UNIVERSITY OF CALIFORNIA, SAN DIEGO

Biologically Plausible Algorithms for Motion Saliency and Tracking

A dissertation submitted in partial satisfaction of the
requirements for the degree
Doctor of Philosophy

in

Electrical Engineering (Signal and Image Processing)

by

Vijay Mahadevan

Committee in charge:

Professor Nuno Vasconcelos, Chair
Professor Garrison W. Cottrell
Professor Kenneth Kreutz-Delgado
Professor Truong Nguyen
Professor Lawrence K. Saul

2011

The dissertation of Vijay Mahadevan is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

_____

_____

Chair

University of California, San Diego

2011

DEDICATION

To my parents, Lalitha and V. Mahadevan.

EPIGRAPH

Nothing happens until something moves.

—Albert Einstein

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would first like to thank my advisor Dr. Nuno Vasconcelos for his invaluable guidance and inspiration. It is my hope that some of his passion for research and his perfectionist zeal would have rubbed off on me in these past years. I also thank the members of my doctoral committee, Professors Kenneth Kreutz-Delgado, Truong Nguyen, Gary Cottrell and Lawrence Saul, for their support and active involvement in my research progress. Thanks also to Professor John T. Serences for very helpful discussions and feedback on my research.

My sincere gratitude is due to my past mentors Professor Badrinath Roysam and Dr. Khaled El-Maleh. Professor Roysam's strong conviction that I should pursue a Ph.D was a big motivation in my returning to school. Khaled's encouragement while at Qualcomm was no less helpful. In fact, Khaled's suggestions on what I should work on 6 years ago eventually led to the first part of my dissertation on motion saliency.

I gratefully acknowledge the support from NSF awards IIS-0448609, and IIS-0534985, and a gift from SONY. Rob Rome, Travis Spackman, Patti Amos, M'Lissa Michaelson and Priscilla Haase have been of great help in dealing with administrivia.

I thank all my colleagues - past and present, at SVCL, Dashan, Antoni, Hamed, Sunhyoung, Nikhil, Jose, Kritika, Ehsan, Weixin, Mandar and Mulloy for their help and support. Even the sun-less and frigid confines of SVCL-I became a cheerful workplace due to their presence.

My many friends in San Diego have made living here for the past seven years a total pleasure. Hanging out with the Qualcomm folks - Arjun and Anu, Kapil and Reena, Sitaraman and Aruna, and Shiva made it seem like I had never left either IIT or Qualcomm behind. Friends at UCSD - Ankit, Sethu, Adarsh, the two Nikhils, Jose, the two Mayanks, Gaurav and Kirti, Rathinakumar, Himanshu, Anshu, Kowsik, Natan, Aneesh, Arun, Bharath and many others made it fun to come to school everyday. Most of my time at UCSD was probably spent having lunch/coffee or going to the gym with a time-varying subset of these friends. A special thanks to my friends elsewhere in the world - especially Srinivas, Avinash, Gautham, Jagadish and Bharath, for braving long exposures to cellphone radiation to lend me a patient ear whenever I needed it.

Most of all I thank my family. Their support and unwavering faith in me has

been invaluable in my journey so far. It must be a big relief for everyone that I finally graduated. My parents, especially, would be glad to see that I no longer need to go to school, with a backpack and lunch bag, a routine that has remained unchanged for over a quarter of a century.

The text of Chapter 3, in full, is based on the material as it appears in: V. Mahadevan, and N. Vasconcelos, "Spatiotemporal saliency in dynamic scenes", IEEE Transactions on Pattern Analysis and Machine Intelligence, 32(1), pp. 171-177, 2010; V. Mahadevan, and N. Vasconcelos, "Background Subtraction in Highly Dynamic Scenes", IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 1-6, 2008. The dissertation author was a primary researcher and an author of the cited material.

The text of Chapter 4, in full, is based on the material as it appears in: D. Gao, V. Mahadevan, and N. Vasconcelos, "On the plausibility of the discriminant center-surround hypothesis for visual saliency", Journal of Vision, 8(7):13, 1-18, June 2008. The dissertation author was a primary researcher and an author of the cited material.

The text of Chapter 5, in full, is based on the material as it appears in: V. Mahadevan, and N. Vasconcelos, "Biologically inspired object tracking using center-surround saliency mechanisms", in review, IEEE Transactions on Pattern Analysis and Machine Intelligence; V. Mahadevan, and N. Vasconcelos, "Saliency Based Discriminant Tracking", IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 1007-1013, 2009. The dissertation author was a primary researcher and an author of the cited material.

The text of Chapter 6, in full, is based on the material as it appears in: V. Mahadevan, and N. Vasconcelos, "Biological plausibility of the saliency hypothesis for visual tracking", in preparation. The dissertation author was a primary researcher and an author of the cited material.

The text of Chapter 7, in full, is based on the material as it appears in: V. Mahadevan, and N. Vasconcelos, "Biological plausibility of the saliency hypothesis for visual tracking", in preparation. The dissertation author was a primary researcher and an author of the cited material.

The text of Chapter 8, in full, is based on the material as it appears in: V. Mahadevan, and N. Vasconcelos, "Biological plausibility of the saliency hypothesis for

visual tracking", in preparation. The dissertation author was a primary researcher and an author of the cited material.

The text of Chapter 9, in full, is based on the material as it appears in: V. Mahadevan, and N. Vasconcelos, "Biological plausibility of the saliency hypothesis for visual tracking", in preparation. The dissertation author was a primary researcher and an author of the cited material.

VITA

2002	Bachelor of Technology Electrical Engineering, Indian Institute of Technology, Madras.
2003	Master of Science Electrical Engineering, Rensselaer Polytechnic Institute, Troy, NY.
2004–2006	Engineer Qualcomm Inc, San Diego, CA.
2006–2011	Research Assistant Statistical and Visual Computing Laboratory Department of Electrical and Computer Engineering University of California, San Diego
2011	Doctor of Philosophy Electrical and Computer Engineering, University of California, San Diego

PUBLICATIONS

V. Mahadevan and N. Vasconcelos, Biological plausibility of the saliency hypothesis for tracking, *i*n preparation.

V. Mahadevan and N. Vasconcelos, Biologically inspired object tracking using center-surround saliency mechanisms, *in review, IEEE Trans. on Pattern Analysis and Machine Intelligence*.

V. Mahadevan C-W. Wong, J Costa-Pereira, T.T. Liu, N. Vasconcelos and L.K. Saul, Maximum Covariance Unfolding - Manifold Learning for Bimodal Data, *Neural Information Processing Systems*, Granada, Spain, Dec 2011.

V. Mahadevan and N. Vasconcelos, Automatic Initialization and Tracking Using Attentional Mechanisms, *Workshop on Biologically-Consistent Vision*, Colorado Springs, CO, 2011.

A.B. Chan, V. Mahadevan and N. Vasconcelos, Generalized Stauffer-Grimson Background Subtraction for Dynamic Scenes, *Machine Vision and Applications*, vol. 22(5), 751-766, 2011.

N. Jacobson, Y-L. Lee, V. Mahadevan, N. Vasconcelos and T.Q. Nguyen, A Novel Approach to FRUC using Discriminant Saliency and Frame Segmentation, *IEEE Transactions on Image Processing*, vol. 19(11) ,2924-2934, Nov. 2010.

N. Jacobson, Y-L. Lee, V. Mahadevan, N. Vasconcelos and T.Q. Nguyen, Motion Vector Refinement for FRUC Using Saliency and Segmentation, *IEEE International Conference on Multimedia and Expo (ICME)*, Singapore, Jul 2010.

V. Mahadevan, W. Li, V. Bhalodia and N. Vasconcelos, Anomaly Detection in Crowded Scenes, *IEEE Conference on Computer Vision and Pattern Recognition*, San Francisco, CA, 2010.

H. Masnadi-Shirazi, V. Mahadevan and N. Vasconcelos, On the Design of Robust Classifiers for Computer Vision, *IEEE Conference on Computer Vision and Pattern Recognition*, San Francisco, CA, 2010.

V. Mahadevan and N. Vasconcelos, Spatiotemporal Saliency in Highly Dynamic Scenes, *IEEE Trans. on Pattern Analysis and Machine Intelligence*,32(1), pp. 171-177, 2010.

V. Mahadevan and N. Vasconcelos, Saliency based discriminant tracking, *IEEE Conference on Computer Vision and Pattern Recognition*, Miami, FL, 2009.

V. Mahadevan and N. Vasconcelos, Unsupervised Moving Target Detection in Dynamic Scenes , *Army Science Conference*, Orlando, FL, 2008.

D. Gao, V. Mahadevan and N. Vasconcelos On the plausibility of the discriminant center-surround hypothesis for visual saliency. *Journal of Vision*, 8(7), pp. 1-18, 2008.

V. Mahadevan and N. Vasconcelos, Background subtraction in highly dynamic scenes, *IEEE Conference on Computer Vision and Pattern Recognition*, Anchorage, AK, 2008.

