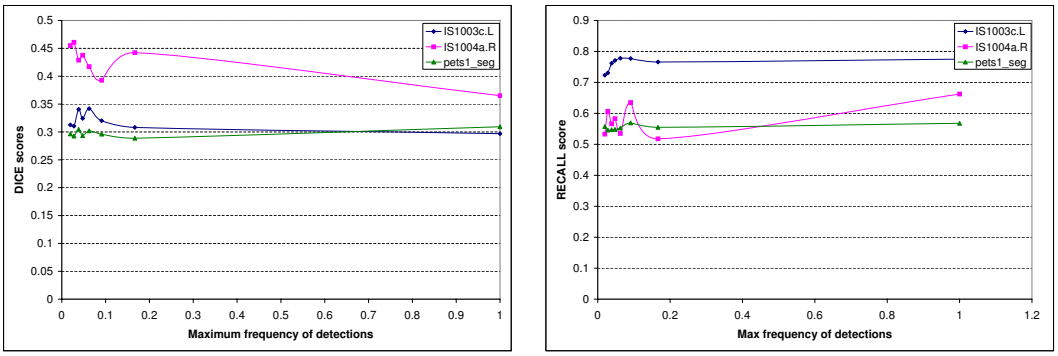


Figure 3.27: Example tracks under different conditions. (a) Integration with detection (INT) yields better results than particle filtering alone (PF), especially for face sequences. (b) Sometimes the particle filter outperforms integration in the short run, whereas integration is better in the long run. (c) One effect of track management is segmented tracks (IT). (d) In many cases thought the effect of termination and reinitialization is refresh of the track. When the track degenerates in the GT condition it is never properly recovered.

Figure 3.28: Targets are better estimated with integration (b) than with particle filtering alone (a).

(a) (b)

Figure 3.29: Frame 270 with and without track management from the tracks in Fig. 3.27a. (a) Using ground truth for initialisation and termination prevents refresh of tracks. (b) Due to termination and reinitialisation the target is estimated more accurately in the long run.



(a) (b)

Figure 3.30: Performance depending on maximum frequency of detections. (a) $d_{\mathcal{D}}$ scores. (b) $\bar{R}$ scores.

quency of detections was set by requiring a minimum gap between detections. Results of running with different maximum frequencies are presented in Fig. 3.30. In no case is the performance improving consistently with maximum frequency of detections, rather it varies.

3.5.3 Investigated applications

The applications of tracking are plenty. For example the face tracking maintains information about pose and identity. This can be used to gather face examples of people in video (see Fig. 3.31), for direct identification purposes or for storage in a database. In

		Faces					
		NOGT		GTIT		GTIO	NN
Seq.		Int.	PF	Int.	PF	Int.	Int.
1	$d_{\mathcal{D}}(\sigma_{d_{\mathcal{D}}})$	0.29(0.17)	0.17(0.01)	0.30(0.01)	0.19(0.01)		0.33
	$d_{\mathcal{D}ist}(\sigma_{d_{\mathcal{D}ist}})$	0.04(0.04)	0.08(0.01)	0.18(0.02)	0.10(0.01)		0.20
	$\bar{P}(\sigma_{\bar{P}})$	0.96(0.05)	1(0)	0.91(0.05)	1(0)		0.91
	$\bar{R}(\sigma_{\bar{R}})$	0.96(0.05)	1(0)	0.91(0.05)	1(0)		0.91
2	$d_{\mathcal{D}}(\sigma_{d_{\mathcal{D}}})$	0.27(0.01)	0.37(0.02)	0.30(0.02)	0.6(0.01)	0.20(0.02)	0.27
	$d_{\mathcal{D}ist}(\sigma_{d_{\mathcal{D}ist}})$	0.10(0.004)	0.30(0.04)	0.13(0.01)	0.43(0.002)	0.09(0.005)	0.11
	$\bar{P}(\sigma_{\bar{P}})$	0.92(0.08)	0.95(0.002)	0.94(0.02)	0.89(0.004)	0.90(0)	0.89
	$\bar{R}(\sigma_{\bar{R}})$	0.88(0.06)	0.87(0.04)	0.99(0.02)	0.94(0.02)	0.13(0)	0.96
3	$d_{\mathcal{D}}(\sigma_{d_{\mathcal{D}}})$	0.24(0.01)	0.26(0.03)	0.48(0.05)	0.48(0.04)	0.21(0.03)	0.28
	$d_{\mathcal{D}ist}(\sigma_{d_{\mathcal{D}ist}})$	0.12(0.002)	0.19(0.03)	0.26(0.04)	0.26(0.02)	0.16(0.02)	0.12
	$\bar{P}(\sigma_{\bar{P}})$	0.89(0.004)	0.63(0.007)	0.73(0.07)	0.46(0.08)	0.99(0.003)	0.73
	$\bar{R}(\sigma_{\bar{R}})$	0.71(0.02)	0.68(0.02)	0.82(0.08)	0.51(0.09)	0.13(0.01)	0.92
4	$d_{\mathcal{D}}(\sigma_{d_{\mathcal{D}}})$	0.32(0.02)	0.42(0.01)	0.32(0.02)	0.45(0.01)	0.29(0.03)	0.25
	$d_{\mathcal{D}ist}(\sigma_{d_{\mathcal{D}ist}})$	0.23(0.02)	0.37(0.01)	0.23(0.02)	0.30(0.01)	0.20(0.04)	0.17
	$\bar{P}(\sigma_{\bar{P}})$	0.81(0.05)	0.61(0.05)	0.84(0.07)	0.76(0.05)	0.98(0.04)	0.25
	$\bar{R}(\sigma_{\bar{R}})$	0.74(0.06)	0.40(0.02)	0.85(0.07)	0.77(0.5)	0.40(0.12)	0.59
		People					
		NOGT		GTIT		GTIO	NN
Seq.		Int.	PF	Int.	PF	Int.	Int.
7b	$d_{\mathcal{D}}(\sigma_{d_{\mathcal{D}}})$	0.21(0.04)	0.19(0.01)	0.37(0.02)	0.23(0.01)	0.16(0.01)	0.26
	$d_{\mathcal{D}ist}(\sigma_{d_{\mathcal{D}ist}})$	0.17(0.01)	0.14(0.01)	0.41(0.03)	0.20(0.02)	0.13(0.01)	0.22
	$\bar{P}(\sigma_{\bar{P}})$	0.97(0.01)	0.95(0.01)	0.32(0.01)	0.39(0.01)	1.0(0.001)	0.38
	$\bar{R}(\sigma_{\bar{R}})$	0.43(0.02)	0.49(0.03)	0.60(0.02)	0.72(0.02)	0.15(0.01)	0.95
8b	$d_{\mathcal{D}}(\sigma_{d_{\mathcal{D}}})$	0.42(0.02)	0.43(0.02)	0.33(0.02)	0.25(0.02)	0.15(0.02)	0.37
	$d_{\mathcal{D}ist}(\sigma_{d_{\mathcal{D}ist}})$	0.21(0.01)	0.24(0.01)	0.22(0.05)	0.10(0.01)	0.10(0.02)	0.16
	$\bar{P}(\sigma_{\bar{P}})$	0.91(0.01)	0.91(0.01)	0.41(0.04)	0.97(0.05)	1(0)	0.74
	$\bar{R}(\sigma_{\bar{R}})$	0.74(0.01)	0.75(0.01)	0.46(0.04)	0.37(0.05)	0.22(0.03)	1
13	$d_{\mathcal{D}}(\sigma_{d_{\mathcal{D}}})$	0.36(0.02)	0.39(0.05)	0.28(0.01)	0.26(0.01)	0.28(0.02)	0.38
	$d_{\mathcal{D}ist}(\sigma_{d_{\mathcal{D}ist}})$	0.20(0.01)	0.25(0.04)	0.18(0.02)	0.18(0.003)	0.18(0.02)	0.19
	$\bar{P}(\sigma_{\bar{P}})$	0.76(0.01)	0.75(0.01)	0.59(0.01)	0.72(0.01)	0.98(0.004)	0.51
	$\bar{R}(\sigma_{\bar{R}})$	0.65(0.01)	0.63(0.02)	0.68(0.01)	0.78(0.01)	0.55(0.03)	0.83

Table 3.3: Comparison of tracking performance

an example based indexing application, each face example could be linked to the video sequence, where that person appears.

Sampled trajectories of two face sequences are displayed in Fig. 3.32. It is possible to tell quite a lot about the sequence by analysing the trajectories. One possibility is to use this information for video shot classification, for example separating meeting and surveillance shots. This can also be used for event detection and to extract information about the camera. For example analysis might reveal that two or more people are meeting

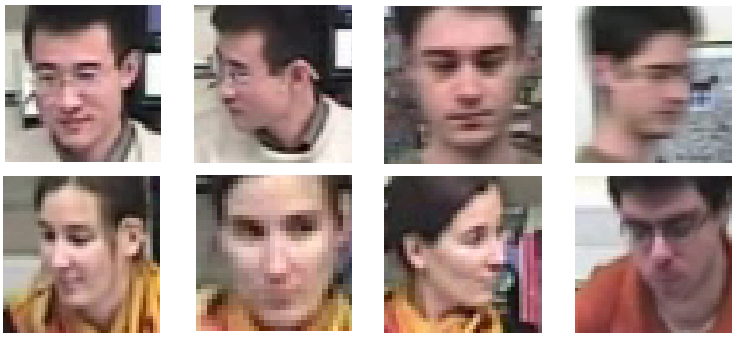
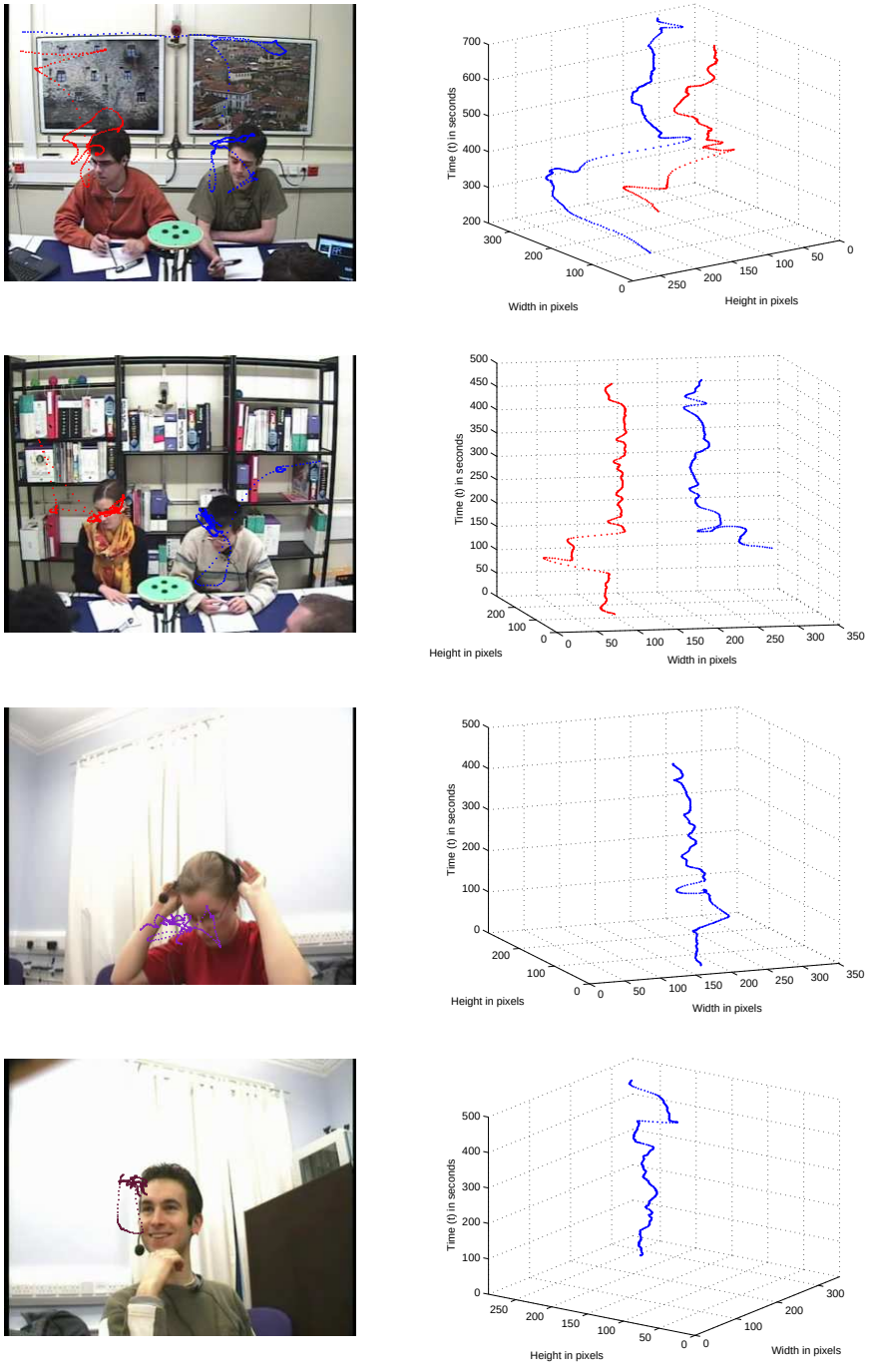


Figure 3.31: Examples of extracted faces of frontal, left and right profile.



(a) (b)

Figure 3.32: Results of face tracking illustrated as tracks projected on an example frame (a) and as tracks evolving in time (b). Different colours are assigned to each track.

(a) (b) (c) (d)

Figure 3.33: Example of a passage monitoring application. When people approach the camera they are tracked (a-c). Just before passage the face is tracked (c) and face examples extracted (d).

for a conversation in a particular part of a video sequence. Another application area is surveillance of passage, through doors or gates. Using people tracking and face tracking in combination can be used to first detect approaching persons, then locate faces as persons approach, and at that stage extract several face examples of passing persons as illustrated in Fig. 3.33. Notification that a person is arriving can be sent to a security guard and enlarged presentation of the shots to the guard would enable identification of the person passing trough.

3.5.4 Moving object tracking

The four and five dimensional moving object trackers have been evaluated with MOTP and MOTA scores with different parameter settings. The results are divided into people and vehicle tracking. For four dimensional tracking results (see Table 3.4) for people tracking is first that it is unclear which value is best for α, $\beta = 0.1$ gives highest value in two cases and $\gamma = 0.1$ is definitely the best choice. For vehicle tracking $\alpha = 0.5$, $\beta = 0.25$ and $\gamma = 0.25$ is clearly best.

Further, some other aspects of the four dimensional tracker has been tested. First, an experiment to test what happens if you only use the area covered by change pixels to build and update the colour histogram model (condition Blob in table 3.4). It turns out that this reduced performance in both people and vehicle tracking. Then, it was tested if treating border object specifically improves performance, and it turns out it does not for either people or vehicle tracking. Finally, results where obtained with and without mask. The mask improves accuracy scores (A) in 7 cases out of 8, whereas it reduces precision scores (P) in 6 out of 8 sequences.

		People							
		5		6		7a		8a	
		P	A	P	A	P	A	P	A
α	0.3	0.60	0.01	0.50	0.01	0.58	0.24	0.57	0.24
	0.5	0.62	0.10	0.48	0.001	0.58	0.24	0.55	0.25
	0.7	0.57	0.09	0.49	-0.01	0.59	0.23	0.57	0.25
β	0	0.57	0.08	0.50	0.008	0.55	0.22	0.55	0.25
	0.1	0.54	0.09	0.47	0.024	0.58	0.24	0.57	0.24
	0.25	0.62	0.10	0.48	0.001	0.58	0.24	0.55	0.25
γ	0.1	0.70	0.11	0.57	0.08	0.66	0.19	0.66	0.20
	0.25	0.62	0.09	0.53	0.05	0.60	0.21	0.62	0.23
	0.5	0.62	0.10	0.48	0.001	0.58	0.24	0.55	0.25
	0.75	0.57	0.10	0.44	0.02	0.55	0.25	0.54	0.25
Blob	Y	0.49	0.09	0.44	0.10	0.47	0.20	0.48	0.23
	N	0.62	0.10	0.48	0.001	0.58	0.24	0.55	0.25
Border	Y	0.48	0.008	0.46	0.11	0.48	0.37	0.51	0.44
	N	0.62	0.10	0.48	0.001	0.58	0.24	0.55	0.25
Mask	Y	0.45	0.15	0.47	-0.11	0.55	0.27	0.55	0.44
	N	0.62	0.10	0.48	0.001	0.58	0.24	0.55	0.25
		Vehicle							
		9		10		11		12	
		P	A	P	A	P	A	P	A
α	0.3	0.65	0.18	0.62	-0.31	0.51	0.06	0.61	0.20
	0.5	0.65	0.20	0.66	-0.30	0.54	0.09	0.61	0.20
	0.7	0.63	0.20	0.58	-0.28	0.50	0.07	0.58	0.21
β	0	0.65	0.19	0.56	-0.46	0.47	0.06	0.58	0.21
	0.1	0.62	0.14	0.62	-0.31	0.54	0.07	0.60	0.21
	0.25	0.65	0.20	0.66	-0.30	0.54	0.08	0.61	0.20
γ	0.1	0.66	0.08	0.67	-0.24	0.63	0.06	0.68	0.09
	0.25	0.68	0.17	0.59	-0.37	0.63	0.05	0.63	0.19
	0.5	0.65	0.20	0.66	-0.30	0.54	0.09	0.61	0.20
	0.75	0.66	0.18	0.60	-0.31	0.60	0.08	0.61	0.21
Blob	Y	0.56	0.15	0.52	-0.42	0.43	0.07	0.58	0.20
	N	0.65	0.20	0.66	-0.30	0.54	0.09	0.61	0.20
Border N	0.56	0.16	0.47	-0.35	0.54	0.04	0.57	0.21	
	Y	0.65	0.20	0.66	-0.30	0.54	0.09	0.61	0.20
Mask	Y	0.45	0.15	0.47	-0.11	0.55	0.27	0.55	0.44
	N	0.65	0.20	0.66	-0.30	0.54	0.09	0.61	0.20

Table 3.4: Comparison MOTA (A) and MOTP (P) scores using different parameter settings for four dimensional moving object tracking.

When it comes to the five dimensional tracker it is difficult to say which α value is best since it differs according to sequence and tracker. Perhaps spread of particles around the detection area is not improving performance since almost the same measure is already expressed into the likelihood. Further, the selection of state from particles is modulated in the SELECTION condition. The conditions are BEST (B), SELECTED AVERAGE (SA) and AVERAGE (A). In the BEST condition the particle with highest likelihood is used as state, in the SELECTED AVERAGE condition the average of the particles with weight higher than the total average is selected and in the AVERAGE condition the total average is used. For people SELECTED AVERAGE has more top scores than the other conditions. For vehicles AVERAGE is the best choice according to the results.

		People							
		5		6		7a		8a	
		P	A	P	A	P	A	P	A
α	0	0.54	-0.21	0.47	0.04	0.47	0.29	0.51	0.21
	0.25	0.55	-0.18	0.46	0.03	0.46	0.32	0.51	0.21
	0.5	0.55	-0.17	0.47	0.03	0.47	0.31	0.50	0.21
Selection	B	0.53	-0.22	0.48	0.07	0.45	0.30	0.49	0.20
	SA	0.55	-0.18	0.46	0.03	0.46	0.32	0.51	0.21
	A	0.53	-0.21	0.47	0.04	0.47	0.29	0.51	0.21
		Vehicle							
		9		10		11		12	
		P	A	P	A	P	A	P	A
α	0	0.66	0.54	0.51	-0.33	0.66	0.12	0.63	0.64
	0.25	0.63	0.62	0.55	-0.37	0.68	0.21	0.65	0.60
	0.5	0.64	0.58	0.51	-0.25	0.70	0.20	0.60	0.63
SELECTION	B	0.60	0.59	0.57	-0.31	0.70	0.15	0.58	0.53
	SA	0.63	0.62	0.55	-0.37	0.68	0.21	0.65	0.60
	A	0.68	0.50	0.58	-0.28	0.72	0.12	0.62	0.62

Table 3.5: Comparison MOTA (A) and MOTP (P) scored using different parameter settings for five dimensional moving object tracking.

3.6 Conclusions

The presented tracking framework does not only annotate video in terms of object and trajectories, it is also able to produce additional information about the tracked object and extract pictorial examples. Moreover, it is intended to be able to track other objects in addition to faces, people and moving objects. The requirement for tracking any object type is only an appropriate detector for that particular object type. A system that can track and classify a larger amount of objects has potential to be utilised in semantic annotation of video. With enough information, in the end, complete story lines could be produced, describing person interaction and other important events. With a face recognition module, video could be annotated semantically with identity information of the appearing persons.

It has been shown that segmentation increase robustness of detection. Still the main limitation of the tracking algorithm is the accuracy of the detectors. To improve the detection of faces, which is needed primarily to reduce false positives, solutions are first to work more with colour segmentation. Possibly the hue saturation colour space is better for this purpose than the YC_bC_r space. It would also be interesting to analyse the behaviour of system by adding more feature detectors for eyes, mouth and nose, to use the contrast contour of the faces as input to the algorithm, or to take advantage of the fact that the face is connected to a body, e.g. to detect shoulders with edge detection.

One of the limitations of the used person detector is that it fails to detect people against a dark background. Possible solutions to this are to use edges for the Adaboost training or to train several differently tuned detectors and integrate the output. There is also a

problem of too large people detections, which possibly could be fixed by adjusting the bounding box with either edge content in the image, or the bounding boxes of the motion segmentation.

Integrating detections with particle filtering is obviously not only limited to using Adaboost trained detectors. With the moving object tracker it is shown that the framework can be used with other types of detections as well as be adjusted to work in five dimensional mode.

The initialisation and termination rules presented does filter out lots of false tracks, first by not starting tracks on false positives outputted from the detector and second to stop tracks once the object disappears from view. There are still limitations to the rules since all false positives are not removed, and tracks are segmented.

The presented work differs from previous work first since any type of object can be tracked, and in particular face, pedestrians and moving objects have been tracked. Further, several track management rules have been implemented.

Chapter 4

High- and low-level visual attention

4.1 Introduction

Modelling visual attention given current knowledge and processing powers in a standard PC is a difficult task. Previous models have mostly focused on low-level saliency, sometimes with top-down modulation of saliency and the inclusion of context as a factor. The interplay between the observers' goals, expectations, ideas and the outer visual world is still to be depicted.

In watching video sequences imaging meeting scenarios, the expectation is most likely to see faces and the goal is to follow the interplay between humans in the meeting and in a real scenario to interact with other participants. When it comes to surveillance scenarios you would expect to see people, and in the context of traffic, cars. Thus the addition of high–level features introduces top–down factors in the interplay between observer and stimuli.

As proposed, top-down influence could simply be an additional contribution to a final saliency map encoding both bottom-up and top-down information before a motor action is selected, and then the most salient feature is scanned for relevant information. Another possibility is that bottom-up and top-down factors influence the selection of focus points in parallel, where one of them takes control in a competitive way, for example reflexive saccades initially and volitional saccades at a later stage. In both cases, location of the target is the most interesting information, and either before top–down saliency arises or during competition between several top-down informations the type of object is of impor-

tance. Since both types of information in the trackers is extracted one can easily integrate this information with low–level saliency.

Considering only low-level features do not discriminate between background and foreground. For this reason, models relying on low-level features will not only allocate the majority of fixation on the interesting objects part of the scene, where people presumably look, and will allocate far too much attention on the background. Further, on videos objects are moving. Previous measurements have illustrated that flicker and change attract far more attention than spatial features[1]. The object trackers described in chapter 3 have been developed to allow for incorporation of these high–level features in models of visual attention. With the developed tracker tools, moving high–level features as such are combined with the saliency map generated using low-level features only. The end result is a video where saliency from low- and high-level features are added up pixel-wise. These are further processed to generate a scan-path with the IOR mechanism described in section 2.3.3, which has been utilised with some success in the past[196] using mainly low-level feature analysis. Another model, developed by the author, utilises the output from tracker modules more directly, by using the centre of objects as the candidate targets of attentional fixation.

It must be noted that the detection of high–level objects can in part or even completely be a bottom-up process. The detection of faces could possibly be special in such a way as the detection is made without expectations to see faces and without prior exposition to faces. The latter is the case since infants look for faces and direct their gaze towards them early in development. Because of the importance of faces there could be hard–coded neural pathways to assign saliency to areas where faces appear. Further, bottom–up grouping processes can identify objects without top–down information. Here is an area where the study of visual attention in the context of scene interpretation is required.

First in this chapter it is outlines how saliency can be generated from low-level feature, high-level features and the combination as well as variations. These saliency models are validated with eye-tracking experiments. Also, in this chapter the collected eye-traces will be described qualitatively and quantitatively as well as one model based on winner-take-it-all and inhibition-of-return. Furthermore, an object based attention module based on these data will be described.

4.2 Saliency with high- and low-level features

4.2.1 Low-level features

Primary bottom-up processes work similar to simple image processing techniques like edge and corner extraction. More complicated processes involve both bottom-up and top-down like processes at different levels of abstraction for any type of visual processing of stimuli i.e. object detection, visual search or scene interpretation. It is believed that a primary mechanism exists to swiftly direct attention towards features like corners and crosses and a low-level feature extractor is needed to mimic this mechanism. Much work has been done in this area before with different models[1, 13].

In this work, low-level features are extracted with the Itti et. al. model[1]. Twelve neuronal features extracts colour contrast (red/green and blue/yellow, separately), temporal flicker (onset and offset of light intensity, combined), intensity contrast (light-on-dark and dark-on-light, combined), four orientations (0°, 45°, 90°, 135°), and four motion energies (up, down, left and right). Centre-surround differences are then calculated as differences between dyadic pyramid scale levels. This yields in total 72 feature maps. Finally, there is within-scale, within-feature and across-scale competition in each feature map before they are added up into one saliency map.

4.2.2 High-level features

The limitation of using low–level features alone is that processes on a higher abstraction level are involved in the selection of interest points. In processes like scene interpretation, higher level patterns and objects must be dealt with. Locations and outlines of objects can be found with low-level processes[10]. As these kind of methods are limited in their success in this work we have relied on the tracker tool[14] presented in 3. The types of objects extracted are three: faces, pedestrians and unclassified moving objects. The tracker utilises particle filtering integrated with detection and is able to track faces and pedestrians, moving object during the entire sequence.

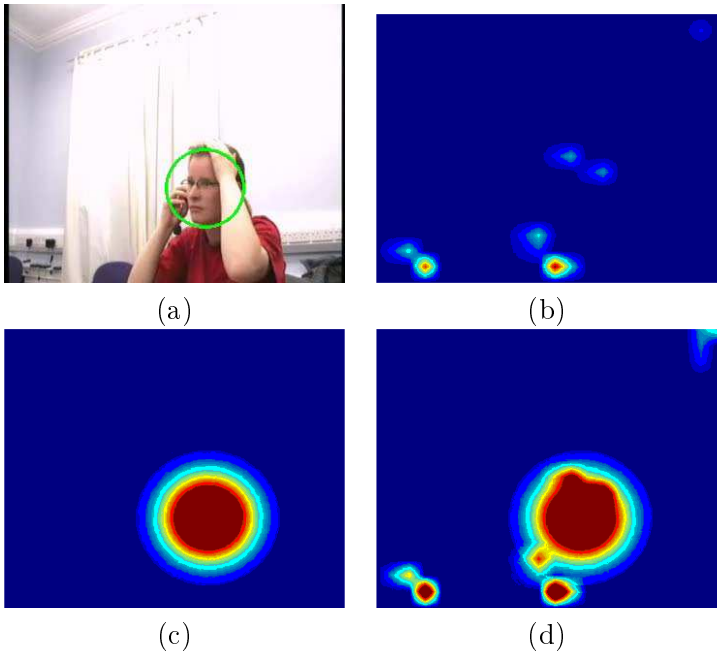


Figure 4.1: Example of saliency generated from image 326 from sequence S2. (a) Image, (b) low-level features only, (c) high-level and (d) low-level and high-level combined.

4.2.3 Combination

With the object tracker tools, described herein, moving high–level features as such are combined with the saliency map generated with low-level features only. The end result is a video where saliency with low- and high-level features with the current model (Fig. 1.1) are linearly combined pixel-wise.

To add up low-level and high-level features the following procedure is followed. First faces, pedestrians and moving objects are tracked. After that low-level features are extracted. Finally, a tool developed to add up low- and high-level features is used which projects Gaussians corresponding to the tracks generated by each tracker onto the low-level features only map. Examples of generated saliency maps are displayed in Fig. 4.1-4.4.

4.2.4 Variation

Variations of the combined map have been generated to optimise the match between saliency map and real eye-tracking data. The first variation is to use a bridge between high-level objects as illustrated in Fig. 4.5. Here a quadrilateral bridge is generated with

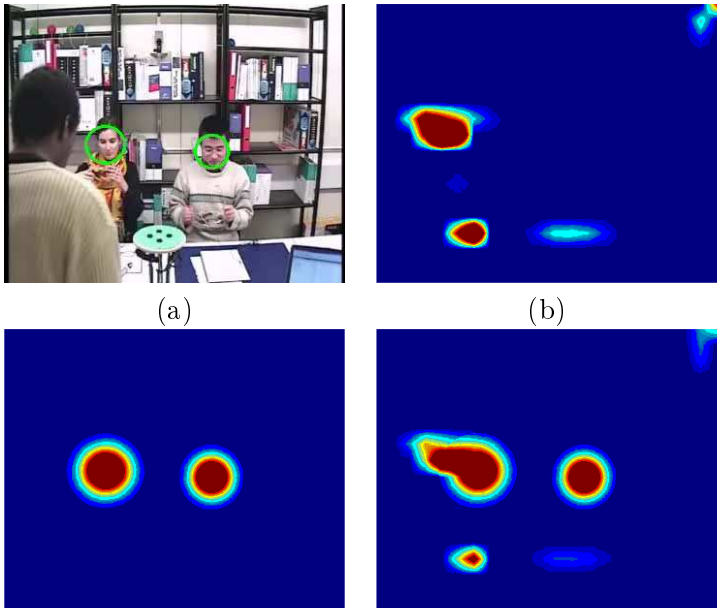


Figure 4.2: Example of saliency generated from image 300 from sequence S3. (a) Image, (b) low-level features only, (c) high-level and (d) low-level and high-level combined.

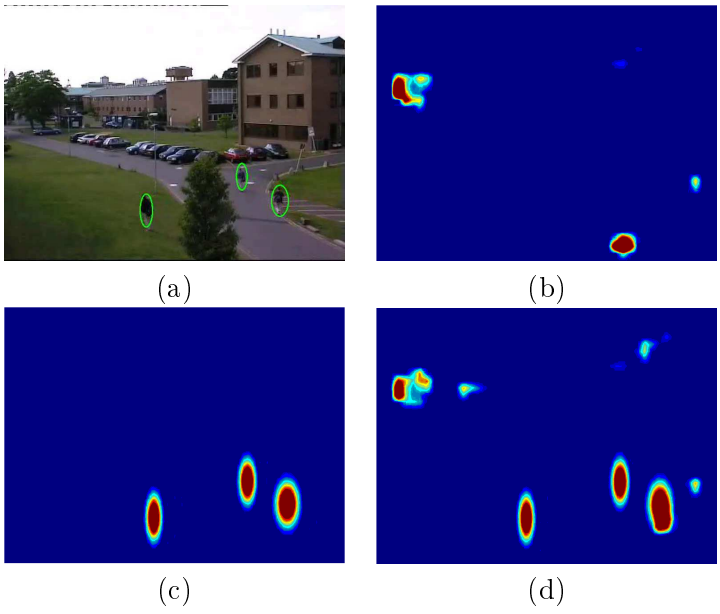


Figure 4.3: Example of saliency generated from image 87 from sequence S13. (a) Image, (b) low-level features only, (c) high-level and (d) low-level and high-level combined.

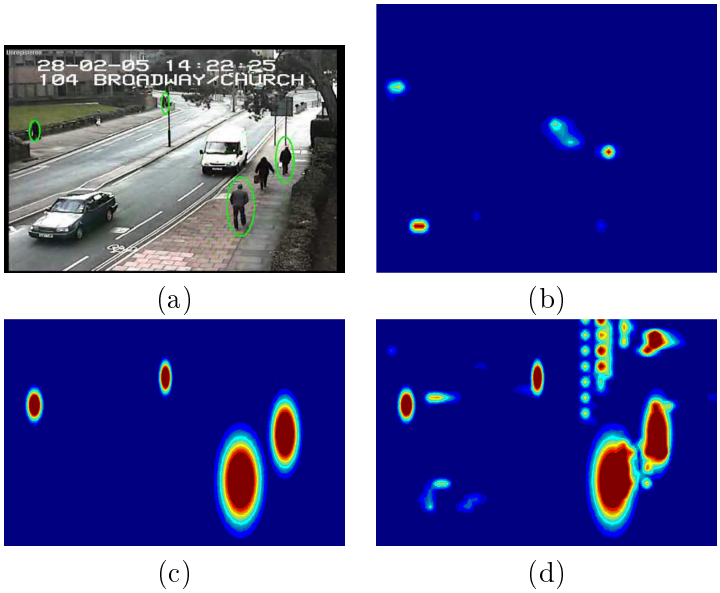


Figure 4.4: Example of saliency generated from image 231 from sequence S8b. (a) Image, (b) low-level features only, (c) high-level and (d) low-level and high-level combined.

the saliency value 100 in the interior of the polygon. The quadrilateral is created with two sides perpendicular to the line between detection centres with a length $l_i = h_i + w_i$, where w_i and h_i are the width and height of the respective detection. Another variation is to only include low-level features on top of objects as illustrated in Fig. 4.6 (abbreviated OOO for only on objects). Finally the saliency maps are also tested against fixed saliency maps with the same image throughout the video (see Fig. 4.7)

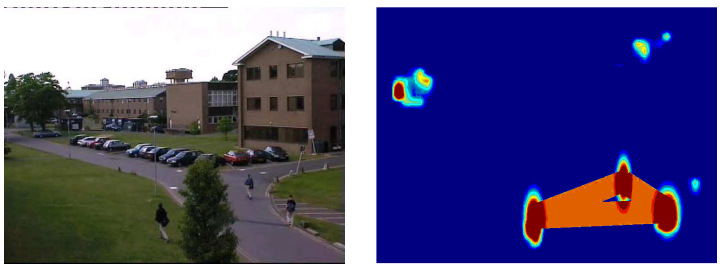
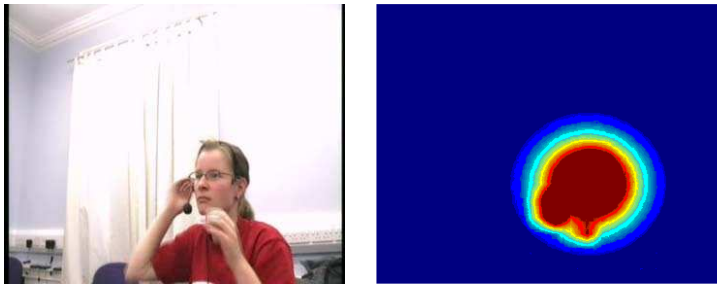
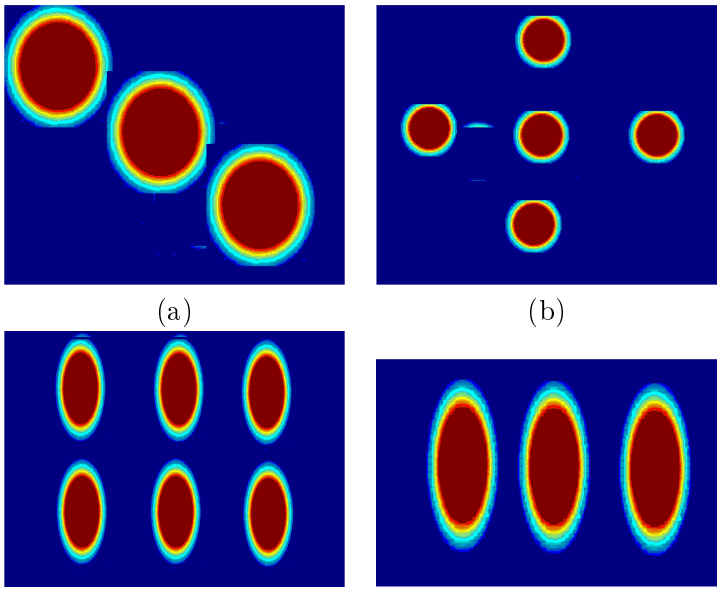


Figure 4.5: (a) Original frame. (b) Saliency map with bridge between objects.



(a) (b)

Figure 4.6: (a) Original frame. (b) Saliency map with low-level features added only on objects.



(a) (b)

(c) (d)

Figure 4.7: Manually created saliency maps for the Fake condition. (a) Used for sequence S2. (b) Used for sequence S3. (c) Used for sequence S13. (d) Used for sequence S7b and S8b.

Figure 4.8: Experimental setup.

4.3 Eye-tracking

4.3.1 Experimental setup

Eye-tracking is necessary to compare saliency models with real fixations. For the collection of eye-traces data a desktop computer equipped with an IR light emitter as well as an IR camera has been utilised. Subjects have been seated in a comfortable chair with their chin resting in a chin-rest and the eyes located approximately 62cm from the screen. The entire set-up is illustrated in Fig. 4.8. The hardware utilised is called the Eyegaze Analysis System and is developed by LC Technologies Inc. It consists of the computer with the camera attached to the $15in$ LCD monitor. A further black and white monitor is connected directly to the camera to show the captured image. A small IR light emitting diode is attached to the camera. The camera captures at a rate of 60Hz. The Eyegaze System uses the Pupil-Centre/Corneal-Reflection method to determine the eye's gaze direction.

Eye-tracking software has been utilised to display the video. It consists of a program that displays videos and records the eye-fixations in a text file. The timing and position is written on each line. Before recording the eye-tracker must be calibrated. This consists of instructions given to the subject to fixate on a fixation point on a black screen that moves to 10 different locations. Fixation points that were missed are redisplayed until all points have been fixated by the subject reasonably well. The calibration accuracy is typically $0.10 - 0.20in$. The sequences used for the sampling are described in Table 4.1.