D. Gao, V. Mahadevan and N. Vasconcelos, The discriminant center-surround hypothesis for bottom-up saliency. In *Proc. Neural Information Processing Systems (NIPS)*, Vancouver, Canada, 2007.

H. Narasimha Iyer, V. Mahadevan, J. M. Beach and B. Roysam, Improved Detection of the Central Reflex in Retinal Vessels Using a Generalized Dual Gaussian Model and Robust Hypothesis Testing, *IEEE Transactions on Information Technology in Biomedicine*, vol. 12, no. 3,pp. 406-410, May 2008.

J. A. Tyrell, V. Mahadevan, R. Tong, E. Brown, B. Roysam and R. K. Jain, 3D Model-Based Complexity Analysis of Tumor Microvasculature from in vivo Multiphoton Confocal Images, *Journal of Microvascular Research*, vol. 70, pp. 165-178, 2005.

V. Mahadevan, J. A. Tyrell, R. Tong, E. Brown, B. Roysam and R. K. Jain, Complexity Analysis for Angiogenesis Vasculature, *SPIE Conference on Medical Imaging*, San Diego, Feb 2005.

V. Mahadevan, H. Narasimha Iyer, B. Roysam and H. Tanenbaum, Robust Model-Based Vasculature Detection in Noisy Biomedical Images, *IEEE Transactions on Information Technology in Biomedicine*, vol. 8, no. 3, pp. 360-376, Sept. 2004.

ABSTRACT OF THE DISSERTATION

Biologically Plausible Algorithms for Motion Saliency and Tracking

by

Vijay Mahadevan

Doctor of Philosophy in Electrical Engineering (Signal and Image Processing)

University of California, San Diego, 2011

Professor Nuno Vasconcelos, Chair

Biologically plausible algorithms for motion saliency and visual tracking are proposed. First a spatiotemporal saliency algorithm, based on a center-surround framework, is introduced. The algorithm is inspired by biological mechanisms of motion-based perceptual grouping, and extends a discriminant formulation of center-surround saliency previously proposed for static imagery. Under this formulation, the saliency of a location is equated to the power of a pre-defined set of features to discriminate between the visual stimuli in a center and a surround window, centered at that location. The features are spatiotemporal video patches, and are modeled as dynamic textures, to achieve a principled joint characterization of the spatial and temporal components of saliency. The combination of discriminant center-surround saliency with the modeling

power of dynamic textures yields a robust, versatile, and fully unsupervised spatiotemporal saliency algorithm, applicable to scenes with highly dynamic backgrounds and moving cameras. The related problem of background subtraction is treated as the complement of saliency detection, by classifying non-salient (with respect to appearance and motion dynamics) points in the visual field as background. The algorithm is tested for background subtraction on challenging sequences, and shown to substantially outperform various state of the art techniques. The biological plausibility of the framework is demonstrated by showing that it can predict human psychophysics data on salient moving stimuli.

Second, a biologically inspired discriminant object tracker is proposed. It is argued that discriminant tracking is a consequence of top-down tuning of the saliency mechanisms that guide the deployment of visual attention. The principle of discriminant saliency is then used to derive a tracker that implements a combination of center-surround saliency, a spatial spotlight of attention, and feature based attention. In this framework, the tracking problem is formulated as one of continuous target-background classification, implemented in two stages. The first, or learning stage, combines a focus of attention mechanism and bottom-up saliency to identify a maximally discriminant set of features for target detection. The second, or detection stage, uses a feature based attention mechanism and a target-tuned top-down discriminant saliency detector, to detect the target. Overall, the tracker iterates between learning discriminant features from the target location in a video frame and detecting the location of the target in the next. The statistics of natural images are exploited to derive an implementation which is conceptually simple and computationally efficient. The saliency formulation is also shown to establish a unified framework for classifier design, target detection, automatic tracker initialization, and scale adaptation. Experimental results show that the proposed discriminant saliency tracker outperforms a number of state-of-the art trackers in the literature.

Finally, it is hypothesized that such saliency based tracking model is biologically plausible and could underlie tracking in primate visual systems. This hypothesis, denoted the *saliency hypothesis for tracking*, is tested for plausibility in three ways. First, results from a set of human behavior studies on the connection between saliency

and tracking show that 1) successful tracking requires targets to be salient, 2) tracking success has a dependence on feature contrast, between target and background, that is remarkably similar to that of saliency, and 3) as for widely accepted models of saliency, tracking also involves a center-surround mechanism with the involvement of a localized background. Second, saliency based tracking is shown to be neurophysiologically plausible, by derivation of a tracking network that is fully compliant with the standard physiological models of V1 and MT, and with what is known about attentional control in area LIP. Finally, this network is shown to 1) replicate electrophysiological recordings from MT neurons in feature-based attention experiments, and 2) explain the results of the psychophysics experiments.

Chapter 1

Introduction

1.1 Introduction

The consumer electronics revolution combined with the advent of the Internet age has led to an explosion in the amount of multimedia content being generated. Youtube alone has over 100 million videos most of which are user generated clips with vast variability in terms of video quality. The ubiquity of surveillance cameras has also produced large amounts of video data. This data represents a massive source of information that can be mined for valuable information. For instance, indexing and categorizing videos in Youtube can help in better monetization potential, while scrutinizing surveillance videos is essential for security applications, detection of anomalous events and even modeling consumer behavior.

However, analyzing and extracting useful information from such videos has been a challenge. Video data is far more voluminous than images, and hence the need for automatic analysis is acute. Further, handling video data is significantly harder than for images. In videos, the common problems associated with images, such as lighting, view point variability, and scale are compounded with many other challenges. For instance, the video could be shot using hand-held cameras and could include egomotion, as seen in the videos shot by users in Youtube. Additionally, the objects of interest might be amidst background clutter that itself could be moving or varying from frame to frame. This has prevented the development of reliable solutions for video analysis.

Most of the current approaches for video analysis aim to solve a specific problem

in isolation. For instance, background subtraction in videos has been studied extensively, but the most no solution exists for situations with egomotion. Similarly, tracking of objects in videos has also received considerable attention for over two decades. However, few approaches provide an *integrated framework for automated video analysis* that can be applied to diverse scenarios and for multiple application domains.

On the other hand, the human visual system is extremely efficient at the task of perceiving and processing moving stimuli. In the human visual system, one of the most important mechanisms driving rapid scene perception is visual saliency. This enables higher level cognitive processing to focus attention only on the most salient locations of the visual field, and allows complex visual tasks to be solved with modest amounts of computation. Further, in biological visual systems, saliency and tracking tasks are not fundamentally different. Once a target is declared salient, it is likely to stay salient for some period of time. It appears sensible to use the computations already performed for saliency to keep track of where the object is. Hence, it can be argued that there is some evolutionary pressure for a common solution to the two problems.

Recent work on the computational modeling of saliency [65] has led to efficient saliency algorithms that have been applied to computer vision problems involving static images. This has led to improved approaches for interest point detection [62], and object recognition [59, 69] in images. Inspired by this work, we propose an extension to the model that can compute motion saliency. We further show that, as in biological systems, the same framework used for motion saliency can be used to track objects in videos. We show this by constructing an algorithm where tracking could be performed through top-down tuning of the mechanisms already in place for bottom-up saliency. This provides a unified view of motion saliency and tracking. Further it reduces target initialization to a special case of discriminant tracking that can be handled using the same computational principles. The framework results in an automated tracker that can be used for surveillance and monitoring.

Finally, we study the connection between tracking and saliency and try to find evidence of its biological plausibility. For this we perform human behavior studies to show that tracking and saliency could be sharing a common underlying mechanism. We also attempt to construct a biologically plausible network model that uses the same

architecture to solve both saliency and tracking.

The specific contributions of this work are discussed below.

1.2 Contributions

1.2.1 Motion saliency and background subtraction

We propose an algorithm for motion saliency and show that this algorithm replicates the psychophysics of salient moving stimuli. Based on this motion saliency algorithm, we derive a robust and versatile procedure for background subtraction, which is successful even for scenes with highly dynamic backgrounds and those shot using moving cameras. This is the first such solution for scenes with moving cameras and extremely dynamic backgrounds.

1.2.2 Discriminant tracking

We show that discriminant tracking follows naturally from the discriminant formulation of visual saliency. In particular, tracking can be implemented with a combination of bottom-up center-surround discriminant saliency and spatial attention for learning, feature-based attention for feature selection, and top-down saliency for target detection. We also show how the same framework can be used to automatically identify the size of the target in each frame. This provides a unified solution to the problems of classifier design, target detection, automatic tracker initialization, and scale adaptation.

1.2.3 Evidence of biological plausibility of discriminant tracking

We suggest that the connections between saliency and tracking exploited in the discriminant saliency tracker could be the basis of tracking in biological visual systems. We provide evidence that supports this hypothesis in three ways. First, human behavior experiments show that tracking requires discrimination between target and background using a center-surround mechanism, and that tracking reliability and saliency have a common dependence on feature contrast. Second, the hypothesis is shown to be neurophysiologically plausible, through construction of a tracking model that can be imple-

mented with widely accepted models of cortical computation. Specifically, a tracking model based on MT neurons is constructed, and it is shown that saliency based tracking can be implemented with a feature selection mechanism akin to the well known phenomenon of feature-based attention in MT. Finally, this tracker is shown to accurately replicate electrophysiological data from MT neurons, and the results of the human behavior experiments.