Dataset	Seq. nbr.	Sequence	Frames
AMI	S1.	EN2001b.Closeup1	100–600
	S2.	EN2001b.Closeup4	1–500
	S3.	IS1003c.L	1-500
	S4.	IS1004a.R	250-750
VACE	S5.	PVTRA102a09	500–3001
	S6.	PVTRA102a10	3007–5701
	S7a.	PVTRA102a11	1003–3010
	S7b.	PVTRA102a11	1–500
	S8a.	PVTRA102a12	3000–5107
	S8b.	PVTRA102a12	1000–1500
	S9.	CMU_20050912–0900.cam3	20005–23605
	S10.	EDI_20050216–1051.cam1	11800–15300
	S11.	EDI_20050216–1051.cam3	25000–30000
	S12.	VT_20051027–1400.cam2	75000–76200
PETS	S13.	Camera1	2045–2545

Table 4.1: Short information about video sequences used for quantitative measures

4.3.2 Procedure

Before the experiment starts instructions are given to the subject:

> Two types of videos are shown. 5 sequences are surveillance sequences that contain pedestrians and moving vehicles. 5 sequences are meeting scenarios containing talking people. Watch the surveillance sequences as if you are doing a surveillance task monitoring the events on the video. For the meeting scenarios pretend that you are a part of a teleconference, equipped with headphones and a microphone. Before each video there is a calibration, with instructions given in text on the screen.

For each video the eye-tracks are recorded preceded by the calibration, and then the playback of the video is started. The output is a text file for each video containing the timing and position of recorded fixation. The meeting scenarios are sequences S9, S2, S3, S4 and S12 and the surveillance scenarios are sequences S13, S5, S6, S7b and S8b, played in the given order.

4.4 Saliency based on fixations

To compare the automatic model with real data, saliency based on several measurements (see Fig. 4.9-4.10) have been used. This saliency map is constructed by adding up Gaussians around each fixation point into one saliency map with a standard deviation of 10 pixels. This was selected to cover a reasonable angle of focus according to the spotlight model of attention as well as the limitation of accuracy of the eye-tracker which is about $0.1-0.15in$.

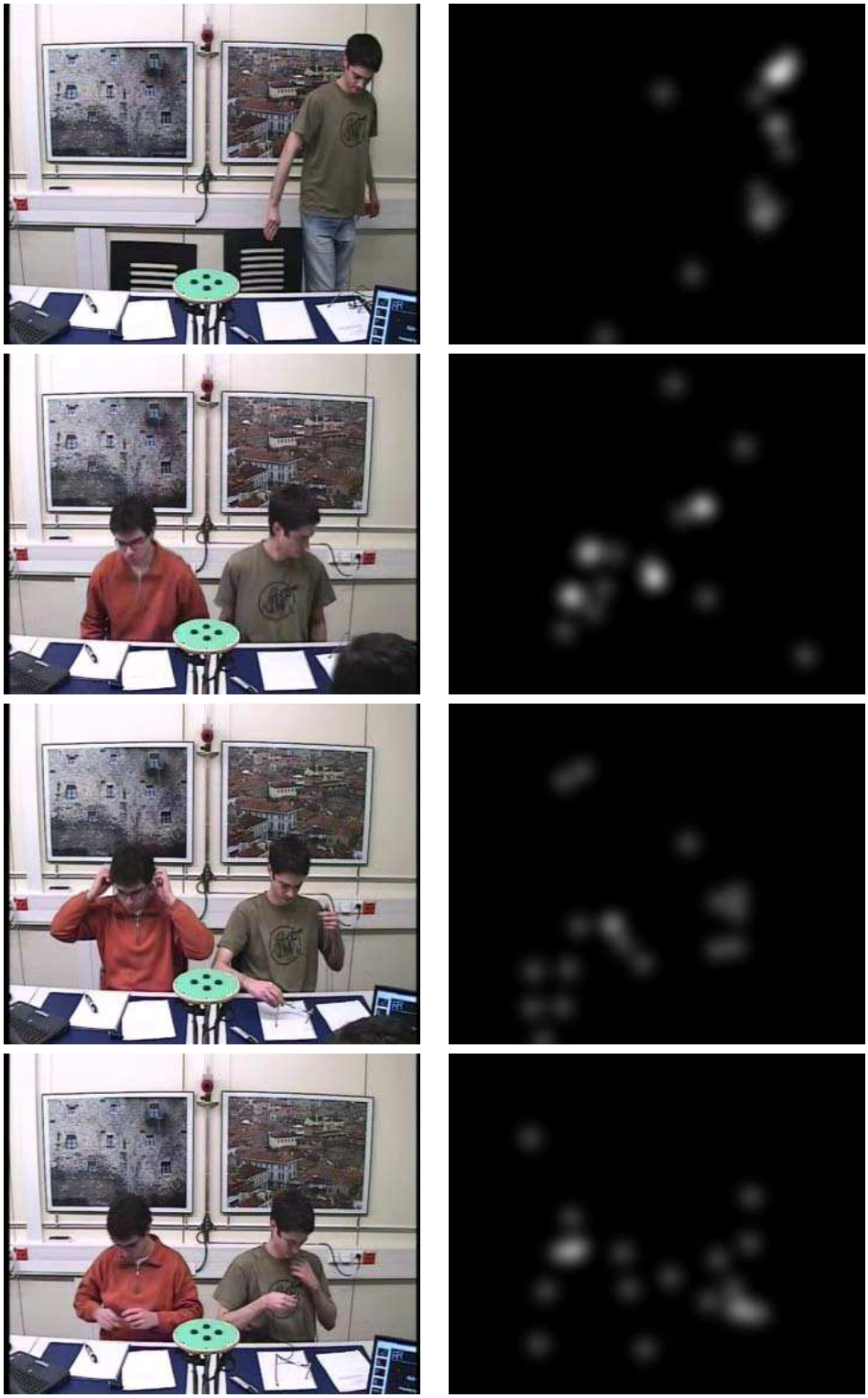
4.5 Eye-tracks based on winner-take-it-all

A simple program that generates tracks on images as well as video have been developed that utilises the winner-take-it-all and IOR mechanism to generate a scan–path on videos similar to previous models[196]. The purpose is to test the developed model in plausibility of generated eye-traces as well as the generation model itself. Contrary to previous work we applied it on videos instead of still images. The problem of using this simple mechanism is obvious for two reasons. First, fixations on for example faces are continuous on one area for extended periods of time. Second, fixations on moving objects follow the object in smooth pursuit. For this reason only traces generated on images are displayed in Fig. 4.11.

The selection of fixation point is first done with a winner-take-it all mechanism that chooses the brightest point in the saliency map of that frame for fixation. Once a point has been fixated an IOR mechanism is applied to an area around the fixation, with a Gaussian added to an inhibition map. In subsequent frames selection of fixation point is based on the brightest point in the product of the saliency map and the inhibition map. As time pass the inhibition map is relaxed, i.e. values are decremented until they reach zero.

4.6 Characterisation of collected data

To be able to qualitatively and quantitatively describe eye-tracking data with respect to stimuli and internal states of mind, would be to solve the problem of automatically generating eye-traces. With generalizability of such description, a model of visual attention would be obtained.

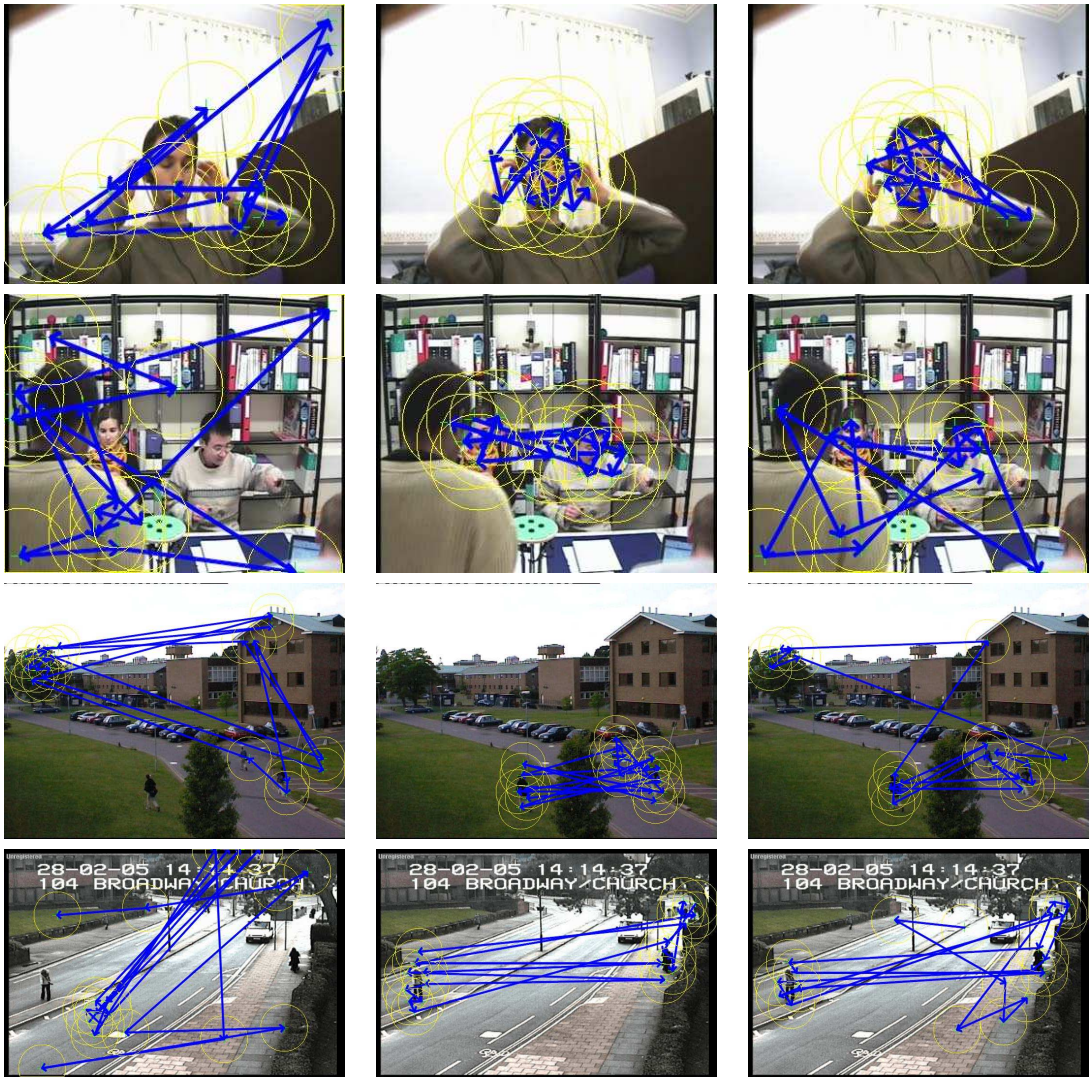


(a) (b)

Figure 4.9: Sample results of saliency mapping from eye tracking data of 20 persons. (a) Original frames from top to bottom: 85, 224, 368 and 432 from sequence S4. (b) Saliency type maps for each frame generated by projecting Gaussians from 20 subjects on eye-fixation points.

(a) (b)

Figure 4.10: Sample results of saliency mapping from eye tracking data of 20 persons. (a) Original frames from top to bottom: 33, 11, 324 and 447 from sequence S5. (b) Saliency type maps for each frame generated by projecting Gaussians from 20 subjects on eye-fixation points.



(a) (b) (c)

Figure 4.11: Here trajectories have been automatically generated on different images from top to bottom. (a) From low-level features only. (b) From high-level features only. (c) From the combination of features.

Qualitative description

A qualitative description of eye–traces can give guideline in building theoretical models by capturing the types of processes that underlie gaze patterns. Unfortunately, qualitative research methods can suffer from lack of objectivity and might be afflicted by the authors personal beliefs etc. Especially in example based presentations an author could possibly put forward only examples that confirm a given hypothesis masking a true interpretation of the results. Perhaps the theme of object based attention in this work might just affect such conclusions.

In an attempt to understand the basic mechanisms I have written down notes on the different types of scenarios in Table 4.2 and Table 4.3. I will here try to summarize these and to some extent discuss the results. This involves two components of phenomenological analysis. First, in making the "transcriptions" of the eye-traces and, second, in summarizing, attempting to derive the essence of visual attention.

When it comes to meeting scenarios people tend to look at faces, in particular eyes, mouth and nose. People also look at items on tables, walls, on the floor, and look at hands. Moreover, there are more advanced patterns in that people follow conversation by following the gaze of meeting participants, and watching the one who is speaking at the moment. So, an eye for social interaction seems to be fairly noticeable. Further, people look at items, and if that is because of spatial properties like contrast content or that these are physical objects, that could be of interest to the observer in interpreting the scene or both, is a question left unanswered. People also look at objects participants hold.

In surveillance scenarios participants look at pedestrians, moving cars and cyclists. Many times the gaze is directed toward the centre of the object, however on pedestrians gaze is often directed towards the head and occasionally towards the feet. On cars it happens that people look in front of the car. People tend to shortly fixate areas with strong contrast for example occupied by poles. Further they follow edges with their gaze and look at windows and corners.

The major conclusion that can be drawn from this is that people look at faces, and at moving objects. Further, not surprisingly people look at social interactions, possibly trying to figure out the social/communicative significance of these events. Also people look at spatial events like poles and curb edges. Finally objects (e.g. paper) seem to attract

Subject No.	Observations
1	mouth, centre of head, following gaze. Ear, nose, eyes. Items on table
2	items on table, follow gaze, attention to particular social interaction, eyes, mouth, centre of head, hands,papers on table, centre of head, hands
3	centre of head, items on table, following conversation, eyes, nose, mouth, items on background with high frequency, browsing background, faces, papers on table, browse background,hands
4	centre of head, following gaze, shifting between persons, hands head
5	Tend to watch faces and hands. Especially following the gaze of meeting participants. Some in centre of face but some tend to lie a bit outside of the face. Eye nose ear and sometimes centre of body. With two faces, fairly fast switches between faces. Items on table.
6	Eyes, hands, item in hands, mouth, item on wall, hands, follow gaze to item/person, centre of head
7	head, following gaze, eyes, centre of head, mouth, shifting between heads, hands, head and body scanning, centre of face, movement of face, hands
8	browsing the scene, follow centre of head, items on wall etc.
9	Head follow gaze, following conversation, centre of head body, shift between heads, heads, hands, heads, hands, items on table, suit, hair, eyes, heads, suits, papers on table
10	Follow gaze, head, items on table, items on floor, eyes, forehead

Table 4.2: Observations of eye-traces on meeting scenarios.

attention and not the background (e.g. table), possibly due to their spatial properties and possibly to due to object as a meaningful conceptualisations of the scene in terms of relevance to the beholder.

4.6.1 Classification of eye–traces

The distinction between saccades and smooth–pursuit on dynamic stimuli is protrudent in eye-gaze patterns. To be able to quantitatively describe saccades and smooth pursuit a classification method has been developed, based on speed. Here the speed is calculated at each sample point. If speed in normalised device coordinates ($width = height = 1.0$) per second is above 0.5, the sample is considered a part of a saccade and else-wise part of smooth pursuit. The tracks are then divided into saccades and smooth pursuit, making it possible to sample the duration in time of each. Further, speed of saccade is sampled, with speeds above 10000 pixels per second in the data are considered outliers and is discarded.

Subject No.	Observations
1	centre of person, cyclist, standing cars, moving cars, shift between centre of pedestrians, feet, head, following walls,centre of person head of person, poles, in front of moving car
2	pedestrians, moving cards, cyclist poles,heads of pedestrians, feet of pedestrians edges, moving car, switching between pedestrians and moving cars as well as one edge
3	heads, windows, moving car, poles, cyclists, bus, head, moving cars, centre of cars, in front of cars, corners on cars, centre of moving cars
4	browsing of houses, parked cars, house surroundings,head of pedestrian, wall, centre of pedestrian, head, feet, centre of moving bus, cars,centre of pedestrian, centre of moving car, head, pole, corner of car, centre of pedestrian, corner of wall
5	heads, corners edges, poles, pedestrians, moving cars, in front of moving cars
6	pedestrians head centre, along street, poles, contrasts, shift between moving cars, in front of pedestrian
7	Head of pedestrian cyclist moving car, centre of pedestrian, head of pedestrian, corner, Shifting between centre of pedestrians, cars, especially moving, centre of bus, centre of cars
8	Following pedestrian, cyclist, looking along cars, houses, heads, cars centre of person, poles, moving cars, speed, heads, centre of cars
9	centre of pedestrians, head of pedestrian, centre of moving car, parked cars, cyclist, in front of cyclist,shift between pedestrian heads, poles shift between centre of pedestrians, moving cars, poles
10	centre of pedestrian, cyclist, feet, head, contrast (curb edges), following edges,in front of car, cars, in front of pedestrians, centre of pedestrians, corners, poles

Table 4.3: Observations of eye-traces on surveillance scenarios.

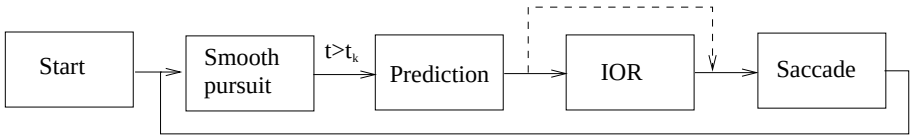


Figure 4.12: Flowchart of automatic trace generation

4.7 A statistical model

A statistical model is build upon the results from the classifier. Speed during saccade $\hat{s}_s = 856.262 pixels/s$ as well as mean duration of smooth pursuit $\hat{t}_{sp} = 0.347932$ and standard deviation $\sigma_{sp} = 0.663629$ have been estimated. The idea is to test theory on visual attention, by comparing real eye-traces with a model that outputs data with similar statistical characters.

From this an automatic visual attention mechanism is implemented which involves high–level object knowledge, inhibition of return, prediction of target and the statistical properties derived illustrated in Fig. 4.12. The system starts by finding a high–level object to follow in smooth pursuit. Such a target is taken from the output of the tracker, that has been calculated in advance. A random number is taken from a Gaussian distribution with mean and standard deviation taken as estimations from the sampled data. This is used as the duration for which smooth pursuit is continued. After that a saccade is done with the speed $|v_s| = \hat{s}_s$. It is traversed in a predictive manner in that it collides with the target in the short future. When it is sufficiently close to the target the system goes into smooth pursuit, and the cycle restarts. The selection of next object to follow is completely random in case the IOR module is inactive, and it is just as likely that the same object is selected as any other. If the IOR module is active an inhibition map is maintained that encodes the relative probability of a point becoming a future target given an object being centred there. When a point is fixated a Gaussian with standard deviation 25 pixels and maximum value 128 is added to the inhibition map. Every frame each pixel in the inhibition map is subtracted by 1. So for each possible saccade endpoint the relative probability is retrieved. Each target endpoint is evaluated in the inhibition map with respect to its relative probability and a final choice of fixation point is done accordingly.