1.3 Organization of the thesis

A brief review of discriminant center surround saliency is presented in Chapter 2. This is followed by the motion saliency and background subtraction algorithm in Chapter 3. The biological plausibility of the motion saliency algorithm is discussed in Chapter 4. The saliency based approach for tracking algorithm is developed in Chapter 5. Chapter 6 introduces the saliency hypothesis for tracking, while Chapter 7 details the human behavior studies that validate the hypothesis. The biologically plausible version of the saliency based tracker is constructed in Chapter 8. Validation of the model on the human behavior data and electrophyisiological data is presented in Chapter 9.

Chapter 2

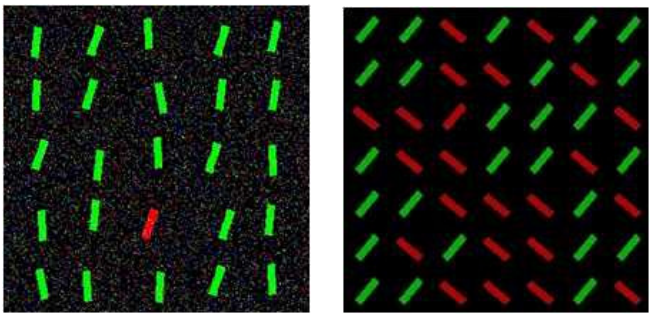
Discriminant Saliency

2.1 Visual saliency

The perception of complex scenes by biological vision systems is heavily dependent on attentional mechanisms. These mechanisms allocate the limited perceptual resources available to the scene regions that matter the most, increasing efficiency and robustness to clutter. Attention is itself driven by saliency mechanisms, which assign to each region of the visual field a degree of saliency, or importance. The different regions of the scene are then explored sequentially, according to their saliency. There are two types of saliency mechanisms, commonly denoted *bottom-up* and *top-down*. Bottom-up saliency is completely stimulus driven, i.e. independent of the higher level goals of the perceptual system. It is, for example, responsible for the high saliency of a "danger" sign posted on a wall, which *pops-out* [124] even when we are not looking for danger signs.

One of its common manifestations is the *pop-out* phenomena [124] illustrated by the left display of Figure 2.1. When subjects are instructed to find an outlying bar (target) in this display, they locate the red bar immediately, independently of the total number of green bars (distractors). The red bar is highly salient and "pops-out", commanding attention.

Top-down saliency mechanisms can be tuned by feedback from high-level cortical areas, according to the tasks to be performed. For example, the eye fixations of a subject trying to identify a person in a photograph will be overwhelmingly located on



(a) (b)

Figure 2.1: Two types of saliency mechanisms (a) bottom-up: the red bar is salient among the green distractors and "pops-out" (b) top-down: there is no immediate pop-out, but when attention is focused on red-bars, the bar in the 3^{rd} row and 3^{rd} column can easily be singled as the odd one out.

the faces present in that picture [188]. Two main types of tuning are possible: a *spatial focus of attention* mechanism, also known as the spotlight of attention [137], and *feature-based attention* [168] which manipulates attention by inhibiting or enhancing groups of features.

Both can be perceived by inspecting the display on the right of Figure 2.1. Like the one on the left, this display contains a target that is different from all distractors. However, the difference can only be perceived through a conjunction of color and orientation (the target shares the color of some distractors and the orientation of others), and because feature conjunctions are not salient [167], there is no effect of pop-out. This makes the target much more difficult to identify than on the left display. Nevertheless, as the reader can verify, the target is easily identifiable when the subject is instructed to look at the vicinity of the bar in the 3^{rd} row and 3^{rd} column. This results from the subject directing his/her spotlight of attention at the target, in response to the additional location information. While this spotlight narrows down the field of view, feature-based attention performs the equivalent of feature selection. Its mechanisms manipulate attention by inhibiting or enhancing groups of features. This can, again, be perceived by inspecting the display of Figure 2.1(b), where the non-salient target becomes salient when subjects are instructed to concentrate on the red bars. Once the green bars are ignored, the target differs from the distractors only in terms of orientation, and there is a percept of pop-out.

2.2 Computational Models for Saliency

While many bottom-up [99, 88, 105, 139, 113, 93, 27, 108, 152, 70, 74, 57] and top-down [63, 149, 179, 183, 55, 20, 9, 120] saliency algorithms have been proposed in the computer and biological vision literature, most approaches lack a unified computational theory to explain both the saliency mechanisms. However, there have been Bayesian formulations for saliency [163, 192, 65] that provide a unified framework for both modes.

Bayesian models for saliency are rooted in a decision theoretic interpretation of perception: that perceptual systems evolve with the goal of producing decisions about the state of the surrounding environment that are *optimal in a decision-theoretic sense*. In this context, the computation of saliency leads to a classification based framework: that salient features are those which best allow visual systems to decide between different hypothesis, regarding the nature of the visual stimulus. For bottom-up saliency search, these two hypotheses are that the stimulus either belongs to the target or background classes, while for top-down saliency search,they correspond to the stimulus belonging to either the target or the distractor classes.

In addition to the several saliency models that have an explicit Bayesian formulation [163, 192], there are many others which can be interpreted as specific cases of the formulation [26].

In the next section, we review the recently proposed decision theoretic formulation for saliency termed *discriminant saliency* [65].

2.3 Discriminant saliency

Discriminant saliency is a mathematical formulation for visual saliency. Two classes of visual stimuli are first defined: a *target* class of stimuli of interest, and a *background* or null hypothesis of non-salient stimuli. The visual stimulus is not observed directly, but through projection into a number of features. Saliency is the result of optimal classification (in a decision-theoretic sense) of feature responses into the *target* and *background* hypotheses [63]. More precisely, the saliency of each location of the visual field is equated to the *expected classification accuracy* for the features extracted from it.

Locations of smallest probability of error are most salient.

This formulation can be applied to various vision problems, by suitable definition of target and null hypotheses. For example, it can be used to implement one-vs-all object detection, by defining the target as an object class, and the null hypothesis as the set of natural images [63]. This is an instance of top-down saliency, due to the necessity of specifying task-related object classes. Alternatively, target and null hypotheses can be defined as the visual stimuli contained in a pair of *center* and *surround* windows, at every location of the visual field [65]. This is a purely stimulus driven definition, which implements bottom-up saliency. Implementations of the discriminant saliency principle have various properties of interest for both biological and computer vision. In the area of biological modeling, they can be mapped into a biologically plausible neural architecture, which has been shown to 1) replicate the computations of the standard neurophysiological model of the visual cortex [65], 2) predict a large body of psychophysics of human saliency [60], and 3) accurately predict human fixations in natural scenes [61]. In computer vision, they have been shown successful for interest point detection [62], and object recognition [59, 69].

2.4 Mathematical Formulation of Discriminant Saliency

Discriminant saliency is defined with respect to two classes of stimuli: the class of *stimuli of interest*, and the *background* or null hypothesis, consisting of stimuli that are not salient. The locations of the visual field that can be classified, with lowest expected probability of error, as containing stimuli of interest are denoted as salient. This is accomplished by setting up a binary classification problem which opposes the stimuli of interest to the null hypothesis. The saliency of each location in the visual field is then equated to the discriminant power (expected classification error) of the visual features extracted from that location in differentiating the two classes.

Formally, let $\mathcal{V}$ be a d dimensional dataset indexed by location vector $l \in L \subset \mathcal{R}^d$ and consider the responses to visual stimuli of a predefined set of features $\boldsymbol{Y}$ (e.g. raw pixel values, Gabor or Fourier features), computed from $\mathcal{V}$ at all locations $l \in L$. A classification problem opposing two classes, of class label $C(l) \in \{0, 1\}$, is posed at