4.8 Results

4.8.1 Evaluation

The above mentioned saliency models have been validated with eye-tracking data, with slightly different techniques. In most studies, some sort of experimental validation has been applied[1, 12, 17, 13]. The need for quantitative measurements is needed not only to test a certain technique, but also to adjust the model and mathematical parameters of the model.

In a study[13] a correlational approach is used. Here the stimuli are still images and a Gaussian smoothed mean human attention map is calculated from saccade locations. A correlation score is calculated between the human attention map and the final saliency map. This technique has the advantage of being intuitive since one can easily compare the output of the human map and the saliency map, and a correlational score is easily interpreted. The correlational score works fine for still images since on every frame each individual is directing their attention to several points. By taking the mean of several people a human attention map is obtained which is similar in character to the saliency map itself.

4.8.2 Measurement

Since videos are used in the experiment and comparisons are done between different combinations of features, correlation has been chosen to measure the similarity between model and data. The correlation coefficient r_{xy} is calculated with the following formula:

$$r_{xy} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\sqrt{n\sum y_i^2 - (\sum y_i)^2}}, \tag{4.1}$$

where x_i and y_i are the pixel values ordered in scan order through each frame in the automatically generated saliency map and the saliency map generated from eye-tracks and n is the number of total pixels in the videos. Using the second measurement in section 2.4.3 gets a answer to the question whether the difference between saliency at fixation points and at random locations is statistically significant.

Scores show that the correspondences with high-level features are higher than with

	PF	Both GM	Veh GM	Ped GM
S1		2	3	
S2		2	3	
S3		2	3	
S4		2	3	
S15	30	30	30	30
S5	35	40	50	40
S6	60	70	60	70
S7b	50	40	55	40
S8b	60	80	60	65

Table 4.4: Optimal sizes of Gaussian on eye-traces.

	PF	Both GM	Veh GM	Ped GM
S15	3	2	1.8	1.0
S5	1.2	1	1	1.0
S6	1.6	1.2	1.2	1.2
S7b	1.2	1.2	1.4	1.2
S8b	1.2	1	1.2	1.0

Table 4.5: Optimal relative size of Gaussian on objects measured as ratio between standard deviation and track width and height.

low-level features alone. The number of samples are in terms of millions or billions since it represents all pixels in the video and even small differences are significant. The outcome of the calculations is shown in table 4.6-4.9. Since the number of samples are in the order of $n \approx 10^7$, any difference in the third digit is considered significant. This means that all differences between correlations in table 4.6 and 4.7 are significant.

4.8.3 Optimizing comparison

To optimize the comparison between saliency maps and eye-traces several experiments have been made. To start with initial experiments have made clear that the Gaussian on objects should cover the entire object and not parts of it. After that, several experiments have been made with varying size of Gaussians on objects as well as on eye-traces. As can be seen in the table 4.4 optimal size of Gaussians on each eye fixation differs quite substantially between sequences. For meeting scenarios (S1-S4) the best score between 2-3 pixels and for surveillance scenarios (S15, S5, S6, S7b, S8b) between 30-80 pixels. This is probably due to faces being smaller in combination with the fact that these sequences

have lower resolution than surveillance ones. A contributing factor should also be the strong aggregation of fixations on the centres of the faces. The one that stands out of the surveillance scenarios is S15 which has lower resolution. An influence could also be due to the different set-up of the camera with respect to traffic.

Table 4.5 reveals the relative size of Gaussians on objects which varies between 1.0 and 3. Especially deviating is the surveillance sequence S15, in cases of using only PF or in using vehicles with and without persons in combination. In the first case it could be due to degenerating sized of tracks from PF and in the second case it could be due to that the sequence in question is inappropriate for vehicle tracking. A value of 1.3 is a sound value to use and was used in subsequent experiments.

4.8.4 Combination of features

First, it was tested how to combine high-level information from face and pedestrian tracking and low-level features. Results are displayed in table 4.6. The conditions are low-level features alone (Low-level), high-level features alone (High-level), combination of low- and high-level features (Combination), combination with bridge between all high-level features (Bridge), combination with low-level features only on objects (OOO), fake saliency map (Fake) and saliency maps generated from tracking ground-truth alone (GT). Examples of combined saliency maps and the corresponding eye-tracks are illustrated in Fig. 4.13. In the Bridge condition a wide connection has been made between objects and in the OOO condition low-level features are added only onto high-level features. The fake saliency maps used have been created by hand by the author and contains only one image, each illustrated in Fig. 4.7, throughout the videos.

Table 4.6 indicates strongly that the combination of low- and high-level features are much better than low-level features alone, especially for simple meeting scenarios like sequence S2. Furthermore even using high-level features alone gives higher scores than the combination in two cases (S2 and S4) and worse in only one of the cases (S13), which is due to failure of the tracker. The correlation measure does not give a strong difference (presented equal) between low-level features alone and the combination (S13, S7b and S8b). Also, the correlation scores in the Fake condition are significantly higher than change levels, which must be due to overlap between generated Fake saliency maps and real fixations,

perhaps due to objects appearing in the middle of the camera since the camera is placed so that this occurs, and the fact that faked maps have saliency more in the centres. It is lower than the scores in the High-level condition except in sequence S13 and S8b though. In S13 the difference is not that significant, but it is in S8b, which should be due to fake saliency in Fig. 4.7d.

The Bridge condition improves scores especially for surveillance sequences. This is not surprising since, as we shall see later, subjects move their gaze more frequently between pedestrians and other moving objects, on these sequences and in S2 there is only one relevant high-level object attended. The OOO condition does improve in comparison to the Combination condition, however not in comparison to the high-level features alone. Finally, the ground-truth gives considerably better scores on sequence S3 and S13, most likely due to inaccurate tracking of high-level features in the model. This indicates that it is actually the detection of objects that improve the similarity to real eye-traces. On the other sequences except for S4 we have similar results. The exception must be due to a random successful correlation between false positives from tracking and real fixations.

In table 4.7 a combination of high-level and low-level features are compared with different weights given to the low-level features. The table shows that, in 3 of 8 sequences, lower weights on low-level features yield higher scores in comparison to higher weight. Only in one case (seq. S6) does the correlation improve, but then only to decay again.

The conditions used in table 4.8 are similar to the ones in 4.6 except for the fake and GT condition since these would be duplicates. Here change is tracked on five surveillance sequences. Similarly to 4.6 high-level features produce higher scores than low-level features and also more than the combination in one case, however similar in the other four. Further, the Bridge improve scores in 4 of 5 sequences significantly and the OOO condition introduce improvement in one of four sequence. The other sequences show no change with respect to using high-level features alone. As in the case of pedestrian sequences the Bridge condition is successful due to many movements between objects. The improvement in the OOO condition could possibly be due to the observed tendency of subjects to fixate on low–level features on vehicles, e.g. corners of cars.

Table 4.9 illustrates that introducing low-level features decrease correlation scores in one case and keeps it constant in the other. As the results with faces and pedestrians

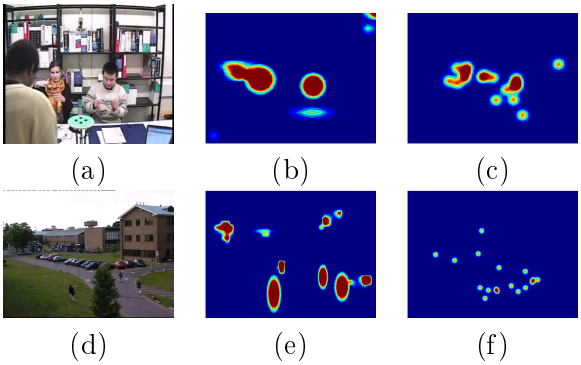


Figure 4.13: Saliency maps have been evaluated against eye-traces that have been projected as Gaussians and added up from 20 subjects. (a,d) Original frame. (b,e) Saliency map with low- and high-level features. (c,f) Saliency map generated from eye-traces.

Seq.	Low-level	High-level	Combination	Bridge	OOO	Fake	GT
S2	0.09	0.63	0.60	0.60	0.61	0.17	0.60
S3	0.01	0.24	0.24	0.26	0.25	0.01	0.51
S4	-0.01	0.22	0.17	0.19	0.18	-0.02	0.13
S13	0.02	0.01	0.08	0.15	0.10	0.09	0.15
S5	0.03	0.12	0.12	0.16	0.12	0.09	0.11
S6	0.04	0.09	0.09	0.17	0.09	0.08	0.08
S7b	0.03	0.10	0.10	0.16	0.10	0.08	0.08
S8b	0.04	0.11	0.11	0.18	0.11	0.30	0.12

Table 4.6: Correlations scores in 7 different conditions between automatically generated saliency maps and a saliency like map generated from initial experiments with 20 subjects. The high-level features used are faces in sequence S2, S3, S4, S13 and pedestrians in sequence S5, S6, S7b and S8b.

low-level features does not improve correlation scores.

The results show that the combination of high and low-level features is much better than low-level features alone. This is especially evident for meeting scenarios. However, higher scores are obtained with high-level features alone. The experiments with different weights on low-level features show clearly that lower weights give higher correlations. This is probably because of the fact that the majority of fixations are on high-level features and that low-level features introduce much excitation on the saliency map that does not produce real attentional attraction. Low-level features might still give guidance for spurious fixations on the background. Adding a bridge between objects is obviously improving results as shown in Table 4.7, which is not much of a surprise since attention shifts between objects.

Seq.	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
S2	0.63	0.63	0.63	0.63	0.62	0.62	0.62	0.61	0.61	0.60
S3	0.24	0.24	0.24	0.24	0.24	0.24	0.24	0.24	0.24	0.24
S4	0.18	0.18	0.18	0.18	0.18	0.18	0.18	0.18	0.18	0.17
S13	0.10	0.10	0.10	0.10	0.09	0.09	0.09	0.09	0.09	0.84
S5	0.12	0.12	0.12	0.12	0.12	0.12	0.12	0.12	0.12	0.12
S6	0.09	0.09	0.09	0.09	0.09	0.09	0.10	0.09	0.09	0.09
S7b	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10
S8b	0.11	0.11	0.11	0.11	0.11	0.11	0.11	0.11	0.11	0.11

Table 4.7: Correlations scores between automatically generated saliency maps and a saliency map generated from experiments with 20 subjects with different weights on low-level feature contribution. The high-level features used are faces in sequence S2, S3, S4, S13 and pedestrians in sequence S5, S6, S7b and S8b.

Seq.	Low-level	High-level	Combination	Bridge	OOO
S13	0.02	0.08	0.07	0.13	0.09
S5	0.02	0.07	0.07	0.10	0.07
S6	0.03	0.12	0.12	0.15	0.12
S7b	0.03	0.07	0.07	0.09	0.07
S8b	0.04	0.12	0.12	0.12	0.12

Table 4.8: Correlations scores in 5 different conditions between automatically generated saliency maps with moving objects as high-level features and a saliency like map generated from initial experiments with 20 subjects.

Seq.	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
S13	0.08	0.08	0.08	0.08	0.08	0.08	0.08	0.07	0.07	0.07
S5	0.07	0.07	0.07	0.07	0.07	0.07	0.07	0.07	0.07	0.07
S6	0.12	0.12	0.12	0.12	0.12	0.12	0.12	0.12	0.12	0.12
S7b	0.09	0.07	0.07	0.07	0.07	0.07	0.07	0.07	0.07	0.07
S8b	0.12	0.12	0.12	0.12	0.12	0.12	0.12	0.12	0.12	0.12

Table 4.9: Correlations scores between automatically generated saliency maps with moving objects as high-level features and a saliency like map generated from initial experiments with 20 subjects with different weights on low-level feature contribution.

4.8.5 Analysis of automatic model

The model based on eye-tracking statistics is compared with the eye tracking data itself in hope that the difference will comprise of qualities that reflect relevant mechanisms in visual attention. Lines have been drawn between fixation points indicating the generated scan–paths in both automatically generated movements and real movements. Results are depicted in Fig. 4.14-4.17.

One can see in Fig. 4.14-4.17 (see also 4.18), that tracking in the human visual attentional system produces more stable tracks than the presented tracker at least at higher frequencies. Thus tremor and micro-saccades are hardly visible in current displays. More importantly, the current model does not include any fixations to points in the background. Although most fixations are on faces, pedestrians and moving objects, some can be seen actually on the background as well. One could here speculate whether other objects are of interest to viewers and in such case which, lets say papers on the table or hands (see table 4.2), or if it is low-level features that attract attention (see table 4.3). Furthermore, in Fig. 4.14-4.15 one can see that the entire frame is scanned, which could be interpreted as a task of surveillance. This is not done in Fig. 4.16-4.17, where traces are more constrained to faces and to some extent body parts, with exception for subject 3 and 8 in table 4.2.

Further, real eye-traces produce longer saccades on surveillance scenarios, which also reflects the task of surveillance since it is reasonable to scan one area and then go to another area, but further investigation is needed to draw any definite conclusions. Thus a model of visual attention should reflect a distribution of real saccade length and not do saccades of any length with the same probability. Moreover, it seams the collected traces follow pattens with periods of movement in predominantly horizontal or vertical orientations (or more precisely in a similar orientation), thus previous fixation point and previous saccade could both be factors that contribute to the determination to the selection of the next fixation point. I also propose that global orientation in the image as in Fig. 4.15c is also a contributing factor, which in this case is determined by the road as a near strait path and the building surrounding it symmetrically.

Also, frame-wise comparison between fixation points and automatically generated traces can possibly give further indication to the underlying processes. In Fig. 4.18 one can see the comparison in a sequence containing one face only. First of all, the similarity between

Figure 4.14: Comparison between automatically generated eye traces (a) no IOR and (b) with IOR and eye–traces from two subjects (c) and (d) on sequence S7b.

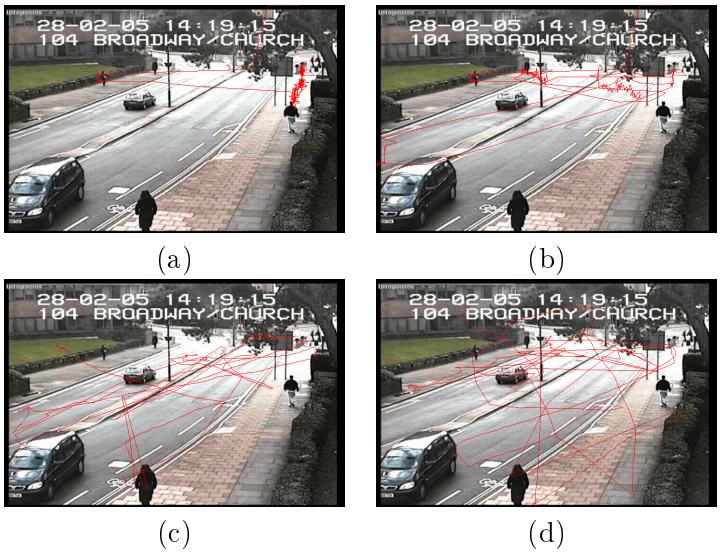


Figure 4.15: Comparison between automatically generated eye traces (a) no IOR and (b) with IOR and eye–traces from two subjects (c) and (d) on sequence S6 frames 4500–5000.

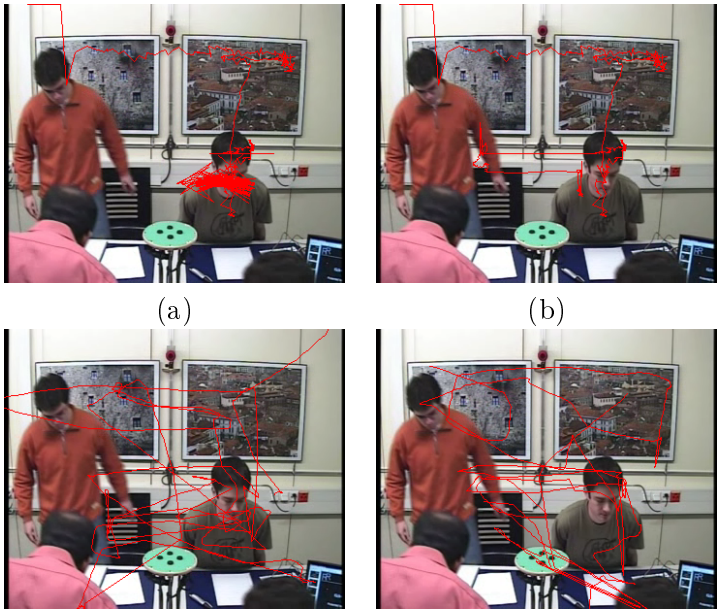


Figure 4.16: Comparison between automatically generated eye traces (a) no IOR and (b) with IOR and eye–traces from two subjects (c) and (d) on sequence S4.

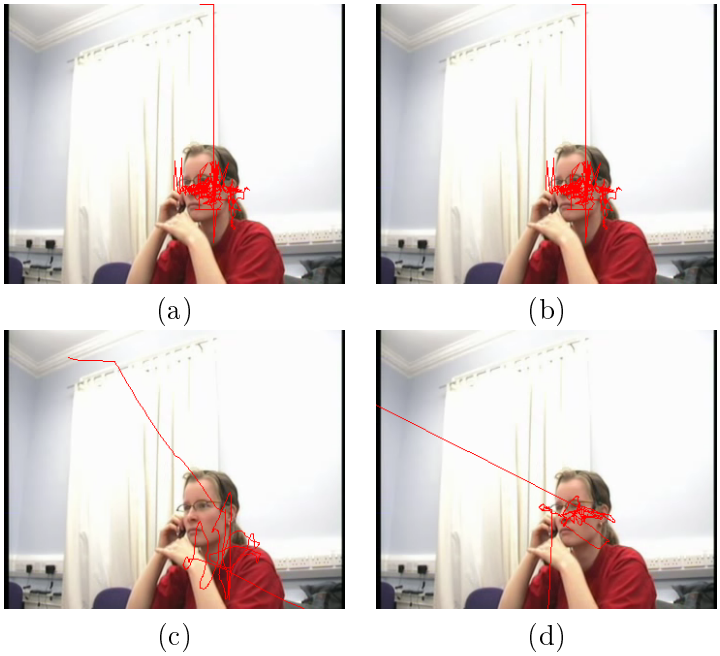


Figure 4.17: Comparison between automatically generated eye traces (a) no IOR and (b) with IOR and eye–traces from two subjects (c) and (d) on sequence S2.

model and collected data is striking. However, the automatically generated traces contain a consistent higher frequency component. But the real trace does instead contain small within-object shifts of attention with a lower frequency.

The mismatch between model and data is obvious in other sequences. For example, in Fig. 4.19 the automatically generated traces are following only one or a few objects whereas the subject appears to follow several and/or to fixate on the background, as also indicated in Fig. 4.14-4.17. There could be two reasons the automatically generated traces do not change object of smooth–pursuit as often as real subjects. Possibly the tracker is, due to limitations, not tracking all objects in every scene and so real fixations are spread on more objects. When only one tracked high–level object is tracked there are no shift of attention in the automatic model. It is also possible that the proposed classification of traces into saccades and smooth-pursuit is inaccurate, and in a better model more saccades should be generated per minute. In fact the model deviates from the sampled data as well since a saccade can be done to the currently tracked object, i.e. not producing a true saccade, but still being registered as a saccade in the statistical model.