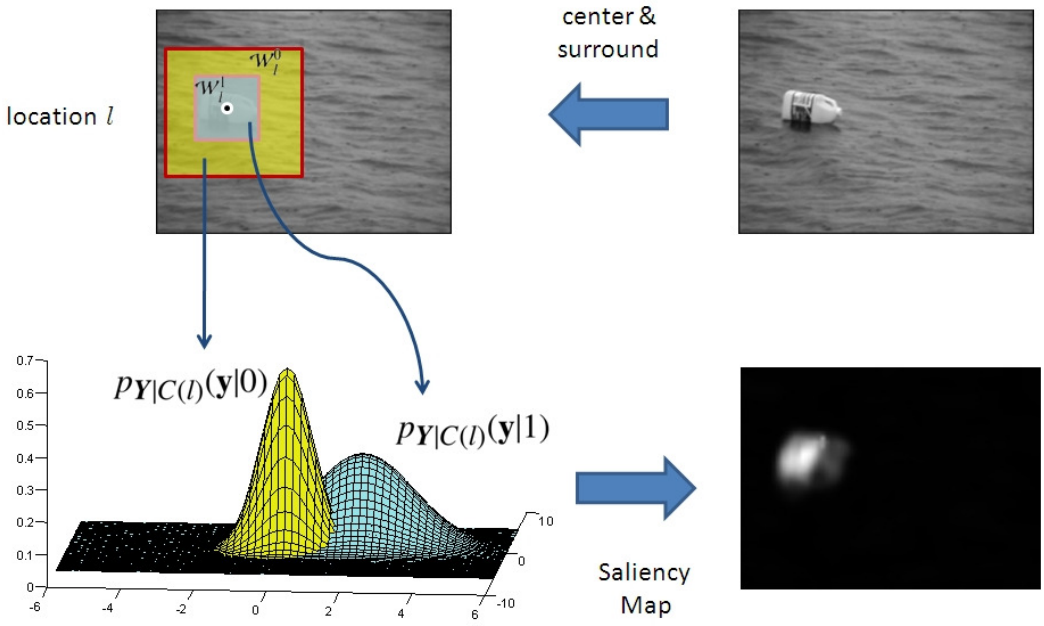


Figure 2.2: Illustration of discriminant center-surround saliency. Center and surround windows are defined around each image location, and the distribution of a previously defined set of features $\boldsymbol{Y}$ estimated from the two windows. The saliency of the location is a measure of how disjoint the two feature distributions are.

location l. Two windows are defined: a neighborhood $\mathcal{W}_l^1$ of l which is denoted as *center*, and a surrounding annular window $\mathcal{W}_l^0$ which is denoted as the *surround*. The union of the two windows is denoted the *total* window, $\mathcal{W}_l = \mathcal{W}_l^0 \cup \mathcal{W}_l^1$. Let $\mathbf{y}^{(j)}$ be the vector of feature responses at location j. Features in the center, $\{\mathbf{y}^{(j)} | j \in \mathcal{W}_l^1\}$, are drawn from the class of interest (or alternate hypothesis) $C(l) = 1$, with probability density $p_{\boldsymbol{Y}|C(l)}(\mathbf{y}|1)$. Features in the surround, $\{\mathbf{y}^{(j)} | j \in \mathcal{W}_l^0\}$, are drawn from the null hypothesis $C(l) = 0$, with probability density $p_{\boldsymbol{Y}|C(l)}(\mathbf{y}|0)$. An illustration of the center-surround classification problem, for a static image, is shown in Figure 2.2.

The saliency of location l, $S(l)$, is the extent to which the features $\boldsymbol{Y}$ can discriminate between *center* and *surround*. This is quantified by the mutual information between features, $\boldsymbol{Y}$, and class label, C,

$$S(l) \quad = \quad I_l(\boldsymbol{Y}; C) = \sum_{c=0}^{1} \int p_{\boldsymbol{Y},C(l)}(\mathbf{y}, c) \log \frac{p_{\boldsymbol{Y},C(l)}(\mathbf{y}, c)}{p_{\boldsymbol{Y}}(\mathbf{y}) p_{C(l)}(c)} d\mathbf{y}. \tag{2.1}$$

This mutual information is an approximation to the expected probability of correct clas-

sification (more precisely one minus the Bayes error rate) of the classification problem that opposes center to surround [172]. Hence, a large $S(l)$ implies that center and surround have a large disparity of feature responses, i.e. large *local feature contrast*, which enables their discrimination with low probability of error. Conversely, the locations where the classification has the smallest expected probability of error can be identified by searching for maxima of $S(l)$. The function $S(l), l \in L$ is referred to as the *saliency map* of the dataset $\mathcal{V}$. It can also be written as

$$S(l) = \sum_{c=0}^{1} p_{C(l)}(c) \int p_{\boldsymbol{Y}|C(l)}(\mathbf{y}|c) \log \frac{p_{\boldsymbol{Y}|C(l)}(\mathbf{y}|c)}{p_{\boldsymbol{Y}}(\mathbf{y})} d\mathbf{y} \tag{2.2}$$

$$= \sum_{c=0}^{1} p_{C(l)}(c) \mathrm{KL}\left(p_{\boldsymbol{Y}|C(l)}(\mathbf{y}|c) \,\|p_{\boldsymbol{Y}}(\mathbf{y})\right) \tag{2.3}$$

where

$$\mathrm{KL}(p\,\|q) = \int_{\mathcal{X}} p_{\boldsymbol{X}}(x) \log \frac{p_{\boldsymbol{X}}(x)}{q_{\boldsymbol{X}}(x)} dx.$$

is the Kullback-Leibler (KL) divergence between the probability distributions $p_{\boldsymbol{X}}(x)$ and $q_{\boldsymbol{X}}(x)$ [101]. This allows an alternative interpretation of saliency as a measure of the average distance between the feature distribution over each window and the average of the two distributions. This is a measure of the (lack of) overlap between the distributions associated with center and surround.

2.4.1 Mathematical formulation of top down saliency

For top-down saliency problems, such as object recognition [63, 59], the target class, of label $C = 1$, is the object class to recognize, and the background class, with label $C = 0$, the class of natural images. Features $\boldsymbol{Y}$ have probability $p_{\boldsymbol{Y}|C}(\mathbf{y}|1)$ under the target hypothesis and probability $p_{\boldsymbol{Y}|C}(\mathbf{y}|0)$ under the background hypothesis. Unlike bottom-up saliency, where the absence of any objects can be salient (e.g. a void region is salient within a textured background), recognition requires the detection of the object of interest. This implies that top-down saliency measures must have a bias towards target presence.

This bias is accomplished with a two-step saliency measure. A likelihood ratio test is first used to identify the set of likely target locations

$$\mathbf{S} = \left\{ l \left| \frac{P_{C,\boldsymbol{Y}}(1, \mathbf{y}(l))}{P_C(1) P_{\boldsymbol{Y}}(\mathbf{y}(l))} > \frac{P_{C,\boldsymbol{Y}}(0, \mathbf{y}(l))}{P_C(0) P_{\boldsymbol{Y}}(\mathbf{y}(l))} \right. \right\}. \tag{2.4}$$

These are the locations where the likelihood of the feature responses is larger under the hypothesis of target presence than target absence. As before, the saliency of location l is defined by the amount of information in the visual stimulus for optimal classification into one of the two classes, using the information measure

$$I(C; \boldsymbol{Y} = \mathbf{y}(l)) \quad = \quad \sum_{i=0}^{1} p_{C|\boldsymbol{Y}}(\mathbf{y}(l)|i) \log \frac{p_{\boldsymbol{Y},C}(\mathbf{y}(l), i)}{p_{\boldsymbol{Y}}(\mathbf{y}(l)) p_C(i)}. \tag{2.5}$$

However, to guarantee that only locations likely to contain the target are declared salient, the saliency computation is restricted to $\mathbf{S}$. This leads to the saliency measure [59, 69]

$$S(l) = \begin{cases} I(C; \boldsymbol{Y} = \mathbf{y}(l)) & \text{if } l \in \mathbf{S} \\ 0, & \text{otherwise.} \end{cases} \tag{2.6}$$

Locations where this measure is large have both 1) larger likelihood under the target than background hypothesis, and 2) feature responses that are highly informative for classification.

In the following Chapters, we use these formulations to construct biologically plausible algorithms for motion saliency and visual tracking.

Chapter 3

Motion Saliency and Background Subtraction

3.1 Introduction

Natural scenes are usually composed of several dynamic entities. As illustrated by Figure 3.1, objects of interest often move amidst complicated backgrounds that are themselves moving, e.g. swaying trees, other objects such as a crowd, or a flock of birds, moving water, waves, and snow, rain, or smoke-filled environments.

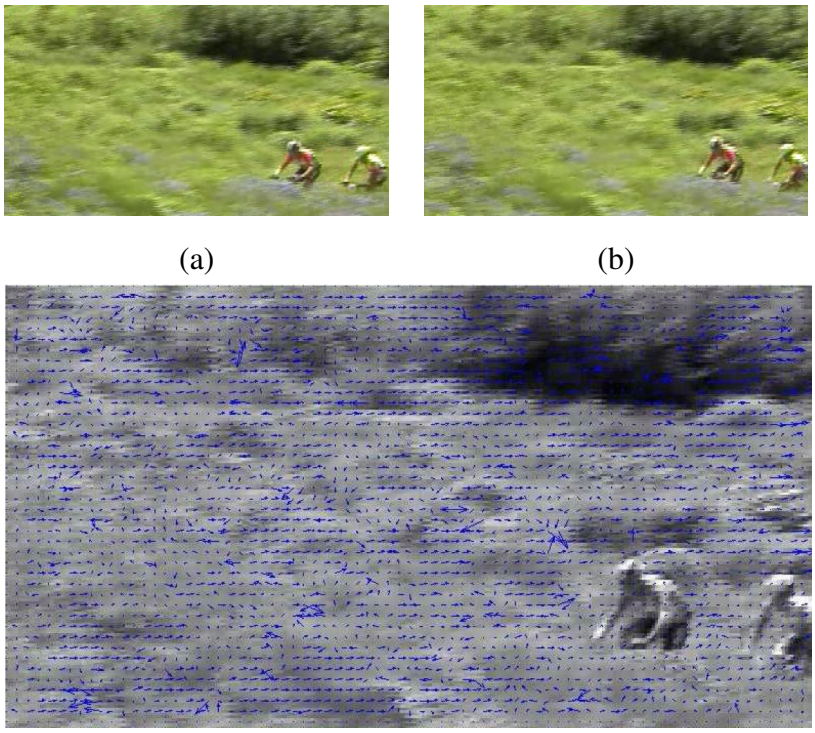
Even when the scene is static, egomotion of the imaging sensor can originate a highly variable background, as shown in Figure 3.2. In the most extreme situations, egomotion and scene motion can combine to produce quite complex background motion patterns. We refer to scenes with any of these types of variability as *dynamic scenes*. Since such scenes are plentiful in the natural world, successful discrimination between the background motions they induce and moving objects of interest (henceforth denoted as *foreground objects*) is a strong survival advantage, for example in terms of being able to identify potential predators or prey. Not surprisingly, biological visual systems have evolved to be extremely efficient in this task [14, 79].