Results have suggested that people follow the gaze of the other attendees as indicated by table 4.2. It is not completely ruled out though, that people direct their attention, due to for example hands moving an object in that area, is the cause of fixation.

There is further a difference between subjects in the duration of fixations or smooth pursuit phases as illustrated in Fig. 4.21, possibly revealing differences in state and trait variables between subjects. Also some subjects fixate more on the background (see Fig. 4.22).

The IOR mechanism affects the traces in two opposite directions. In some cases there are more changes of fixations, as in Fig. 4.23, possibly due to the current position being inhibited for further pursuit. In (a-b) there are just more saccades in the end of the sequence, with the IOR effect. In (c-d) there is a saccade earlier in (d) than (c). Also, obviously return to a previous location is inhibited, making it more likely that a current position is in pursuit in the short run with the IOR mechanism, as illustrated in Fig. 4.24

4.9 Conclusions

Introduction of high-level features on saliency maps strongly improve the general feasibility of the saliency map. Eye-tracking measurements show that a majority of fixations are on

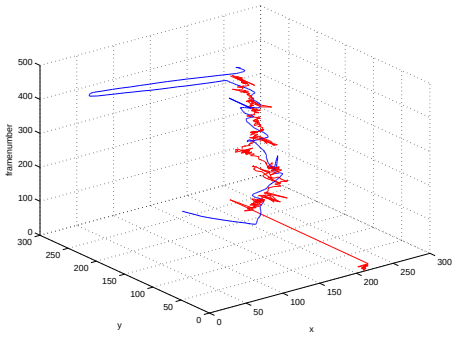


Figure 4.18: Comparison between real traces (blue) and automatically generated (red) on sequence S2.

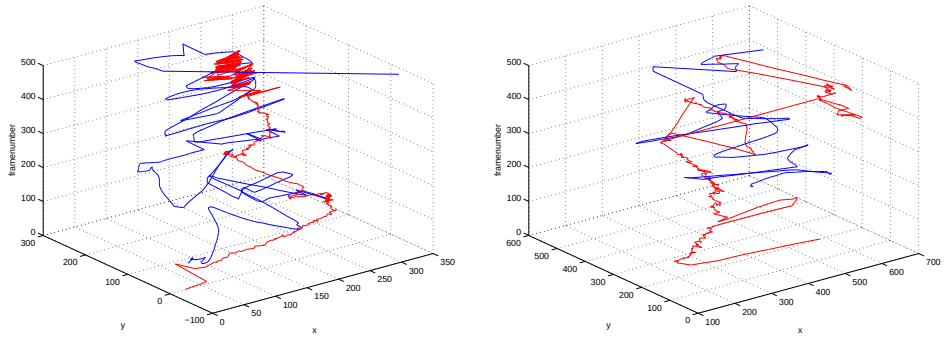


Figure 4.19: In some sequences (a) S4 (b) S6) subjects (blue) follow more objects than the automatic model (red).

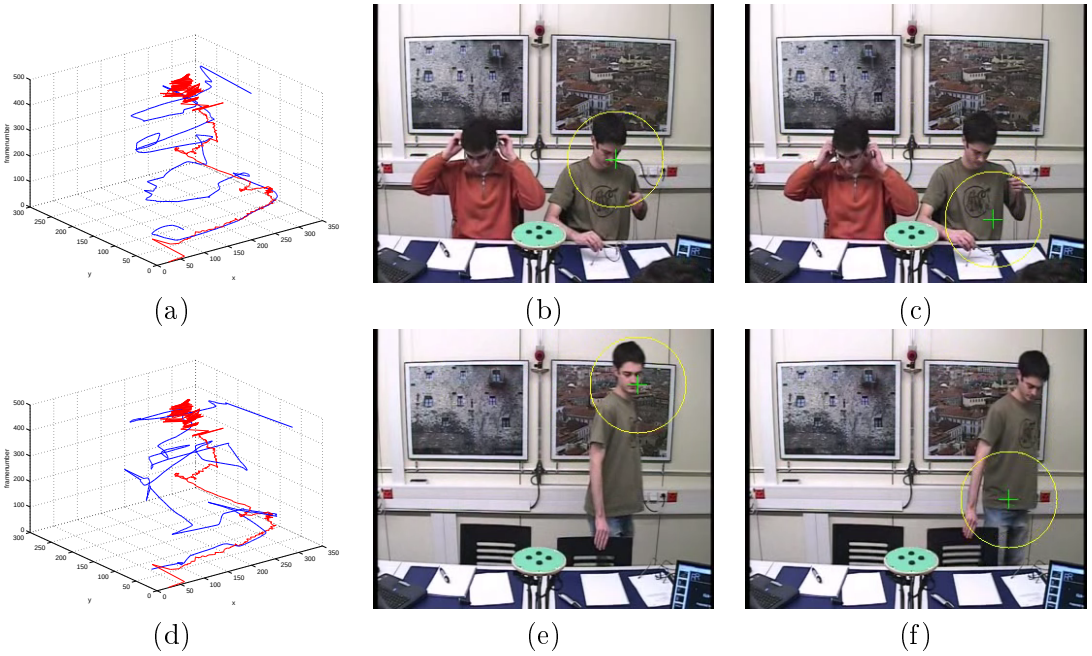


Figure 4.20: The frame-wise comparison between automatically generated (red) and real eye-traces (blue) in (a,d) reveal a deviation that is explained by fixations that follow gaze (b,e-c-f)

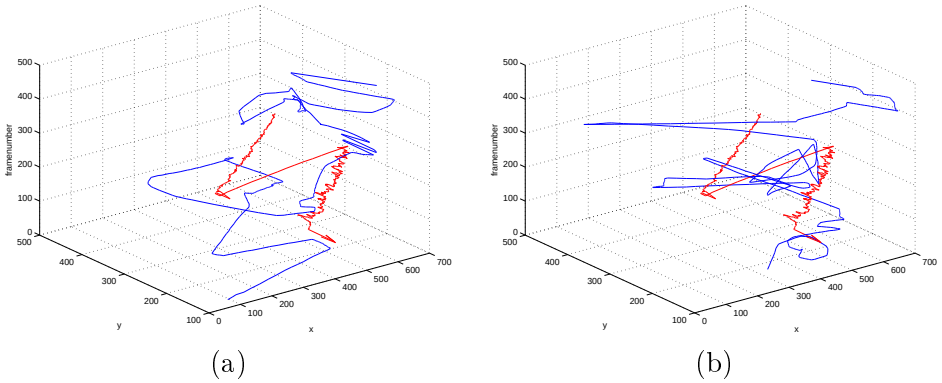


Figure 4.21: Some subjects (a) spend longer time in smooth pursuit per object than others (b). This is shown by more frequent abrupt changes in the track of (b). Red indicates automatically generated traces and blue subject gaze patterns.

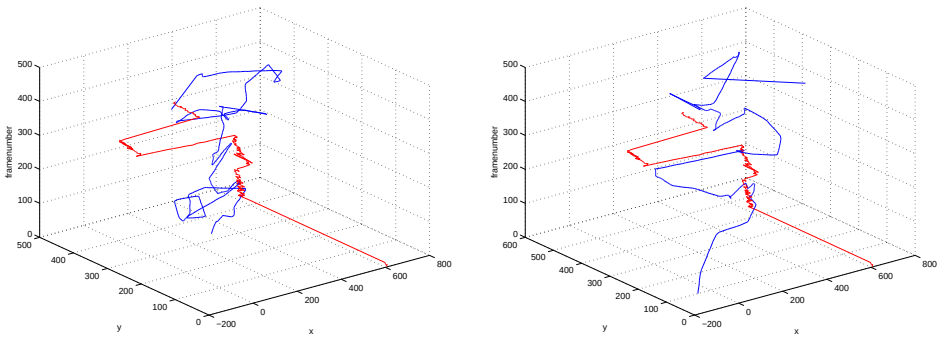


Figure 4.22: Some people (a) tend to spend more time on the background than others (b). Red indicates automatically generated traces and blue subject gaze patterns.

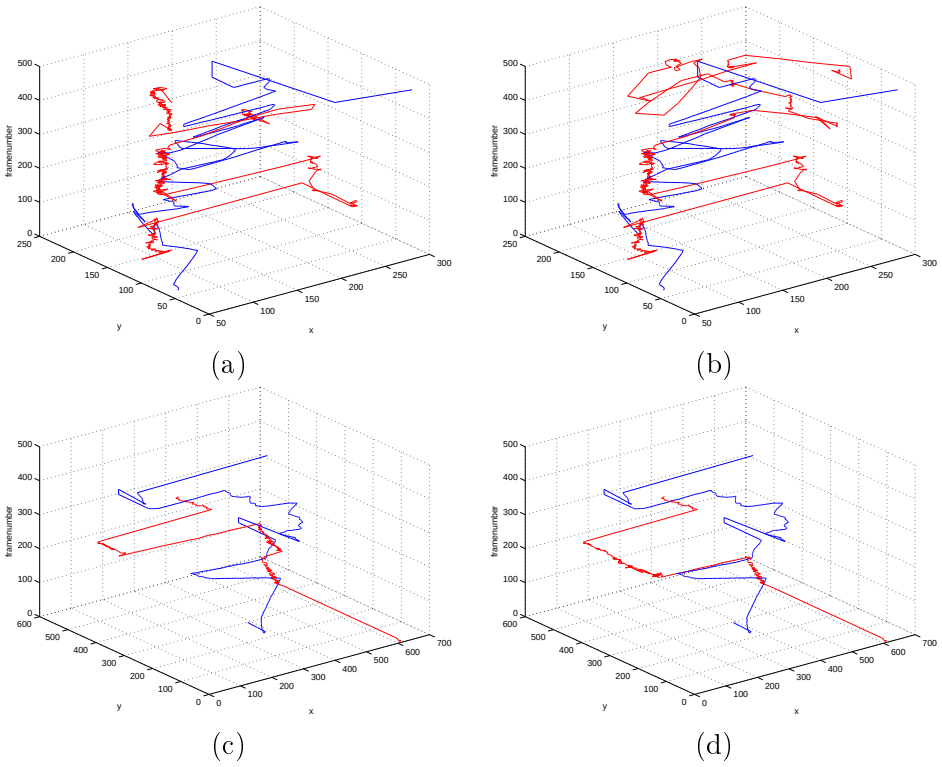
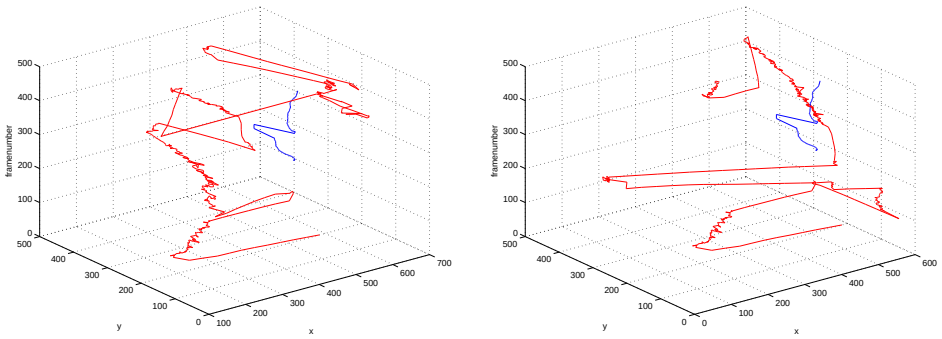


Figure 4.23: (a) Results without the IOR mechanism on sequence S3. (b) With the IOR mechanism on sequence S3. (c) Without the IOR mechanism on sequence S13. (d) With the IOR mechanism on sequence S13. Red indicates automatically generated traces and blue subject gaze patterns.



(a) (b)

Figure 4.24: (a) Results without the IOR mechanism on sequence S6. (b) With the IOR mechanism on sequence S6.

these high-level features, which presently are limited to faces, pedestrians and moving objects. Automatically generated eye-tracks on images using high- and low-level features are much more realistic than either alone. It has clearly been demonstrated that the addition or use of high-level features instead on low-level in a model of visual attention is promising.

It has been clearly demonstrated that high-level features are better than low-level features in a model of visual attention for videos. Videos can certainly be categorised by the generated scan-paths. Further, the extension of the above tracking framework using more high-level features would make it possible to model visual attention in other frameworks as well. It could also be experimented with how audiovisual events affect attention[278].

We are definitely following the most relevant objects, i.e. faces, pedestrians and moving objects. Which additional objects or features that are attended to should be further investigated. This could be investigated with task dependence in mind, for example in a video compression module in teleconferencing where the task is especially to follow the important objects of the conference. Also contextually defined objects like objects on the table, or objects that participants touch should be included in such a model as well as other task relevant objects.

Further, the model should be extended to exhibit fixations according to low-level features as well. Then we would have variation in type of integration of low-level and high-level features in the two presented models for automatic generation of traces. However, low-level contribution to correlation between model and data is low.

Further, using a Gaussian distribution of fixation duration might not be the best choice since there are no fixations with less duration than $0s$. Sampling of fixation lengths across subjects has been done, showing that it is a shrewd distribution[279]. A Gaussian assumption was made for the scope of this thesis. An obvious limitation of a model of saccade timing is that fixation might be reflecting the time it takes to analyse the target[280]. It should thus not be a random variable, but reflecting the output of several different processes generating such a statistical distribution, the unknowns that we would like to investigate. However, since the output of several competing processes of complicated and unpredictable nature, a random model is applicable. We do find a statistical regularity in the real tracks with a frequency in medium range though (see Fig. 4.19-4.24).

The IOR effect could be a mechanism that determines frequency of saccades in such a way that after a while the current fixated position is inhibited for further pursuit and thus a change of object occurs. This could be further studied in relation to possibly explaining regularities in eye-tracking data in more detail.

It is important to analyse differences between subjects in relation to fixations to background/foreground and also length of smooth pursuit periods (see Fig. 4.21-4.22. What variables does this depend on? Is it consistent within persons? If so does it depend on trait personality variables (for example compliance to authorities) or (for example temporary tiredness)? So, does it depend the person conducting the eye-tracking experiments, if it is a a professor or a student. Does anxiety levels influence the results? For example it could be an indicator of the motivation the subject has in following the instructions. In such case it would be of interest to change instructions and see if there is a change in behaviour. One could speculate whether a subject is really focusing attention in the broad sense on executing the task. For example lack of concentration could induce fixations on task irrelevant points. Another explanation is that subjects interpret instructions differently and thus fixate on different objects/features on the scene.

Results on meeting scenarios indicate that saccades are dependent on events, like for example that one person starts talking, noticeable from mouth movements, then looks at another person. What happens in this case is that attention is first allocated to the mouth area and then allocated to the eyes and then to the person where the gaze is directed towards. This pattern is most likely due to the fact that the subject is trying to follow the

conversation in the meeting scenario. The goal with a visual attention system that mimics the human visual attention system must be to capture such processes. The current model only involves task factors in so far as the higher–level objects tracked are task–congruent. Thus, further studies of visual attention in the context of scene/event interpretation would be fruitful. In general the eye-traces are dependent on task as it differs in surveillance scenarios, where except for objects being tracked, areas are scanned, possibly for new suspicious events.

Chapter 5

Conclusions

5.1 Summary of achievements

A fully automatic multi-object tracker has been implemented. Not only does detection initialize and terminated tracks, but the statistical formulation can integrate with detectors of any sort of object and any type of classification method. Further, low-level segmentation of chosen type can easily be used as validation of tracks and structure information can be integrated in the model matching of the tracked object. Specifically, a four dimensional face and human tracker has been developed that uses a cascade of Adaboost trained classifiers for detection and skin chromaticity and motion segmentation respectively to validate tracks. Further, a four and five dimensional moving object tracker that uses blobs from motion detection and, in the case of the five dimensional tracker, a blob measure as structure information, has been developed. It has been shown that the proposed system successfully initializes and terminates tracks automatically and have higher precision scores than particle filtering alone.

Moreover, a saliency model combining low- and high level features has been developed and evaluated against eye-tracking data. Here the major accomplishment is validation of the saliency model, and the finding that high-level objects account for the majority of the correlation with eye-tracking data, and low-level features even lower correlation scores. Also, the span of object based attention has been investigated.

The eye-traces collected has been evaluated qualitatively, giving better understanding of allocation of human attention on surveillance and meeting scenarios. A simple classifier

that identifies saccades and fixation/smooth-pursuit has been developed. Also statistics of the collected eye-traces have been extracted.

Perhaps the most interesting results come from models of visual attention on images and video, that have been implemented illustrating the benefits of adding high–level objects, as direct or by saliency mediated attractors. Also, similarities/dissimilarities between automatically generated traces and real traces on the same video illustrate some interesting qualities.

5.2 Philosophical considerations

The theory on visual attention in chapter 2 covers much more that is relevant for the experiments presented in this thesis. I would also like to take the opportunity to write down philosophical conclusions that I have made during reading this theory. This is important to be able to understand visual attention as well as modelling it computationally. We need to be able to interpret psychological, neurological and psychophysical findings, to draw conclusions about how to structure a visual attention system, what are the constituents and how do they relate to each other. I also would like to prepare for a look at the future, which directions I would like to take in case I continue to study visual attention later on.

5.2.1 Search for meaning

First, I would like to discuss the definition of attention as a mechanism of search optimisation. In my perspective one subtask of visual processing is to construct meaning out of the surroundings with respect to internal motives. Humans and other animals are not primarily engaged in a task of finding a particular target, although this ought to be a task they engage in at occasion. One of the subtasks of visual attention is definitely to interpret the world in terms of categories that are accessible to higher-level processes like goal achievement for example grouping spatially dispersed object parts into a whole. I have myself[8] done experiments showing how visual attention can speed up search for a object scene decomposition. The synthesis of the two perspectives is then that meaning arises naturally as a match between internal motives and external stimuli, where the goal is to find a match. The search for meaning involves not only the sensory data as to manipulate in parallel or as memory scratch pad objects but also the internal top-down flow of data,

i.e. goals, ideas and desires, and visual attention can be viewed as an optimisation strategy to find this meaning.

5.2.2 Covert attention

From the definition of attention given[22], and considering that covert attention is where the brain is focusing, covert attention is any selection process except overt orienting of the eyes. From these definitions I find it difficult to believe that covert attention can be readily defined as one distinct process, as opposed to consisting of several different selection processes, given that several different levels of processing and pathways towards experiencing and reacting to outer stimuli in the brain. Instead there should be different mechanism that all are given the name covert attention.

One possible type of covert attention is volitional attentional enhancement of an area that a human is not looking directly at. Another type is exogenously primed targets [281, 282]. A third type is the simple pop-out effects of early processing. Since some link covert attention directly to saccade preparation I propose looking for a saccade target involves a distinguishable "covert" attention mechanism instead of all covert attention being a mechanism that prepares saccades.

Experimental evidence has also been provided[283], for in this model, two sources of visual attentional control. The first one explains higher performance due to attention on noisy stimuli (integration mask), and the second on interruption mask. In the study with the conditions of with and without each masking and with both types of masking, performance is always improved by cuing the target. It has been argued[283] that since the combination does not give the strongest response it must be two different processes underlying. Although not explicitly proven, the first type is believed to be early visual enhancement and the second type late, corresponding to transference of stimuli to visual short term memory.

5.2.3 Object detection and recognition

Given that attention to objects, as meaningful units, mediates the quest of interacting with the outer world and achieving goals, objects must be detected and recognised. First, we would like to know how saliency relate to object based attention. For example, saliency

driven attention could be a mechanism that reveals the structure and is a guide to detecting and recognizing objects as such, thus allocating attention to objects in the scene. In such case objects arise from the stimuli, as a sort of grouping process before being matched with an internal object template. The question here is how much the stimuli is processed bottom-up for object detection and before recognition occurs. This is an area called mid–level processing. If saliency and objects attract attention completely orthogonally there should little processing before recognition.

One improvement to using detection modules that try to detect every type of object in every position, every 3D-orientation and scale is to find possible objects by grouping processes that find position and scale and perhaps even 3D-orientation of objects. Only type of object must now be established, and the computational complexity has been significantly reduced.

I would also like to suggest to interpret the findings of the neurological dissociation of where and what information processing in terms of object detection and recognition. In such a way possibly the "where" stream does detection and the "what" stream recognition. If so the what stream does not provide top-down information about the particular object category or identity before detection, and detection of objects is done as a grouping process before recognition starts. In this way reallocation of attention is done due to detection with the purpose of recognizing an object. The two pathways is also a possible solution to whether object recognition is serial or parallel in that parts of the processes are disparate and can have different type of processing.

From the above it follows that advanced models of grouping of visual stimuli should be incorporated into a model of visual attention, and visual attention is guided to points of interest due to bottom-up object indications (e.g. a colour segmented red car). Of course if recognition occurs more close to the unprocessed stimuli, we should definitely have object specific detectors that try to find every type of object in every position and at every scale of visual input. Given the potential reduction of computational complexity later object recognition is a viable alternative for a visual processing system. Research on the conspicuity of object-hood (e.g. [284]) does point to that objects are attracting attention in an early grouping face, before recognition. One might want to study attention to artificial displays (e.g. 2.4), introducing low-level as well as mid–level cues, like proximity

and continuity independently. Possibly one could study how attention to objects in videos relate to extracted low-level features.

5.2.4 Semantic gap and mid-level processing

The concept of the semantic gap defines the difference in representation and processing of sensory and higher-level reasoning, a gap the brain bridges per definition. The problems of recalculating data from one representation to the other is difficult theoretically. In my view the gap needs to be filled in with mid-level representations and processes. If we can bridge the semantic gap and do this with systems that operate in real-time we can solve possibly many problems in multi-media processing, for example video compression and building multimedia retrieval systems. Possibly research on mid-level processing will provide relevant results in present day computational model. For example, I am personally investigating using mid-level processes[14] in object tracking in my work at RetCorr AB, Helsingborg, Sweden. Mid-level processing is very much a way to make sensory data available for higher-level conceptualisations, and the importance of such a system is why I stress the relevance of study of scene interpretation in relation to visual attention in the presented thesis. Technologies derived from present day computer vision models can at least provide short-cuts between low-level and high-level representations, and it is not clear whether these can provide just as good or even better computational models than the human brain presents. However, it is at least reasonable to believe that deriving ideas from studies of the human brain is something that is and will be fruitful, since we want to do what the brain does.

5.3 Future work

- *Improved tracking:* People and moving objects are of high interest at least in the studied types of scenarios, thus better trackers of such objects should be developed. The pedestrian tracker has low performance due to many false negatives and the moving object tracker due to inaccurate change detection. A possibility to improve pedestrian tracking is to train the feature based classifier with refined data like edges instead of intensity arrays, or to use other feature based approaches especially tuned for person detection.

- *Gaze following behaviour:* Since it was found that visual attention follows the gaze of persons in meeting scenarios it would be interesting to take advantage of gaze and attentional tracking of people in video[285]. Particularly in the context of video conferences this could be used in the application area of video compression. Further, the tracking of additional objects/body parts should be included in such a model.

- *Task influence:* The influence of task on eye–traces could also be studied. For example giving different instructions to subjects in different groups, perhaps none in one group, higher-level concepts influencing on direction of attention at the population level could be investigated. For example does task relevance affect allocation of attention to particular classes of objects. The question then is how and when object recognition occurs before overt orienting.

- *Video compression:* The major area the presented work could be utilised almost immediately in is video compression. By combining high- and low-level features in a saliency map improved compression ratios, given a experienced quality standard, could easily be obtained with current compression frameworks MPEG-2 and MPEG-4 similarly to the work by Itti et. al. [9] using low-level features alone. An extension to this would be to encode objects of interest to visual attention separately in the MPEG-4, improving over motion prediction. Possibly objects could even be expressed as modifications of the features that distinguish them from other objects, e.g. the features in the presented tracker framework, thus significantly reducing dimensionality.

- *Semantic content retrieval:* Another possible application area is semantic description of multimedia, e.g. to build a multimedia retrieval system. In a multimedia retrieval system we would like to ask for media that match our criteria, for example in surveillance videos short sequences that contain threatening or suspicious behaviour.I think current state of the art, in object and event detection, can provide the tools to extract relevant information to some extent, however a visual attention mechanism might be necessary to extract only the information that is relevant to the person who searches for videos, based on automatically annotated videos. For example the following of a conversation, who is speaking and to whom is that person directing their speech.

In the more general case a complete story line should be done, and doing so by the automatic system involves filtering out unimportant data with respect to important data, both generally but also with the story line itself considered. In most motion pictures it is not significant if mosquito enters the scene, since this is probably a random unimportant event, but in a program about nature it is of utmost important. For a more refined abstraction of sensory data, given internal model of the world, then perhaps more advances technologies taking advantage of evolved systems in the human brain could be utilised.

Bibliography

[1] L. Itti. Quantifying the contribution of low-level saliency to human eye movements in dynamic scenes. *Visual Cognition*, 12, Aug 2005.

[2] C. R. Noback, N. L. Strominger R. J. Demarest, and D. A. Ruggiero. *The human nervous system: structure and function.* Humana Press Inc., New Jersey, USA, 6 edition, 2005.

[3] R. D. Wright and L. M. Ward. *Orienting of Attention.* Oxford University Press, 2008.

[4] B. Wandell. *Foundations of vision.* Sinauer Associates, Sunderland, Massachusetts, 1995.

[5] L. Itti. Models of bottom-up attention and saliency. In J. K. Tsotsos L. Itti, G. Rees, editor, *Neurobiology of Attention*, volume 1, pages 576–582, Jan 2005.

[6] C. A. Rothkopf, D. H. Ballard, and M. M. Hayhoe. Task and context determine where you look. *Journal of Vision*, 7(14):1–20, 2007.

[7] S. B. Perlman, J. P. Morris, W. Vander, C. Brent, S. R. Green, J. L. Doyle, and K. A. Pelphrey. Individual differences in personality predict how people look at faces. *PLoS ONE*, 4(6):e5952, 06 2009.

[8] S. Karlsson. Monocular depth from occluding edges. Master's thesis, Department of Mathematics, Lund Institute of Technology, 2004.

[9] L. Itti. Automatic foveation for video compression using a neurobiological model of visual attention. *IEEE Transactions on Image Processing*, 13:1304–1318, 2004.

[10] A. Oliva, A.Torralba, M. S.Castelhano, and J. M. Henderson. Top down control of visual attention in object detection. In *IEEE Proceedings of the International Conference on Image Processing, Vol I*, pages 253–256, Veenendaal, The netherlands, 2003. Universal Press.

[11] D. Walther, U. Rutishauser, C. Koch, and P. Perona. Selective visual attention enables learning and recognition of multiple objects in cluttered scenes. *Computer Vision and Image Understanding*, 100:41–63, December 2005.

[12] L. Itti and C. Koch. Computational modelling of visual attention. *Nature Reviews Neuroscience*, 2:194–203, March 2001.

[13] N. Ouerhani, R. von Wartburg, H. Hügli, and R.M. Müri. Empirical validation of saliency-based model of visual attention. *Electronic Letters on Computer Vision and Image Analysis*, 3:13–24, 2003.

[14] S. Karlsson, M. Taj, and A. Cavallaro. Detection and tracking of humans and faces. *EURASIP Journal On Image And Video Processing*, 2008:1–9, 2008.

[15] E. Maggio and A. Cavallaro. Hybrid particle filter and mean shift tracker with adaptive transition model. In *Proceedings of IEEE Signal Processing Society International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 221–224, 2005.

[16] M. Taj, E. Maggio, and A. Cavallaro. Objective evaluation of pedestrian and vehicle tracking on the clear surveillance dataset. In *Proceedings of CLEAR, Springer LNCS, Baltimore, USA*, 2007.

[17] T. Jost, N. Ouerhani, R. von Wartburg, René Müri, and H. Hügli. Contribution of depth to visual attention: comparison of a computer model and human. In *Early cognitive vision workshop 2004*, Isle of Skye, Scotland, May 2004. Early cognitive vision workshop.

[18] D. D. Salvucci. A model of eye movements and visual attention. In *Proceedings of the International Conference on Cognitive Modeling*, pages 252–259, Veenendaal, The netherlands, 2000. Universal Press.

[19] Y. Gong. Maximum entropy model-based baseball highlight detection and classification. *Computer Vision and Image Understanding*, 96(2):181–199, May 2004.

[20] C. Baccon, L. Hafemeister, and Ph. Gaussier. Computational model for a task directed attention mechanism without effective recognition of targets. In *ECV Workshop*, 2000.

[21] K. Rapantzikos, Y. Avrithis, and S. Kollias. On the use of spatiotemporal visual attention for video classification. In *ICIP*, 2005.

[22] D. Wegener, F. Ehn, M. K. Aurich, F. O. Galashan, and A. K. Kreiter. Feature-based attention and the suppression of non-relevant object features. *Vision Research*, 48:2696–2707, 2008.

[23] A. Pastukhov, L. Fischer, and J. Braun. Visual attention is a single, integrated resource. *Vision Research*, 49:1166–1173, 2009.

[24] G. d'Avossa, G. L. Shulman, A. Z. Snyder, and M. Corbetta. Attentional selection of moving objects by a serial process. *Vision Research*, 46:3403–3412, 2006.

[25] F. Tong. Splitting the spotlight of visual attention. *Neuron*, 42(4):524–526, 2004.

[26] D. Baldauf and H. Deubel. Visual attention during the preparation of bimanual movements. *Vision Research*, 48:549–563, 2008.

[27] S. Kastner and M. A. Pinsk. Visual attention as a multilevel selection process. *Cognitive, Affective, & Behavioral Neuroscience*, pages 483–500, 2004.

[28] V. A.F. Lamme. Why visual attention and awareness are different. *TRENDS in Cognitive Sciences*, 7, 2003.

[29] A. L. Rothenstein and J. K. Tsotsos. Attention links sensing to recognition. *Image and Vision Computing*, 26:114–126, 2008.

[30] A. Tavassoli, I. van der Linde, A. C. Bovik, and L. K. Cormack. Eye movements selective for spatial frequency and orientation during active visual search. *Vision Research*, 49:173–181, 2009.

[31] T. Töllner, M. Zehetleitner, K. Gramann, and H. J. Müller. Top-down weighting of visual dimensions: Behavioral and electrophysiological evidence. *Vision Research*, 50:1372–1381, 2010.

[32] M. M. Doran and J. E. Hoffman. Event-related potentials reveal "intelligent suppression" during multiple object tracking. *Journal of Vision*, 10(7):303, 2010.

[33] N. B. Carlisle and G. F. Woodman. Do visual working memory representations automatically bias deployments of covert attention? *Journal of Vision*, 10(7):320, 2010.

[34] B. Purcell, R. Heitz, J. Cohen, G. Woodman, and J. Schall. Timing of attentional selection in frontal eye field and event-related potentials over visual cortex during pop-out search. *Journal of Vision*, 10(7):97, 2010.

[35] D. Soto, G. W. Humphreys, and D. Heinke. Working memory can guide pop-out search. *Vision Research*, 46:1010–1018, 2006.

[36] D. Gao, V. Mahadevan, and N. Vasconcelos. On the plausability of the discriminant center-surround hypotheses for visual saliency. *Journal of Vision*, 9(12):1–27, 2008.

[37] D. Gao and N Vasconcelos. Integrated learning of saliency, complex features, and object detectors from cluttered scenes. In *Advances in Neural Information Processing Systems*, volume 17, pages 481–488, 2005.

[38] L. Zhang, M. H. Tong, T. K. Marks, H. Shan, and G. W. Cottrell. Sun: A bayesian framework for saliency using natural statistics. *Journal of Vision*, 8(7):1–20, 2008.

[39] S. Chikkerur, T. Serre, and T. Poggio. Attentive processing improves object recognition. *Computer Science and Artificial Intelligence Laboratory Technical Report*, Oct. 2009.

[40] S. Chikkerur, R. Serre, and T. Poggio. A bayesian inference theory of attention: neuroscience and algotihms. *Computer Science and Artificial Intelligence Laboratory Technical Report*, Oct. 2009.

[41] M. Snorrason, H. Ruda, and J. Hoffman. Modeling cognitive effects on visual search for targets in cluttered backgrounds. In *Proceedings of SPIE 2th Annual AeroSense*, volume 3375, 1998.

[42] L. Elazary and L. Itti. A bayesian model for efficient visual search and recognition. *Vision Research*, 50:1338–1352, 2010.

[43] S. Ahmad and S. Omohundro. Efficient visual search: a connectionist solution. In *Proceeding of the 13th Annual Conference of the Cognitive Science Society*, 1991.

[44] M. Riesenhuber and T. Poggio. Hierarchical models of object recognition in cortex. *Nature Neuroscience*, 2:1019—1025, 1999.

[45] V. Navalpakkam and P. Perona. What, when and how of target detection in visual search. *Journal of Vision*, 10(7):1321, 2010.

[46] Y. Yeshurun and M. Carrasco. Attention improves or impairs visual performance by enhancing spatial resolution. *Nature (London)*, 396(6706):72, 1998.

[47] K. Anton-Erxleben, J. Abrams, and M. Carrasco. Evaluating comparative and equality judgments in contrast perception: Attention alters appearance. *Journal of Vision*, 10(11):1–22, 2010.

[48] B. Montagna, F. Pestilli, and M. Carrasco. Attention trades off spatial acuity. *Vision Research*, 49:735–745, 2009.

[49] James J. Clark. Spatial attention and latencies of saccadic eye movements. *Vision Research*, 39:585–602, 1998.

[50] C.-H. Juan, S. M. Shorter-Jacobi, and J. D. Schall. Dissociation of spatial attention and saccade preparation. *PNAS*, 101:15541–15544, Sep 2004.

[51] J. E. Hoffman and B. Subramaniam. The role of visual attention in saccadic eye movements. *Perception & Psychophysics*, 57(6):787–795, 1995.

[52] M. E. Goldberg, J. W. Bisley, K. D. Powell, and J. Gottlieb. Saccades, salience and attention: the role of the lateral intraparietal area in visual behavior. *Progress in Brain Research*, 155, 2006.

[53] A. C. Nobre, D. R. Gitelman, E. C. Dias, and M. M. Mesulam. Covert visual spatial orienting and saccades: overlapping neural systems. *Neuroimage*, 11:210–216, 2000.

[54] M. Corbetta, E. Akbudak, T. E. Conturo, A. Z. Snyder, J. M Linenweber, S. E. Petersen, M. E Raichle, D. C. Van Essen, and G. L. Shulman. A common network of functional areas for attention and eye movements. *Neuron*, 21:761–773, 1998.

[55] T. Collins and K. Doré-Mazars. Eye movement signals influence perception: Evidence from the adaptation of reactive and volitional saccades. *Vision Research*, 46:3659–3673, 2006.

[56] T. M. Gersch, E. Kowler, B. S. Schnitzer, and B. A. Dosher. Visual memory during pauses between successive saccades. *Journal of Vision*, 8(16):1–18, 2008.

[57] J. C. Rossini and M. von Gränau. Covert and overt selection on visual search. *Journal of Vision*, 10(7):1321, 2010.

[58] A. V. Belopolsky and J. Theeuwes. Where are attention and saccade preparation dissociated? *Psychological Science*, 2009.

[59] R. Godjin and J. Theeuwes. Parallell allocation of attention prior to the execution of saccade sequences. *Journal of Experimental Psychology: Human Perception and Performance*, 29:882–896, 2003.

[60] A. Belopolsky and J. Theeuwes. Differential involvement of the oculomotor system in covert visual search and covert endogenous cueing. *Journal of Vision*, 10(7):164, 2010.

[61] E. McSorley and J. M. Findlay. Visual search in depth. *Vision Research*, 41:3487–3496, 2001.

[62] T. Moore. The neurobiology of visual attention: finding sources. *Current Opinion in Neurobiology*, 16:159–165, 2006.

[63] A. Zenon, S. B. hamed, and J-E Duhanel. Visual search without attentional displacement. *Journal of Vision*, 9(11):1–15, 2009.

[64] J. Dubois, P. F. Hanker, and R. VanRullen. Atentional selection of noncontiguous locations: the spotlight is only transiently split. *Journal of Vision*, 9(5):1–11, 2009.

[65] J. Dubois, J. Macdonald, and R. VanRullen. Reevaluating the sustained division of the attentional spotlight at high temporal resolution. *Journal of Vision*, 10(7):148, 2010.

[66] S. K. Andersen, M. M. Müller, and S.A. Hillyard. Color-selective attention need not be mediated by spatial attention. *Journal of Vision*, 9(6):1–7, 2009.

[67] R. Koenig-Robert and R. VanRullen. Spatio-temporal mapping of exogenous and endogenous attention. *Journal of Vision*, 10(7):1280, 2010.

[68] M. Mancas. Relative influence of bottom-up and top-down attention. In Lucas Paletta and John Tsotsos, editors, *Attention in Cognitive Systems*, volume 5395 of *Lecture Notes in Computer Science*, pages 212–226. Springer Berlin / Heidelberg, 2009.

[69] R. Pedersini, C. Morvan, L. T. Maloney, T. S. Horowitz, and J. M. Wolfe. An abstract equivalent of visual search: Gain maximization fails in the absence of visual judgments. *Journal of Vision*, 10(7):1311, 2010.

[70] A. List, A. Sherman, A. V. Flevaris, M. Grabowecky, and S. Suzuki. Neural signatures of local and global biases induced by automatic versus controlled attention. *Journal of Vision*, 10(7):92, 2010.

[71] A. E. C. Pensky, A. Landau, and W.Prinzmetal. The effects of voluntary attention on the event-related potentials and gamma-band response of eeg. *Journal of Vision*, 10(7):93, 2010.

[72] R. Chakravarthi and R. VanRullen. Beam me up, scotty! exogenous attention teleports but endogenous attention takes the shuttle. *Journal of Vision*, 10(7):244, 2010.

[73] H-I. Liao and S-L Yeh. Interaction between stimulus-driven orienting and top-down modulation in attentional capture. *Journal of Vision*, 10(7):123, 2010.