In computer vision, background subtraction is useful for diverse applications. Algorithms that can produce reliable "figure-ground" segmentation are used as a pre-processing step for object and event detection [44], activity and gesture recognition [185],

Figure 3.1: Examples of dynamic scenes. A skier skiing amidst falling snow, a surfer riding a wave, birds frolicking in moving water, a helicopter flying amidst heavy smoke.

tracking [190], surveillance [52], and video retrieval [176]. For example, in robotic path planning, an autonomous device could benefit from a background subtraction module to simplify the task of identifying objects that approach it. Unlike biological vision, background subtraction has proven quite challenging for computer vision. After decades of research on this problem (see [154] for a review), there has been little progress in the development of methods that are robust and generic enough to handle the complexities of most natural dynamic scenes. In result, even the most advanced techniques exhibit at least one from a number of common shortcomings. For example, various approaches rely on the assumption of a static camera, and are unsuitable for video shot with hand-held cameras or from moving platforms (as in the robotics scenario) [158, 51, 118, 162]. In fact, the dominant approach to background subtraction in the presence of egomotion is to first explicitly [119], or approximately [143], compensate for the camera motion, and then rely on background subtraction techniques that assume a stationary imaging sensor. Accurate compensation of egomotion is, however, cumbersome and can be quite difficult when the background is itself dynamic.

Another frequent shortcoming is the adoption of several (often unjustified) as-



(a) (b)

(c)

Figure 3.2: (a) and (b) Two frames from a video sequence shot with a panning camera that tracks two cyclists riding against a mostly stationary background. (c) the optical flow information overlaid on (a). The background is highly variable, but there is no consistent pattern of optical flow in the region of the foreground objects.

sumptions regarding the motion of the foreground objects. For instance, it is often assumed that these objects move in a consistent direction (an assumption that we denote as *temporal coherence*) [182, 104, 29], and have faster variations in appearance than the background [154]. As illustrated by Figure 3.2, such assumptions are particularly questionable when there is egomotion, e.g., a camera that tracks a moving object. The figure shows two consecutive frames from a video sequence shot with a panning camera, in a manner such that the foreground objects, viz. the two cyclists, undergo very small motion in image coordinates. Figure 3.2 (c) shows the optical flow between the two frames. While the background changes rapidly, there is no consistent pattern of flow in the foreground region, where the flow is indeed close to null. This type of attentional tracking is the norm in biological vision, where eyes tend to follow objects of interest, and desirable for most computer vision applications that involve moving objects, as

demonstrated by the very extensive literature in object tracking [190]. Nevertheless, the inversion (with respect to the stationary camera scenario) of the motion characteristics of background (which is, in this case, fast moving and temporally coherent) and foreground (whose motion is barely existent and mostly random) can be a major challenge for existing background subtraction techniques.

A third common shortcoming is the requirement of an explicit model of the background scene. This implies a bootstrapping phase, where the algorithm is presented with frames containing only the background [118, 158, 196] or where the background is *estimated* by batch processing, e.g. median filtering [44], of a large number of video frames. We refer to these techniques as *implicitly supervised*, and to the initial phase as a *training* step for learning background parameters. This training has several shortcomings, including the facts that 1) it is difficult to perform for dynamic backgrounds, where the background model must be continuously updated, and 2) it must be repeated for each scene where background subtraction is to be deployed. Furthermore, global background models tend to be brittle, and difficult to manage when there is significant egomotion.

To address these limitations, we propose a novel paradigm for background subtraction. This paradigm is inspired by biological vision, where background subtraction is inherent to the task of deploying visual attention. This can be done in multiple ways, but frequently relies on motion saliency mechanisms, which identify regions of the visual field where objects move differently from the background. We equate background subtraction to the problem of detecting salient motion, and propose a solution based on a generic hypothesis for biological salience, which is referred to as the *discriminant center-surround hypothesis*. Under this hypothesis, bottom-up saliency is formulated as the result of optimal discrimination between center and surround stimuli at each location of the visual field. A set of visual features are collected from the center and surround of each location, and the locations where the discrimination between the features of the two types can be performed with smallest expected probability of error are declared as most salient. Background subtraction is then equivalent to simply ignoring the locations declared as non-salient.

This *strictly local* approach to background subtraction has various advantages over the traditional *global* procedures. First, there is no need to train or maintain a

global model of the background. As the latter changes, so do the surround windows at all locations of the visual field. Thus, the local saliency measures are automatically adapted to variations in the background, and there is no need to keep track of, or update, a global model. Second, background modeling is considerably simplified. While, globally, a dynamic background is rarely homogeneous (e.g. different trees have different motion), the assumption of spatial homogeneity is usually accurate locally. This enables the use of much simpler probabilistic models (e.g. unimodal distributions vs. mixtures) which are easier to learn and update. Third, because discriminant saliency compares the center and surround regions, it depends only on the *relative disparity* between their motion characteristics, and therefore is invariant to camera motion. Finally, discriminant saliency can be adapted to various problems by simply modifying the features and probabilistic models used to discriminate between center and surround. For example, motion features can be complemented with depth measurements, if range sensors are available, and different types of models can be chosen to account for different background dynamics. In this work, we choose dynamic texture models, due to their versatility in modeling complex moving patterns, ability to replicate the motion of natural scenes, and the rich statistical formulations they lend themselves to [49, 43, 193].

Overall, the main contributions of this work are three-fold. First, the proposed algorithm is completely unsupervised and does not require initial training with background only frames. In effect, it is a *bottom-up* approach that can adapt to any situation. Second, due to its *locally discriminant* nature, the algorithm is insensitive to egomotion, and applicable to video shot with moving cameras. Third, by relying on dynamic textures as models for the video, it accounts for joint saliency in motion and appearance in a principled manner, and is robust enough to handle backgrounds of complex dynamics. Experimental results on a diverse collection of sequences with such dynamics shows that the proposed algorithm substantially outperforms the current state-of-the-art in background subtraction.

3.2 Previous Work on Background subtraction

Due to its potential application to a wide range of vision problems, background subtraction is a well studied topic in computer vision. Comprehensive reviews of prior work in this area appear in [154, 135]. One possible taxonomy for existing approaches is to group the different methods according to the spatial and temporal models adopted for video representation. In this context, a popular cue for background subtraction is optical flow. For example, Wixson [182] and Tian et al. [106] find salient regions by grouping pixels that move in a consistent direction. Mittal and Paragios [117] complement optical flow with color, modeling their probability distributions with kernel density estimates. These techniques tend not to perform well when all components of the scene (including the background) are dynamic, or when the variability of background optical flow is larger than that of the foreground objects (as in the panning example of Figure 3.2).

A popular alternative is to rely on appearance-based representations. In this case, the background can be represented as a probability distribution of certain features extracted from each pixel. Individual pixel distributions are estimated over time, updated regularly, and the background model is the union of all pixel models. A pixel is classified as belonging to background or foreground according to its probability under the background model. Examples of pixel representations include a single Gaussian for pixel color [185] and a mixture of Gaussians for pixel intensity [158, 196]. Temporal updating of the Gaussian parameters is sometimes achieved using Kalman filters [95, 100]. A strong limitation of all these approaches is that they do not model spatial appearance. Extensions that address this limitation use local features extracted from a neighborhood of each pixel, including texture [75], spatio-temporal volumes [136], local kernel histograms and contour features [122], or even SIFT-like descriptors [195]. Elgammal et al. propose a non-parametric kernel density estimate (KDE) [51] to model pixel intensity, but incorporate spatial information by searching for the best match of a pixel to classify among all neighboring pixel models. While this adds some robustness to background motion, dynamic scenes pose a significant challenge for global background models built as ensembles of spatially rigid pixel models.