[74] M. Pomplun and A Hwang. The dynamics of top-down and bottom-up control of visual attention during search in complex scenes. *Journal of Vision*, 10(7):1275, 2010.

[75] C. Koch and S. Ullman. Shifts in selective visual attention: towards the underlying neural circuitry. *Human Neurobiology*, 4:219–227, 1985.

[76] L. Elazary and L. Itti. Interesting objects are visually salient. *Journal of Vision*, 8(3):1–15, 2008.

[77] T. Foulsham and G. Underwood. What can saliency models predict about eye movements? spatial and sequential aspects of fixations during encoding and recognition. *Journal of Vision*, 8(2):1–17, 2008.

[78] W. Einhäuser, M. Spain, and P. Perona. Objects predicts fixations better that early saliency. *Journal of Vision*, 8(14):1–26, 2008.

[79] J. Theeuwes. Top–down search strategies cannot override attentional capture. *Psychonomic Bulletin & Review*, 11(1):65–70, 2004.

[80] R. Godjin and J. Theeuwes. Programming of endogenous and exogenous saccades: evidence for a competitive integration model. *Journal of Experimental Psychology: Human Perception and Performance*, 28:1039–1054, 2002.

[81] C. Suchy-Dicey and T. Watanabe. Re-thinking the active-passive distinction in attention from a philosophical viewpoint. *Journal of Vision*, 10(7):218, 2010.

[82] J. Gottlieb, P. F. Balan, J. Oristaglio, and D. Schneider. Task specific computations in attentional maps. *Vision Research*, 49:1216–1226, 2009.

[83] G. L. Malcolm and J. M. Henderson. Combining top-down proesses to guide eye movements during real–world scene search. *Journal of Vision*, 10(2):1–11, 2010.

[84] Y. Pertzov, E. Zohary, and G. Avidan. Implicitly perceiving objects attracts gaze during later free viewing. *Journal of Vision*, 9(6):1–12, 2009.

[85] D. D. J. de Grave, C. Hesse, A-M. Brouwer, and V. H. Franz. Fixation locations when grasping partly occluded objects. *Journal of Vision*, 8(7):1–11, 2008.

[86] R. Alexander and G. Zelinsky. Visual similarity predicts categorical search guidance. *Journal of Vision*, 10(7):1316, 2010.

[87] K. Das, F. Guo, B. Geisbrecht, and M. P. Eckstein. Predicting contextual locations in natural scenes from neural activity. *Journal of Vision*, 10(7):1295, 2010.

[88] G. J. Zelinsky, W. Zhang, B. Yu, X. Chen, and D. Samaras. The role of top-down and bottom-up processes in guiding eye movements during visual search. In *Nineteenth Annual Conference on Neural Information Processing Systems (NIPS 2005)*, 2005.

[89] M. Stritzke and J. Trommershauser. Eye movements during rapid pointing under risk. *Vision Research*, 47:2000–2009, 2007.

[90] E. A. Reavis, P. J. Kohler, S. He, and P. U. Tse. Attentional tracking in the absence of consciousness. *Journal of Vision*, 10(7):303, 2010.

[91] A. Hollingworth, M. Matsukura, and S. J. Luck. Visual working memory influences the speed and accuracy of simple saccadic eye movements. *Journal of Vision*, 10(7):550, 2010.

[92] M. Silver, D. Bullock, S. Grossberg, M. Histed, and E. Miller. A neural model of how rank-selective spatial working memory and the supplementary eye fields control sequences of saccadic eye movements. *Journal of Vision*, 10(7):552, 2010.

[93] J. Spencer, S. Schneegans, and A. Hollingworth. Dynamic interactions between visual working memory and saccade planning. *Journal of Vision*, 10(7):537, 2010.

[94] M. Cui, G. Orban, M. Lengyel, and J. Fiser. The effect of previous implicit knowledge on eye movements in free viewing. *Journal of Vision*, 10(7):552, 2010.

[95] R. Bannerman, M. Milders, and A. Sahraie. Attentional bias to brief threat-related stimuli revealed by saccadic eye movements. *Journal of Vision*, 10(7):163, 2010.

[96] J. Lee and S. Shomstein. Reward driven prioritization modulates object-based attention in human visual cortex. *Journal of Vision*, 10(7):241, 2010.

[97] A. Kristjansson, O. Sigurjonsdottir, and J. Driver. "reversals of fortune" in visual search: Fast modulatory effects of financial reward upon visual search performance. *Journal of Vision*, 10(7):239, 2010.

[98] J. O'Brien, J. Raymond, and T. Sanocki. The role of motivational value in competition for attentional resources. *Journal of Vision*, 10(7):246, 2010.

[99] A. C. Schütz Send Mail and K. R. Gegenfurtner. Dynamic integration of saliency and reward information for saccadic eye movements. *Journal of Vision*, 10(7):551, 2010.

[100] L. P. Zapata, A. Aznar-Casanova, and H. Supèr. Dissociation of eye movement signals and perception during fixation. *Journal of Vision*, 10(7):549, 2010.

[101] B. Tamber-Rosenau and J. Moher. Attentional capture by objecthood is unaffected by salience in other dimensions. *Journal of Vision*, 10(7):130, 2010.

[102] S. Shomstein, S. Mayer-Brown, E. Wing, and S. Larsen. Relative contributions of spl and tpj to object-based attentional capture. *Journal of Vision*, 10(7):121, 2010.

[103] A. Nuthmann and J. M. Henderson. Object-based attentional selection in scene viewing. *Journal of Vision*, pages 1–19, 2010.

[104] D. G. Pelli, N. J. Majaj, N. Raizman, C. J. Christian, E. Kim, and M. C. Palomares. Grouping in object recognition: The role of a gestalt law in letter identification. *Cognitive Neuropsychology*, 26:36–49, February 2009.

[105] G. Kanizsa. *Organization in vision: essays on gestalt perception.* Praeger, New York, 1979.

[106] O. Ben-Shahar, B. J. Scholl, and S. W. Zucker. Attention, segregation, and textons: Bridging the gap between object-based attention and texton-based segregation. *Vision Research*, 47:845–860, 2007.

[107] A. Scharff and J. Palmer. Is object recognition serial or parallel? *Journal of Vision*, 10(7):228, 2010.

[108] T. Gao and B. J. Scholl. Chasing vs. stalking: Interrupting the perception of animacy. *Journal of Vision*, 10(7):239, 2010.

[109] T. Horowitz and Y. Kuzmova. Predictability matters for multiple object tracking. *Journal of Vision*, 10(7):243, 2010.

[110] T. Drew, T. S. Horowitz, J. Wolfe, and E. K. Vogel. Neural measures of interhemispheric information transfer during attentive tracking. *Journal of Vision*, 10(7):302, 2010.

[111] I.A. Rybak, V.I. Gusakova, L.N. Podladchikova A.V. Golovan, and N.A. Shevtsoca. A model of attention-guided visual perception and recognition. *Vision Research*, 38:2387–2400, Dec 1997.

[112] G. Deco and J. Zihl. A neurodynamical model of visual attention: Feedback enhancement of spatial resolution in a hierachical system. *Journal of Computational Neuroscience*, 10:231–253, November 2001.

[113] G. Cohen. *Memory In The Real World*. Psychology Press, 2 edition, Dec 1996.

[114] S. M. Crouzet, H. Kirchner, and S. J. Thorpe. Fast saccades toward faces: Face detection in just 100 ms. *Journal of Vision*, 10(4):1–17, 2010.

[115] S. M. Morand, M-H. Grosbras, and R. Caldara amd M. Harvey. Looking away from faces: Influence of high-level visual processes on saccade programming. *Journal of Vision*, 10(3):1–10, 2010.

[116] M. Bindermann, C. Scheepers, and A. M. Burton. Viewpoint and center of gravity affect eye movements to human faces. *Journal of Vision*, 9(2):1–16, 2009.

[117] M. Vo, T. Smith, and J. Henderson. The dynamics of gaze when viewing dynamic faces. *Journal of Vision*, 10(7):135, 2010.

[118] J. Wolfe. Visual search in continuous, naturalistic stimuli. *Vision Research*, 34:1187–95, 1994.

[119] S. Baldasi, S. Burr, M. Carrasco, M. Echstein, and P. Verghese. Editorial: visual attention. *Vision Research*, 44:1189–1191, 2004.

[120] M. C. Mozer. *The perception of multiple objects: A connectionist approach*. PhD thesis, Cambridge, Mass., 1991.

[121] M. C. Mozer and M. Sitton. Computational modeling of spatial attention. In H. Pashler, editor, *Attention*, pages 341–393. London: Psychology Press, 1998.

[122] K. Schill, E. Umkehrer, S. Beinlich, G. Krieger, and C. Zetzsche. Scene analysis with saccadic eye movements: Top-down and bottom-up modeling. *Journal of Electronic Imaging*, 10:152–160, Jan 2001.

[123] J. Najemnik and W. S. Geisler. Eye movement statistics in humans are consistent with an optimal search strategy. *Journal of Vision*, 8(3):1–14, 2008.

[124] I.A. Rybak, V.I. Gusakova, A.V. Golovan, L.N. Podladchikova, and N.A. Shevtsova. A model of attention-guided visual perception and recognition. *Vision Research*, 38:2387–2400, Aug 1998.

[125] D. Noton and L. W. Stark. Scanpaths in saccadic eye movements while viewing and recognizing patterns. *Vision Research*, 11:929–942, 1971.

[126] L. W. Stark and S. R. Ellis. Scanpaths revisited: cognitive models direct active looking. In et al. Fisher, editor, *Eye movements: cognition and visual perception*, pages 193–226. Hillsdale, NJ: Lawrence Erlbaum Associates, 1981.

[127] J. K. Tsotsos, L. Itti, and G. Rees. A brief and selective history of attention. *Neurobiology of Attention*, 2005.

[128] T. S. Horowitz, A. O. Holcombe, J. M. Wolfe, H. C. Arsenio, and J. S. DiMase. Attentional pursuit is faster than attentional saccade. *Journal of Vision*, 4:585–603, July 2004.

[129] D. Vishwanath and E. Kowler. Localization of shapes: eye movements and perception compared. *Vision Research*, 43:1637–1653, 2003.

[130] A. Kaminiarz, K. Königs, and F. Bremmer. Task influeces on the dynamic properties of fast eye movements. *Journal of Vision*, 9(13):1–11, 2009.

[131] R. J. Krauzlis. Target selection, attention and the superios colliculus. *Behavioural and Brain Sciences*, 30, 2007.

[132] R. M. Mcpeek, E. L. Keller, K. Nakayama, and R. M. Mcpeek. Concurrent processing of saccades. *Behavioral and Brain Sciences*, 22:692, 1996.

[133] S. Pannasch, J. Schulz, and B. Velichkovsky. Explaining visual fixation durations in scene perception: Are there indeed two distinct groups of fixations? *Journal of Vision*, 10(7):138, 2010.

[134] C. Wilimzig, T. Palmeri, G. Logan, and J. Schall. Toward an interactive race model of double-step saccades. *Journal of Vision*, 10(7):210, 2010.

[135] T. L. Alvarez, Y. Alkan, S. Gohel, B. D. Ward, and B. B. Biswal. Functional anatomy of predictive vergence and saccade eye movements in humans: A functional mri investigation. *Vision Research*, 50:2163–2175, 2010.

[136] D. Vergilino-Perez and J. M. Findlay. Between-object and within-object saccade programming in a visual search task. *Vision Research*, 46:2204–2216, 2006.

[137] R. Walker and E. McSorley. The parallel programming of voluntary and reflexive saccades. *Vision Research*, 46:2082–2093, 2006.

[138] T. L. Hodgson, B. A. Parris, N. J. Gregory, and T. Jarvis. The saccadic stroop effect: Evidence for involuntary programming of eye movements by linguistic cues. *Vision Research*, 49:569–574, 2009.

[139] S. K. Mannan, C. Kennard, D. Potter, Y. Pan, and D. Soto. Early oculomotor capture by new onsets driven by the contents of working memory. *Vision Research*, 50:1590–1597, 2010.

[140] H. A. Trukenbrod and R. Engbert. Oculomotor control in a sequential search task. *Vision Research*, 47:2426–2443, 2007.

[141] J. M. Findlay and V. Brown. Eye scanning of multi-element displays: I. scanpath planning. *Vision Research*, 46:179–195, 2006.

[142] I. T. C. Hooge, B. N. S. Vlaskamp, E. A. B. Over, and C. J. Erkelens. Coarse-to-fine eye movement strategy in visual search. *Vision Research*, 47:2272–2280, 2007.

[143] B.C. Motter and J. Holsapple. Saccades and covert shifts of attention during active visual search: Spatial distributions, memory, and items per fixation. *Vision Research*, 47:1261–1281, 2007.

[144] M. Rolfs, R. Kliegl, and R. Engbert. Toward a model of microsaccade generation: The case of microsaccadic inhibition. *Journal of Vision*, 8(11):1–23, 2008.

[145] X. G. Troncoso, S. L. Macknik, and S. Martinez-Conde. Microsaccades counteract perceptual filling–in. *Journal of Vision*, 8(14):1–9, 2008.

[146] H. Collewijn and E. Kowler. The significance of microsaccades for vision and oculomotor control. *Journal of Vision*, 8(14):1–21, 2008.

[147] R. J. Krauzlis. Recasting the smooth pursuit eye movement system. *Journal of Neurophysiology*, 91:591–603, 2004.

[148] A. C. Schütz, D. I. Braun, and K. R. Gegenfurtner. Object recognition during foveating eye movements. *Vision Research*, 49:2241–2253, 2009.

[149] J-J. O. de Xivry, M. Missal, and P. Lefèvre. A dynamic representation of target motion drives predictive smooth pursuit during target blinking. *Journal of Vision*, 8(15):1–13, 2008.

[150] U. Biber and U. J. Ilg. Initiation of smooth-pursuit eye movements by real and illusory contours. *Vision Research*, 48:1002–1013, 2009.

[151] P. van Donkelaar and A. S. Drew. The allocation of attention during smooth pursuit eye movements. *Progress in Brain Research*, 140, 2002.

[152] L. P. Lovejoy, G. A. Fowler, and R. J. Krauzlis. Spatial allocation of attention during smooth pursuit eye movements. *Vision Research*, 49:1275–1285, 2009.

[153] D. Kerzel and N .E. Ziegler. Visual short-term memory during smooth pursuit eye movements. *Journal of Experimental Psychology: Human Perception and Performance*, 31:354–372, 2005.

[154] C. J. Erkelens. Coordination of smooth pursuit and saccades. *Vision Research*, 46:163–170, 2006.

[155] C.-S. Ray Li and S.-C. Lin. A perceptual level mechanism of the inhibition of return in oculomotor planning. *Cognitive Brain Research*, 14:269–276, February 2002.

[156] Z. Wang and R. M. Klein. Searching for inhibition of return in visual search: A review. *Vision Research*, 50:220–220, 2010.

[157] A. Kingstone and J. Pratt. Inhibition of return is composed of attentional and oculomotor processes. *Perception & psychophysics*, 61:1046–1056, 1999.

[158] K. Shen and M. Paré. Neural basis of object memory during visual search. *Journal of Vision*, 10(7):1294, 2010.

[159] D. Souto, S. Born, and D. Kerzel. The sensory component of inhibition of return. *Journal of Vision*, 10(7):266, 2010.

[160] J. J. Clark. Spatial attention and saccadic camera motion. In *Proceedings of the 1998 IEEE International Conference on Robotics and Automation*, 1998.

[161] S. P. Leversedge and J. M. Findlay. Saccadic eye movements and cognition. *Trends in Cognitive Sciences*, 4:6–14, 2000.

[162] G.J. Giefing, H. Janssen, and H. Mallot. Saccadic object recognition with an active vision system. In B. Neumann, editor, *Proceedings of the ECAI 92*, pages 803–805. Wiley and Sons, 1992.

[163] Rajesh P.N. Rao. Top-down gaze targeting for space-variant active vision. In *Proceedings ARPA Image Understanding Workshop*, pages 1049–1058, 1994.

[164] J. M. Canas, M. M. de la Casa, and T. Gonzales. An overt visual attention mechanism based on saliency dynamics. *ICMED*, 2:93–100, 2007.

[165] C. Breazeal and B. Scassellati. A context-dependent attention system for a social robot source. In *Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence*, pages 1146 – 1153, 1999.

[166] J.K. Tsotsos and K. Shubina. Attention and visual search : Active robotic vision systems that search. In *Keynote Lecture, The 5th International Conference on Computer Vision Systems, Bielefeld*, March 2007.

[167] A. Bogadhi, A. Montagnini P. Mamassian, L. Perrinet, and G. Masson. A recurrent bayesian model of dynamic motion integration for smooth pursuit. *Journal of Vision*, 10(7):545, 2010.

[168] O. Komogortsev and J. Khan. Predictive perceptual coding for real time video communication. In *ACM Multimedia 2004*, October 2004.

[169] Rafal Mantiuk Karol, Karol Myszkowski, and Sumanta Pattanaik. Attention guided mpeg compression for computer animations. In *Proceedings of the 19th Spring Conference on Computer Graphics*, pages 239–244, 2003.

[170] K.-C. Yang, C. C. Guest, and P. K. Das. Human visual attention map for compressed video. In *Proceedings of the Eighth IEEE International Symposium on Multimedia (ISM'06)*, pages 525–532. IEEE, 2006.

[171] A. Ude, V. Wyart, L-H. Lin, and G. Cheng. Distributed visual attention on a humanoid robot. In *Proceedings of 2005 5th IEEE-RAS International Conference on Humanoid Robots*, pages 381–386, December 2005.

[172] V. Vapnik. *Statistical learning theory*. John Wiley & Sons, New York, 1998.

[173] V. Sundstedt, A. Chalmers, K. Cater, and K. Debattista. Top-down visual attention for efficient rendering of task related scenes. In *Vision, Modeling and Visualization*, pages 209–216, 2004.

[174] S. de Jong. Top-down selective visual attention during slam. Master's thesis, Department of Artificial Intelligence, University of Groningen, 2008.

[175] E. Mendi and M. Milanova. Image segmentation with active contours based on selective visual attention. In *Proceedings of the 8th WSEAS International Conference on Signal Processing*, pages 79–84, May 2009.

[176] W. Wang, Y. Song, and A. Zhang. Semantics-based image retrieval by region saliency. In *Proceeding of International Conference on Image and Video Retrieval*, pages 29–37, 2002.

[177] G. Evangelopoulos, A. Zlatintsi, G. Skoumas, K. Rapantzikos, A. Potamianos, P. Maragos, and Y. Avrithis. Video event detection and summarization using audio, visual and text saliency. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, April 2009.

[178] A. A. Salah, E. Alpaydin, and L. Akarun. A selective attention-based method for visual pattern recognition with application to handwritten digit recognition and face recognition. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 24:420–425, 2002.

[179] Z. Han, Z. Liu, Z. Zhang, W. Li, J. Gu, and Z. Ning. Interesting moving object segmentation based on selective visual attention and markov random field. In *IET Conference on Wireless, Mobile and Sensor Networks (CCWMSN07)*, pages 394–397. IEEE, 2007.