Moving away from pixel representations, various types of region models have also been proposed. These range from image windows, whose statistics are modeled

with time series, to full-fledged models of global image appearance. Among the methods based on time series, Zhong and Scarloff [193] use a patch-based autoregressive moving average (ARMA) model whose parameters are learned with principal component analysis (PCA). A robust Kalman filter is then used to temporally update the model parameters, and predict foreground regions. Monnet et al. [118] also rely on video patches and an autoregressive model, updating the parameters of the latter with an incremental PCA algorithm. The 'Wallflower' method [165] combines a pixel level background model with region and frame level components that allow for spatial homogeneity and increased robustness. Other techniques rely on Markov random fields (MRF) to enforce spatial consistency. In this category, Bugueau and Perez [29] perform mean shift clustering to group optical flow and color information, and use a maximum a posteriori probability (MAP) rule to assign each cluster to either foreground or background. Sheikh and Shah [154] use an MRF to achieve spatial coherence of foreground and background labels. The MRF includes different observation models, learned with KDE, for foreground and background. The foreground model accounts for temporal persistence of foreground objects, and includes an outlier process to identify new foreground regions. Pixels are assigned to background or foreground through an MAP rule. Global methods that use eigenspace models of the entire background have also been proposed [131, 104].

While most of these techniques assume no egomotion, approaches have been developed specifically for non-stationary cameras. For example, camera pan and tilt are modeled probabilistically in [71], where the resulting models are used to compensate for camera motion before resorting to static background subtraction methods. Nevertheless, nearly all the approaches discussed so far rely on an explicit background model, and assume that the algorithm will be initialized with *background-only* frames. Some techniques even rely on strongly supervised training of classifiers, such as support vector machines (SVM), for pixel- or region-level foreground detection, using manually annotated images [162, 36].

A few techniques deviate from region models rigidly defined a priori. These techniques treat video as layers containing foreground and background motion, and equate background subtraction to the extraction of foreground layers [180, 186, 174].

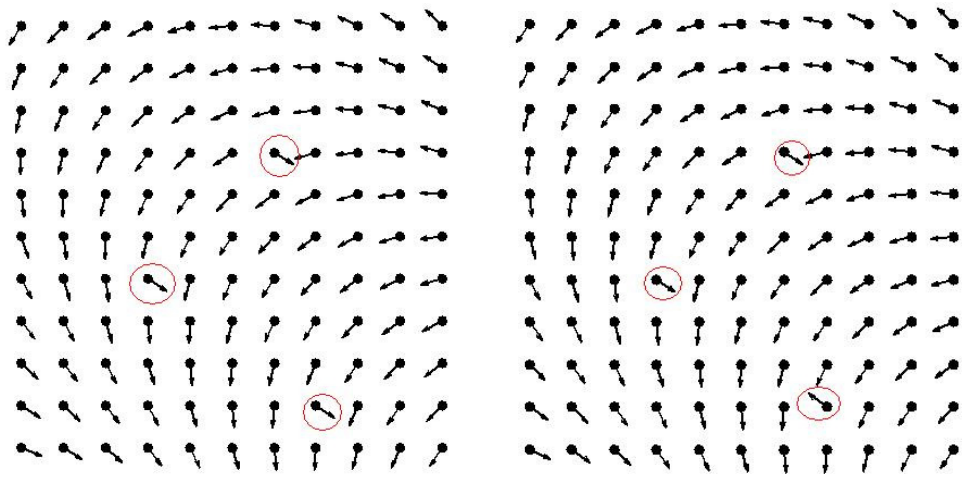
However, the decomposition into layers requires segmentation of the video into all its component objects, a problem which is more complex than background subtraction itself.

Significantly less attention has been given to biologically inspired solutions to background subtraction, or the formulation of background subtraction as the detection of non-salient locations. A notable exception, in this context, is SUNDay [191] and the work of Itti et al., consisting of a number of motion saliency mechanisms [84, 86, 87]. The most popular among these is the "surprise" model of [86, 87] which maintains probabilistic models for several pixel features, and computes divergences between prior and posterior distributions (based on these models) to find regions that are novel. The "surprise" model has been successfully applied to scenes with relatively simple backgrounds, but its performance on scenes with complex dynamic backgrounds, that might themselves trigger spurious "surprise" responses, has not been previously studied in great detail.

3.3 Biological motivation

There is evidence that in biological vision bottom-up saliency is achieved through center surround mechanisms [81, 167, 129, 89], i.e. mechanisms tuned to detect stimuli that are distinct from stimuli in their surround. Extensive psychophysics experiments have shown that these mechanisms can be driven by a variety of features, including intensity, color, orientation or motion, and *local feature contrast* plays a predominant role in the perception of saliency [125]. For example, Nothdurft has shown that simple visual concepts, such as bars, can be very salient when viewed against a background of similar visual concepts that differ from them only in terms of low-level properties, such as color, orientation, or motion [126, 127].

Figure 3.3 shows some displays used in classical psychophysics experiments designed to determine the role of feature contrast on judgments of motion saliency [125, 127]. In one experiment [127], subjects were shown a display of moving dots such as that depicted in Figure 3.3 (a) (the videos are available in [3]). While all dots (whose motion is indicated, in the figure, by arrows) were subject to motion different from that



(a) (b)

Figure 3.3: Saliency perception due to local contrast [127]. Each panel shows a quiver plot of the stimuli (dots, whose direction of motion is indicated by arrows of length proportional to the speed of that motion). In (a), three targets which move in the same direction, amongst a field of distractors, are perceived as the vertices of a moving triangle. When, as in (b), one target moves in a direction different than those of the other two, observers still perceive a moving triangle.

of their immediate neighbors, three (referred to as *the targets*, and indicated by circles in the figure) had *substantially larger motion contrast* than the others. The targets could be in different configurations, two of which are shown in the figure: (i) "similar" (Figure 3.3 (a)) where all three targets moved in the same direction, and (ii) "dissimilar" (Figure 3.3 (b)) where one target moved in a direction *different* than that of the other two. In all cases, subjects reported the percept of *pop-out* of a "moving triangle ", with similar detection rates. While motion pop-out was already well established, these experiments showed that both motion saliency and the perceptual organization of the points into a triangle *do not depend on absolute quantities*, such as the direction of motion of the targets, how coherent their motion is, or the type of background motion. Instead, the fact that the targets are coherently perceived as a triangle, even when the vertex motions are incoherent and the background motion cannot be easily explained by a physical geometric transformation, suggests that both motion saliency and perceptual organization are rooted in measurements of *local motion contrast*. Indeed, neurophysiological experiments on primates have also shown that neurons in the middle temporal visual area (MT) use local motion contrast in a center-surround mechanism that may underlie the

perception of figure-ground motion segmentation [22, 21].

From a computer vision point of view, this also appears to be a good idea. Note that, if motion contrast is defined as dissimilarity of optical flow, the saliency judgments are robust to egomotion. As long as the background motion field is *smooth*, as would typically be the case for a static scene and a moving camera, any foreground objects (either moving or stationary) will be identified as salient with high probability (i.e. barring the unlikely coincidence where object motion is the same as background motion at object location). Furthermore, there is no need for a "global background model", or any type of training. Instead, saliency can be computed *efficiently* with resort to purely local computations, *without any assumptions* on scene geometry, and it *immediately adapts* to previously unseen environments. We will see, in later sections, that these properties still hold for *dynamic scenes*, under a more general definition of motion contrast. On the other hand, these experiments suggest that background subtraction techniques which 1) rely on grouping of features by motion similarity to identify foreground objects, or 2) require compensation of camera motion, will have difficulties to match the performance of biological systems.

3.4 Discriminant Center-Surround Approach for Motion Saliency

Using biology for motivation, we rely on local measurements of motion contrast as the central source of information for the motion saliency detector now proposed. To produce a quantitative measure of saliency we rely on the principle of *discriminant saliency* [63, 61] described in Chapter 2. As described earlier, this is a generic saliency principle, applicable to a broad set of problems. Here we consider bottom-up motion saliency, using a center-surround architecture and motion models which are suitable for dynamic scenes.

The discriminant saliency measure of (2.1) is defined in a generic sense, which does not depend on the type of stimulus or features $\boldsymbol{Y}$. It can be shown that for static saliency, under the popular model of Gabor features and generalized Gaussian natural image statistics [80, 178], it can be mapped into a biologically plausible neural archi-

tecture, which replicates the computations of the standard neurophysiological model of V1 [64, 30]. It is, thus, not surprising that this network also replicates a large body of psychophysics of human saliency [61]. In what follows we show that, by adopting suitable models for spatio-temporal stimulus statistics, this formulation is robust enough to compute motion saliency in highly dynamic scenes. This enables the design of powerful background subtraction algorithms by simple reduction of background subtraction to the complement of saliency detection.

3.5 Background subtraction

Under the definition of saliency as the expected accuracy of the classification problem which opposes stimulus at location and surround, locations of minimal saliency are those where the distinction between the two stimuli can be made with *lowest confidence*. This provides a natural, objective, definition of *background* based on *strictly local* computations: background points are those of lowest center-surround saliency. Under this formulation, the design of a background subtraction algorithm capable of handling highly dynamic scenes only requires the use, in (2.3), of probability models $p_{\mathbf{Y}|C(l)}(\mathbf{y}|c)$ that can capture the variability associated with such scenes. We adopt the dynamic texture (DT) model of [49], due to its ability to account for this variablity, while jointly modeling the spatial and temporal characteristics of the visual stimulus in an elegant unified stochastic framework.

3.5.1 Modeling spatio-temporal stimulus statistics

A DT is an autoregressive generative model that represents the appearance of the stimulus $\mathbf{y}_t \in \mathbb{R}^m$ (the two-dimensional image stimulus is first converted into a column vector of length m), observed at time t, as a linear function of a hidden state process $\mathbf{x}_t \in \mathbb{R}^n$ ($n << m$) subject to Gaussian observation noise. The state and appearance processes form a linear dynamical system (LDS)

$$\begin{aligned} \mathbf{x}_t &= \mathbf{A}\mathbf{x}_{t-1} + \mathbf{v}_t \\ \mathbf{y}_t &= \mathbf{C}\mathbf{x}_t + \mathbf{w}_t \end{aligned} \tag{3.1}$$

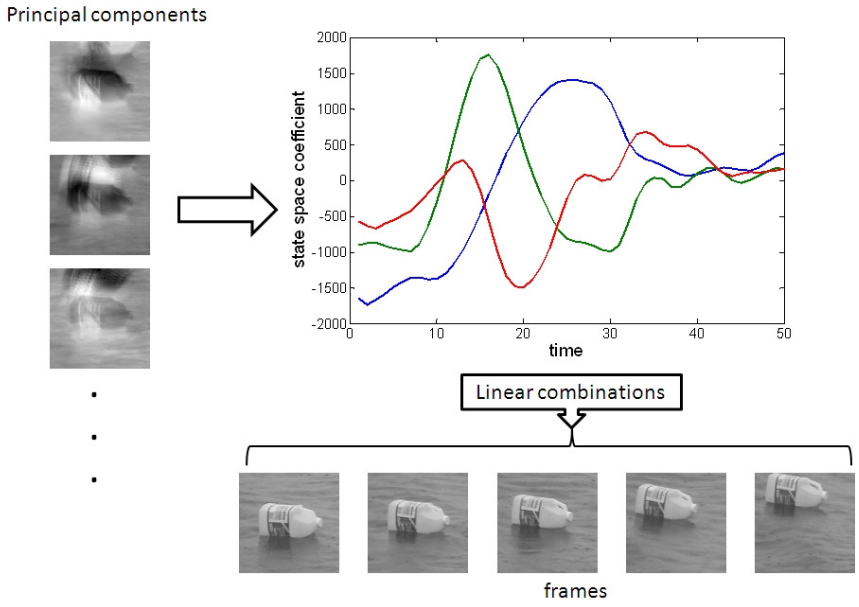


Figure 3.4: Illustration of a dynamic texture model. The first three basis images are shown on the left, and the corresponding state space variables plotted as a function of time. At each time instant, a video frame is represented as a linear combination of the basis images, with weights given by the value of the corresponding state variable.

where $\mathbf{A} \in \mathbb{R}^{n \times n}$ is the state transition matrix, $\mathbf{C} \in \mathbb{R}^{m \times n}$ the observation matrix, and $\mathbf{v}_t \sim_{iid} \mathcal{N}(0, \mathbf{Q})$ and $\mathbf{w}_t \sim_{iid} \mathcal{N}(0, \mathbf{R})$ are Gaussian state and observation noise processes, respectively. The initial state is assumed to be distributed as $\mathbf{x}_1 \sim \mathcal{N}(\boldsymbol{\mu}_1, \mathbf{S}_1)$, and the model is parameterized by $\boldsymbol{\Theta} = (\mathbf{A}, \mathbf{C}, \mathbf{Q}, \mathbf{R}, \boldsymbol{\mu}_1, \mathbf{S}_1)$. The hidden state space sequence $\mathbf{x}_t$ is a first order Markov chain that encodes stimulus dynamics, while $\mathbf{y}_t$ is a linear combination of prototypical basis functions (the columns of $\mathbf{C}$) and encodes the appearance component of the stimulus at time t. Dynamic texture modeling of a sequence of images is illustrated in Figure 3.4[1].

3.5.2 Learning dynamic textures

Given center and surround regions, DT parameters could in principle be learned by maximum likelihood (using expectation-maximization [155], or N4SID [132]). However, due to the high dimensionality of video sequences, these solutions are too complex

[1]The bottle sequence from [193] is used in this example.

for motion saliency. A suboptimal alternative, that works well in practice[49], is to learn the spatial and temporal parameters separately. Given N sequences, $\mathbf{y}_{1:\tau}^{(1)}, \ldots, \mathbf{y}_{1:\tau}^{(N)}$, of τ frames each (where $\mathbf{y}_{1:\tau}^{(i)} = [\mathbf{y}_1^{(i)} \ldots \mathbf{y}_\tau^{(i)}]$), sampled from a DT, let $\mathbf{Y}_{1:\tau} = [\mathbf{y}_{1:\tau}^{(1)}, \ldots, \mathbf{y}_{1:\tau}^{(N)}] \in \mathbb{R}^{m \times N\tau}$ be the matrix composed by concatenating all sequences. If $\mathbf{Y}_{1:\tau} = \mathbf{U}\mathbf{S}\mathbf{V}^T$ is its singular value decomposition (SVD), the DT parameters are estimated as follows,

$$\hat{\mathbf{C}} = \mathbf{U}[1:n] \text{ (first n columns of } \mathbf{U}) \tag{3.2}$$

$$\hat{\mathbf{x}}_{1:\tau}^{(i)} = \hat{\mathbf{C}}^T \mathbf{y}_{1:\tau}^{(i)} \tag{3.3}$$

$$\hat{\mathbf{A}} = \hat{\mathbf{X}}_{2:\tau} (\hat{\mathbf{X}}_{1:\tau-1})^\dagger \tag{3.4}$$

$$\hat{\mathbf{Q}} = \frac{1}{N(\tau-1)} \sum_{i=1}^{N} \sum_{j=1}^{\tau-1} \hat{\mathbf{v}}_j^{(i)} (\hat{\mathbf{v}}_j^{(i)})^T \tag{3.5}$$

$$\hat{\mathbf{R}} = \frac{1}{N(\tau-1)} \sum_{i=1}^{N} \sum_{j=1}^{\tau-1} \hat{\mathbf{w}}_j^{(i)} (\hat{\mathbf{w}}_j^{(i)})^T \tag{3.6}$$

$$\tag{3.7}$$

where, $\hat{\mathbf{X}}_{1:\tau} = [\hat{\mathbf{x}}_{1:\tau}^{(1)}, \ldots, \hat{\mathbf{x}}_{1:\tau}^{(N)}]$ is the matrix of state estimates, $\mathbf{M}^\dagger$ the pseudo-inverse of $\mathbf{M}$, $\hat{\mathbf{v}}_t^{(i)} = \hat{\mathbf{x}}_{t+1}^{(i)} - \hat{\mathbf{A}}\hat{\mathbf{x}}_t^{(i)}$, and $\hat{\mathbf{w}}_t^{(i)} = \mathbf{y}_t^{(i)} - \hat{\mathbf{C}}\hat{\mathbf{x}}_t^{(i)}$, for $t \in 1 \ldots \tau$. Finally, the initial state parameters are estimated as,

$$\hat{\boldsymbol{\mu}}_1 = \frac{1}{N} \sum_{i=1}^{N} \hat{\mathbf{x}}_1^{(i)} \tag{3.8}$$

$$\hat{\mathbf{S}}_1 = \frac{1}{N} \sum_{i=1}^{N} \hat{\mathbf{x}}_1^{(i)} (\hat{\mathbf{x}}_1^{(i)})^T - \hat{\boldsymbol{\mu}}_1 \hat{\boldsymbol{\mu}}_1{}^T \tag{3.9}$$

3.5.3 Probability Distributions

Since the states of a DT form a Markov process with Gaussian conditional probability for $\mathbf{x}_t$ given $\mathbf{x}_{t-1}$ (for any t), and Gaussian initial state conditions, the joint distribution of the state sequence, $\mathbf{x}(\tau) = \begin{bmatrix} \mathbf{x}_1^T & \cdots & \mathbf{x}_\tau^T \end{bmatrix}^T$, is also Gaussian [35]

$$p_{\boldsymbol{X}}(\mathbf{x}(\tau)) \sim \mathcal{N}(\boldsymbol{\mu}(\tau), \boldsymbol{\Sigma}(\tau)) \tag{3.10}$$

with parameters defined by the recursions,

$$\boldsymbol{\mu}(\tau) = \begin{bmatrix} \boldsymbol{\mu}(\tau-1) \\ \boldsymbol{\mu}_\tau \end{bmatrix} \tag{3.11}$$

$$\boldsymbol{\Sigma}(\tau) = \begin{bmatrix} \boldsymbol{\Sigma}(\tau-1) & \boldsymbol{\Upsilon}^T(\tau) \\ \boldsymbol{\Upsilon}(\tau) & \mathbf{S}_\tau \end{bmatrix}, \tag{3.12}$$

where

$$\boldsymbol{\mu}_\tau = \mathbf{A}\boldsymbol{\mu}_{\tau-1} \tag{3.13}$$

$$\mathbf{S}_\tau = \mathbf{A}\mathbf{S}_{\tau-1}\mathbf{A}^T + \mathbf{Q} \tag{3.14}$$

$$\boldsymbol{\Upsilon}(\tau) = \begin{bmatrix} \mathbf{A}\boldsymbol{\Upsilon}(\tau-1) & \mathbf{A}\mathbf{S}_{\tau-1} \end{bmatrix}, \tag{3.15}$$

for $\tau \geq 2$ and, $\boldsymbol{\mu}(1) = \boldsymbol{\mu}_1$, $\boldsymbol{\Sigma}(1) = \mathbf{S}_1$, $\boldsymbol{\Upsilon}(2) = \mathbf{A}\mathbf{S}_1$.