[180] V. Mahadevan and N. Vasconcelos. Spatiotemporal saliency in highly dynamic scenes. *IEEE transactions on pattern analysis and machine intelligence*, 32, January 2010.

[181] S. Van der Stigchel. Recent advances in the study of saccade trajectory deviations. *Vision Research*, 50:1010–1018, 2006.

[182] P.-H. Tseng, Ian Cameron, D. Munoz, and L. Itti. Differentiating patients from controls by gazing patterns. *Journal of Vision*, 10(7):277, 2010.

[183] D. Lamy and L. Zoaris. Task–irrelevant stimulus salience affects visual search. *Vision Research*, 49:1472–1480, 2009.

[184] H. Nothdurft. Salience from feature contrast: variations with texture density. *Vision Research*, 34:3181–3200, 2000.

[185] P. Monnier. Detection of multidimensional targets in visual search. *Vision Research*, 46:4083–4090, 2006.

[186] K. R. Dobkins, A. A. Rezec, and B. Krekelberg. Effects of spatial attention and salience cues on chromatic and achromatic motion processing. *Vision Research*, 47:1893–1906, 2007.

[187] R. Carmi and L. Itti. Visual causes versus correlated of attentional selection in dynamic scenes. *Vision Research*, 46:4333–4345, 2006.

[188] S. E. Christ and R. A. Abrams. The attentional influence of new objects and new motion. *Journal of Vision*, 8(3):1–8, 2008.

[189] R. J. Peters and L. Itti. Applying computational tools to predict gaze direction in interactive visual environments. *ACM Transactions on Applied Perception*, 5, 2008.

[190] L. Jansen, S. Onat, and P. König. influence of disparity on fixation and saccades in free viewing of natural scenes. *Journal of Vision*, 9(1):1–19, 2009.

[191] J. Pratt, P. Radulescu, R. Guo, N-Al-Aidroos, and R. Abrams. Biological motion captures attention. *Journal of Vision*, 10(7):120, 2010.

[192] M. Donk and W. van Zoest. Effects of salience are short-lived. *Psychological Science*, 19:733–739, 2002.

[193] M. Donk. The role of salience-driven control in visual selection. *Journal of Vision*, 10(7):216, 2010.

[194] M. Z. Aziz, B. Mertsching, M. Salah, E.-N. Shafik, and R. Stemmer. Evaluation of visual attention models for robots. In *Proceedings of the Fourth IEEE International Conference on Computer Vision Systems*, 2006.

[195] W. Kenzle, M. O. Franz, B. Schölkopf, and F. A. Wichmann. Center–surround patterns emerge as optimal predictors for human saccade targets. *Journal of Vision*, 9(5):1–15, 2009.

[196] S. Vijayakumar, J. Conradt, T. Shibata, and S. Schaal. Overt visual attention for a humanoid robot. In *International Conference on Intelligence in Robotics and Autonomous Systems (IROS 2001)*, pages 2332–2337, Hawaii, 2001.

[197] M. Cerf, E. P. Frady, and C. Koch. Faces and text attract gaze independent of the task: Experimental data and computer model. *Journal of Vision*, 9(12):1–15, 2009.

[198] F. Baluch and L. Itti. The effects of 2nd–order feature interactions in predicting human gaze. *Journal of Vision*, 10(7):131, 2010.

[199] Anna Lazar. *Modeling Visual Attention.* PhD thesis, Faculty of Information Technology, Pázmány Péter Catholic University, 2008.

[200] E. Birmingham, W. F. Bischof, and A. Kingstone. Saliency does not account for fixations to eyes within social scenes. *Vision Research*, 49:2992–3000, 2009.

[201] X. Chen and G. J. Zelinsky. Real-world visual search is dominated by top-down guidance. *Vision Research*, 46:4118–4133, 2006.

[202] B. Follet, O. Le Meur, and T. Baccino. Modeling visual attention on scenes. *Studia Informatica Universalis*, 8:150–167, 2010.

[203] M. Pomplun. Saccadic selectivity in complex visual search displays. *Vision Research*, 46:1886–1900, 2006.

[204] R. Milanese, H. Wechsler, S. Gil1, J-M. Bost, and T. Pun. Integration of bottom–up and top–down cues for visual attention using non–linear relaxation. In *Proceedings of IEEE Conf. on Computer Vision and Pattern Recognition*, pages 781–785, June 1994.

[205] S-B. Choi, S-W. Ban, and M. Lee. Biologically motivated visual attention system using bottom–up saliency map and top–down inhibition. In *Neural Information Processing-Letters and Review*, 2004.

[206] B. Rasolzadeh, A. T. Targhi, and J.-O. Eklundh. An attentional system combining top-down and bottom-up influences. In L. Paletta and E. Rome, editors, *Attention in Cognitive Systems. Theories and Systems from an Interdisciplinary Viewpoint: 4th International Workshop on Attention in Cognitive Systems, WAPCV 2007 Hyderabad, India, January 8, 2007 Revised Selected Papers*, volume 4840, pages 123–140. Springer–Vorlag Berlin Heidelberg, 2008.

[207] F. H. Hamker. Modeling feature-based attention as an active top-down inference process. *BioSystems*, 86:91–99, 2006.

[208] C. Kanan, M. H. Tong, L. Zhang, and G. W. Cottrell. Sun: Top-down saliency using natural statistics. *Visual Cognition*, 17:979–1003, 2009.

[209] T. Xu, N. Chenkov, K. Kuhnlenz, and M. Buss. Autonomous switching of top-down and bottom-up attention selection for vision guided mobile robots. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 4009–4014, Oct 2009.

[210] M. T. Lopez, A. Fernandez-Caballero, J. Mira, A. E. Delgado, and M. A. Fernandez. Algorithmic lateral inhibition method in dynamic and selective visual attention task: Application to moving objects detection and labelling. *Expert Systems with Applications*, 31:570–594, 2006.

[211] M. B. Neider and G. J. Zelinsky. Scene context guides eye movements during visual search. *Vision Research*, 46:614–621, 2006.

[212] M. Dorr, T. Martinetz, K. R. Gegenfurtner, and E. Barth. Variability of eye movements when viewing dynamic natural scenes. *Journal of Vision*, 10(10):1–17, 2010.

[213] N. Ouerhani and H. Hügli. Computing visual attention from scene depth. *Proc. 15th Int. Conf. on Pattern Recognition, ICPR 2000*, 3:375–378, 2000.

[214] J. Harel, C. Koch, and P. Perona. Graph-based visual saliency. In *Advances in Neural Information Processing Systems*, volume 19, pages 545–552. MIT Press, 2007.

[215] S. Marat, T. Phuoc, L. Granjon, N. Guyader, D. Pellerin, and A. Guerin-Dugue. Modeling spatiotemporal saliency to predict gaze direction for short videos. *International Journal of Computer Vision*, 82:231–243, 2009.

[216] O. Le Meur, P. Le Callet, D. Barba, and D. Thoreau. A coherent computational approach to model bottom-up visual attention. *IEEE Transaction on pattern analysis and machine intelligence*, 28:802–817, 2006.

[217] R.-L. Hsu, M. Abdel-Mottaleb, and A. K. Jain. Face detection in color images. *IEEE Trans. Pattern Anal. Machine Intell.*, 24:1046–1049, May 2002.

[218] N. Bruce and J. Tsotsos. Saliency based on information maximization. In *Advances in Neural Information Processing*, volume 18, pages 155–162, 2006.

[219] Xiaodi Hou and Liqing Zhang. Saliency detection: A spectral residual approach. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR07). IEEE Computer Society*, pages 1–8, 2007.

[220] H. J. Seo and P Milanfar. Static and space–time visual saliency detection by self–resemblance. *Journal of Vision*, 9(12):1–27, 2009.

[221] L. Itti and P. Baldi. Bayesian surprise attracts human attention. *Advances in Neural Information Processing Systems*, 18:1–8, 2006.

[222] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scen analysis. *IEEE Transactions on Pattern Analysus and Machine Intelligence*, 20, 1998.

[223] K. Rapantzikos and N. Tsapatsoulis. On the implementation of visual attention architectures. In *Tales of the Disappearing Computer, Santorini*, June 2003.

[224] D. H. Hubel and T. N. Wiesel. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *Journal of Physiology*, 160:106–154, 1962.

[225] J. Månsson. A computational model of suppressive mechanisms in human contour perception. *Lund University Cognitive Studies*, 81, 2000.

[226] L. Itti, N. Dhavale, and F. Pighin. Realistic avatar eye and head animation using a neurobiological model of visual attention. In B. Bosacchi, D. B. Fogel, and J. C. Bezdek, editors, *Proceedings of SPIE 48th Annual International Symposium on Optical Science and Technology*, volume 5200, pages 64–78, Aug 2003.

[227] R. A. Abrams and S. E. Christ. Motion onset captures attention. *Psychological Science*, 14(5):427–432, 2003.

[228] R. A. Abrams and S. E. Christ. The onset of receding motion captures attention: comment on franconeri and simons (2003). *Perception & Psychophysics*, 67:219–223, 2005.

[229] S. L Franconeri and D. J. Simons. The dynamic events that capture visual attention: A reply to abrams & christ (2005). *Perception & Psychophysics*, 67:962–966, 2005.

[230] R. A. Abrams and S. E. Christ. Motion onset captures attention: A rejoinder to franconeri and simons (2005). *Perception & Psychophysics*, pages 114–117, 2006.

[231] M. Verma and P. W. McOwan. A semi-automated approach to balancing of bottom up salience for predicting change detection performance. *Journal of Vision*, 10(6):1–17, 2010.

[232] O. le Meur, D. Thoreau, P. le Callet D., and Barba. A spatio-temporal model of the selective human visual attention. *ICIP 2005. IEEE International Conference on Image Processing, 2005*, 3:1188–1191, 2005.

[233] D. Gao and N. Vasconcelos. Discriminant saliency for visual recognition from cluttered scenes. In *Proceedings of Neural Information Processing Systems*, pages 481–488, 2004.

[234] T. Kadir and M. Brady. Saliency, scale and image description. *International Journal of Computer Vision*, 45(2):83–105, 2001.

[235] R. Voorhies, L. Elazary, and L. Itti. Application of a bottom-up visual surprise model for event detection in dynamic natural scenes. *Journal of Vision*, 10(7):215, 2010.

[236] L. Itti and P. Baldi. Bayesian surprise attracts human attention. *Vision Research*, 49:1295–1306, 2009.

[237] T. Foulsham, R. Teszka, and A. Kingstone. What is the shape of the visual information that drives saccades in natural images? evidence from a gaze-contingent display. *Journal of Vision*, 10(7):534, 2010.

[238] A. Sherman and G. Alvarez. Real-world statistical regularities guide the deployment of visual attention, even in the absence of semantic scene recognition. *Journal of Vision*, 10(7):1285, 2010.

[239] Y. Sun and R. Fisher. Object-based visual attention for computer vision. *Artificial Intelligence*, 146:77–123, May 2003.

[240] Ming-Hsuan Yang. Face detection. In S. Z. Li and A. K. Jain, editors, *Encyclopedia of Biometrics*, pages 303–308. Springer US, 2009.

[241] H. Rowley, S. Baluja, and T. Kanade. Neural network-based face detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20:23–38, 1998.

[242] M. H. Yang, D. Kriegman, and N. Ahuja. Detecting faces in images: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24:34–58, 2002.

[243] H. Schneiderman and T. Kanade. Object detection using the statistics of parts. *International Journal of Computer Vision*, 56:151–177, 2004.

[244] F. Fleuret and D. Geman. Coarse-to-fine face detection. *International Journal of Computer Vision*, 41:85–107, 2001.

[245] P. Viola and M. Jones. Robust real-time object detection. *International Journal on Computer Vision*, 1:511–518, December 2001.

[246] R. J. Radke, S. Andra, O. Al-Kofahi, and B. Roysam. Image change detection algorithms: a systematic survey. *IEEE Transactions on Image Processing*, 14:294–307, March 2005.

[247] D. Rowe. *Towards robust multiple target tracking in unconstrained human populated environments*, chapter 2 Reviewing detection and tracking approaches. Universitat Autònoma de Barcelona, Spain, 2008.

[248] J. Shen. Motion detection in color image sequence and shadow elimination. *Visual Communications and Image Processing, California, USA*, 5308:731–740, 2004.

[249] A. A. Shafie, H. Fadhlan, and M. H. Ali. Motion detection techniques using optical flow. *World Academy of Science, Engineering and Technology*, 56, 2009.

[250] G. L. Foresti, L. Marcenaro, and C. S. Regazzoni. Automatic detection and indexing of video-event shots for surveillance applications. *Multimedia, IEEE Transactions on*, 4(4):459–471, 2002.

[251] S. Dasiopoulou V., Mezaris I, Kompatsiaris, V.-K. Papastathis, and M. G. Strintzis. Knowledge-assisted semantic video object detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 15(10):1210–1224, 2005.

[252] M. S. Lew. Content-based multimedia information retrieval: State of the art and challenges. In *ACM Transactions on Multimedia Computing, Communications, and Applications*, 2008.

[253] G. Zelinsky and A. Todor. The role of "rescue saccades" in tracking objects through occlusions. *Journal of Vision*, 10(7):132, 2010.

[254] D.A. Forsyth and J. Ponce. *Computer vision: a modern approach*, chapter 2, pages 22–56. Prentice-Hall, 2003.

[255] Y. Rathi, N. Vaswani, A. Tannenbaum, and A. Yezzi. Particle filtering for geometric active contours with application to tracking moving and deforming objects. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, pages 2–9, 2005.

[256] C. Hue, J.-P. Le Cadre, and P. Perez. A particle filter to track multiple objects. In *IEEE Workshop on Multi-Object Tracking*, pages 61–68, July 2001.

[257] P. Withagen, K. Schutte, and F. Groen. Object detection and tracking using a likelihood based approach. In *Advanced School for Computing and Imaging Conference*, volume 2, pages 248–253, Lochem, The Netherlands, June 2002.

[258] M. S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp. A tutorial on particle filters for online nonlinear/non-gaussian bayesian tracking. *IEEE Transactions on Signal Processing*, 50(2):174–188, Feb 2002.

[259] K. Okuma, A. Taleghani, N. de Freitas, J. J. Little, and D. G. Lowe. A boosted particle filter: Multitarget detection and tracking. In *European Conference on Computer Vision*, volume 1, pages 28–39, 2004.

[260] X. Xu and B. Li. Head tracking using particle filter with intensity gradient and color histogram. In *IEEE International Conference on Multimedia and Expo*, pages 888–891, July 2005.

[261] Y. Rubner, J. Puzicha, C. Tomasi, and J. M. Buhmann. Empirical evaluation of dissimilarity measures for color and texture. *Computer Vision and Image Understanding*, 84:25–43, October 2001.

[262] R. van der Merwe, A Doucet, J. F. G. de Freitas, and E. Wan. The unscented particle filter. In *Advances in Neural Information Processing Systems*, volume 8, pages 351–357, 2000.

[263] S. Gangaputra and D. Geman. A unified stochastic model for detecting and tracking faces. In *The 2nd Canadian Conference on Computer and Robot Vision 2005*, pages 306–313, May 2005.

[264] S. McKenna and S. Gong. Tracking faces. In *Second International Conference on Automated Face and Gesture Recognition*, pages 271–276, Killington, Vermont, October 1996.

[265] M. Xu, R. Nniu, and P. K. Varshney. Detection and tracking of moving objects in image sequences with varying illumination. In *International Conference on Image Processing*, volume 4, pages 2595–2598, Oct 2004.

[266] M. G. S. Bruno and J. M. F. Moura. Integration of bayes detection and target tracking in real clutter image sequences. In *IEEE National Radar Conference*, pages 234–238, Atlanta GA, USA, May 2001.

[267] P. Willett, R. Niu, and Y. Bar-Shalom. Integration of bayes detection with target tracking. In *IEEE Transactions on Signal Processing*, volume 49, pages 17–29, Jan 2001.

[268] D. Marr. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information.* San Francisco: W. H. Freeman and Company, 1982.

[269] M. Cerf, J. Harel, W. Einhäuser, and C Koch. Predicting human gaze using low-level saliency combined with face detection. *Advances in neural information processing systems*, 20, 2008.

[270] O. V. Komogortsev, D. V. Gobert, S. Jayarathna, D. Koh, and S. Gowda. Standardization of automated analyses of oculomotor fixation and saccadic behaviors. *IEEE Transactions on Biomedical Engineering*, 57:2635–2645, 2010.

[271] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, pages 511–518, Kauai, Hawaii, Dec 2001.

[272] P. Viola, M. Jones, and D. Snow. Detecting pedestrians using patterns of motion and appearance. In *Proc. of International Conference on Computer Vision Systems*, volume 2, pages 734–741, October 2003.

[273] G. Bradski, A. Kaehler, and V. Pisarevsky. Learning-based computer vision with intel's open source computer vison library. *Intel Technology Journal*, 9:119–130, May 2005.

[274] Y. Freund and R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997.

[275] R. Kasturi. *Performance evaluation protocol for face, person and vehicle detection & tracking in video analysis and content extraction (VACE-II).* Computer Science & Engineering University of South Florida, Tampa FL, USA, January 2006.

[276] A. Cavallaro and T. Ebrahimi. Interaction between high-level and low-level image analysis for semantic video object extraction. *EURASIP Journal on Applied Signal Processing*, 6:786–797, June 2004.

[277] L. R. Dice. Measures of the amount of ecologic association between species. *Ecology*, 26:297–302, 1945.

[278] E. Burg, C. N. L. Olivers, A. W. Brunkhorst, and J. Theeuwes. Audiovisual events capture attention: Evidence from temporal order judgements. *Journal of Vision*, 8(5):1–10, 2008.

[279] R. C. McGivern and J. M. Gibson. Characterisation of ocular fixation in humans by analysis of saccadic intrusions and fixation periods: A pragmatic approach. *Vision Research*, 46:3741–3747, 2006.

[280] J. D. Wilder, C. D. Aitkin, B. S. Schnitzer, A. Cohen, and E. Kowler. The timing of oculomotor fixations. *Journal of Vision*, 10(7):498, 2010.

[281] J. M. Henderson. Stimulus discrimination following covert attentional orienting to an exogenous cue. *Journal of experimental psychology. Human perception and performance*, 17(1):91, 1991.

[282] C. L. Folk, R. W. Remington, and J. C. Johnston. Involuntary covert orienting is contingent on attentional control settings. *Journal of experimental psychology. Human perception and performance*, 18(4):1030, 1992.

[283] P. L. Smith, R. Ellis, D. K. Sewell, and B. J. Wolfgang. Cues detection with compound integration-interruption masks reveals multiple attentional mechanisms. *Jounal of Vision*, 10(5):1–29, 2010.

[284] L. Shahbazyan and B. Kokinov. The effect of objecthood on processing efficiency. In N. Taatgen and H. van Rijn, editors, *COGSCI 2009. The annual meeting of the cognitive science commutity*, pages 373–378, 2009.

[285] R. Stiefelhagen. Tracking focus of attention in meetings. In *Fourth IEEE International Conference on Multimodal Interfaces*, pages 273–280, 2002.