Similarly, the sequence of observations $\mathbf{y}(\tau) = \begin{bmatrix} \mathbf{y}_1^T & \cdots & \mathbf{y}_\tau^T \end{bmatrix}^T$, has joint distribution

$$p_{\boldsymbol{Y}}(\mathbf{y}(\tau)) \sim \mathcal{N}(\boldsymbol{\gamma}(\tau), \boldsymbol{\Phi}(\tau)) \tag{3.16}$$

with parameters defined by the recursions,

$$\boldsymbol{\gamma}(\tau) = \begin{bmatrix} \boldsymbol{\gamma}(\tau-1) \\ \mathbf{C}\boldsymbol{\mu}_\tau \end{bmatrix} \tag{3.17}$$

$$\boldsymbol{\Phi}(\tau) = \begin{bmatrix} \boldsymbol{\Phi}(\tau-1) & \boldsymbol{\zeta}^T(\tau)\mathbf{C}^T \\ \mathbf{C}\boldsymbol{\zeta}(\tau) & \mathbf{C}\mathbf{S}_\tau\mathbf{C}^T + \mathbf{R} \end{bmatrix} \tag{3.18}$$

$$\boldsymbol{\zeta}(\tau) = \begin{bmatrix} \mathbf{A}\boldsymbol{\zeta}(\tau-1) & \mathbf{A}\mathbf{S}_{\tau-1}\mathbf{C}^T \end{bmatrix} \tag{3.19}$$

for $\tau \geq 2$, and $\boldsymbol{\gamma}(1) = \mathbf{C}\boldsymbol{\mu}_1$, $\boldsymbol{\Phi}(1) = \mathbf{C}\mathbf{S}_1\mathbf{C}^T + \mathbf{R}$, $\boldsymbol{\zeta}(2) = \mathbf{A}\mathbf{S}_1\mathbf{C}^T$. Using the parameter estimates obtained with (3.2)-(3.9), from a collection of spatio-temporal patches extracted from the center and surround windows, in (3.10) and (3.16) produces the probability distributions required by (2.3), for the center, surround, and total windows.

3.5.4 KL Divergence between DTs

The final step for the computation of $S(l)$, with (2.3), is the evaluation of the KL divergence between DTs. Let $p_{\boldsymbol{Y}|C(l)}(\mathbf{y}(\tau)|i) \sim \mathcal{N}(\boldsymbol{\gamma}_i(\tau), \boldsymbol{\Phi}_i(\tau)), i \in \{0, 1\}$ be the

class-conditional probabilities of a sequence of τ frames under two DTs parameterized by $\boldsymbol{\Theta}_i(l), i \in \{0, 1\}$, respectively, and $p_{\boldsymbol{Y}}(\mathbf{y}(\tau)) \sim \mathcal{N}(\boldsymbol{\gamma}(\tau), \boldsymbol{\Phi}(\tau))$ the probability under the marginal DT parametrized by $\boldsymbol{\Theta}(l)$.

Since all distributions are Gaussian, the KL divergence between the densities has the closed-form [42]

$$\mathrm{KL}\left(p_{\boldsymbol{Y}|C(l)}(\mathbf{y}(\tau)|i) \,||\, p_{\boldsymbol{Y}}(\mathbf{y}(\tau))\right) = \tag{3.20}$$

$$\frac{1}{2}\left[\log\frac{|\boldsymbol{\Phi}(\tau)|}{|\boldsymbol{\Phi}_i(\tau)|} + \mathrm{tr}\left(\boldsymbol{\Phi}(\tau)^{-1}\boldsymbol{\Phi}_i(\tau)\right) + ||\boldsymbol{\gamma}_i(\tau) - \boldsymbol{\gamma}(\tau)||^2_{\boldsymbol{\Phi}(\tau)} - m\tau\right] \tag{3.21}$$

where m is the number of pixels in each frame, and

$$||\mathbf{z}||_{\mathbf{A}} = \mathbf{z}^T\mathbf{A}^{-1}\mathbf{z}$$

is the Mahalanobis norm of $\mathbf{z}$ with respect to covariance $\mathbf{A}$, $|\mathbf{A}|$ the determinant of $\mathbf{A}$, and $tr(\mathbf{A})$ its trace. Direct evaluation of (3.21) is intractable since the matrices $\boldsymbol{\Phi}(\tau)$, $\boldsymbol{\Phi}_i(\tau)$ have size $m\tau \times m\tau$. Using several matrix identities, it is possible to rewrite all terms in recursive form, which only requires $n\tau \times n\tau$ matrices (recall that n is the dimension of the state space and $n \ll m$). The recursions are derived in full generality in [34]. Here, we only consider the case where the image noise is independently distributed, i.e., where the covariances $\mathbf{R}$, $\mathbf{R}_i$ of the noise term $\mathbf{w}_t$ in (3.1) are diagonal, $\mathbf{R} = \sigma^2\mathbf{I}$, $\mathbf{R}_i = \sigma_i^2\mathbf{I}$.

Using the recursive definitions for means $\boldsymbol{\gamma}(\tau)$ and $\boldsymbol{\mu}(\tau)$ and covariances $\boldsymbol{\Phi}(\tau)$, and $\boldsymbol{\Sigma}(\tau)$ (in (3.17), (3.11), (3.18), (3.12) respectively), the Mahalanobis term of (3.21) can be written as :

$$||\boldsymbol{\gamma}_i(\tau) - \boldsymbol{\gamma}(\tau)||^2_{\boldsymbol{\Phi}(\tau)} = ||\boldsymbol{\gamma}_i(\tau-1) - \boldsymbol{\gamma}(\tau-1)||^2_{\boldsymbol{\Phi}(\tau-1)} + ||\mathbf{z}_i||^2_{\hat{\boldsymbol{\Phi}}} \tag{3.22}$$

where the update term is given by,

$$||\mathbf{z}_i(\tau)||^2_{\hat{\boldsymbol{\Phi}}} = \frac{1}{\sigma^2}||\mathbf{z}_i(\tau)||^2 - \frac{1}{\sigma^4}\mathbf{z}_i^T(\tau)\mathbf{C}\boldsymbol{\Gamma}^{-1}(\tau)\mathbf{C}^T\mathbf{z}_i(\tau) \tag{3.23}$$

$$\mathbf{z}_i(\tau) = \frac{1}{\sigma^2}\mathbf{C}\boldsymbol{\Upsilon}(\tau)\Delta(\tau)\boldsymbol{\nu}_i(\tau-1) - \boldsymbol{\gamma}_{i,\tau} + \boldsymbol{\gamma}_\tau \tag{3.24}$$

$$\boldsymbol{\nu}_i(\tau-1) = \begin{bmatrix} \boldsymbol{\nu}_i(\tau-2) \\ \mathbf{C}^T\mathbf{C}_i\boldsymbol{\mu}_{i,\tau-1} - \boldsymbol{\mu}_{\tau-1} \end{bmatrix}, \boldsymbol{\nu}_i(1) = \mathbf{C}^T\mathbf{C}_i\boldsymbol{\mu}_{i,1} - \boldsymbol{\mu}_1, \tag{3.25}$$

$$\Delta(\tau) = \mathbf{I} - \frac{1}{\sigma^2}\boldsymbol{\beta}(\tau) \tag{3.26}$$

$$\boldsymbol{\Gamma}(\tau) = \left[\mathbf{S}_\tau - \frac{1}{\sigma^2}\boldsymbol{\Upsilon}(\tau)\Delta(\tau)\boldsymbol{\Upsilon}^T(\tau)\right]^{-1} + \frac{1}{\sigma^2}\mathbf{I} \tag{3.27}$$

with,

$$\boldsymbol{\beta}(\tau) = \begin{bmatrix} \mathbf{H}^{-1}(\tau) & \mathbf{H}^{-1}(\tau)\mathbf{G}^T(\tau) \\ \mathbf{G}(\tau)\mathbf{H}^{-1}(\tau) & \boldsymbol{\beta}(\tau-1) + \mathbf{G}(\tau)\mathbf{H}^{-1}(\tau)\mathbf{G}^T(\tau) \end{bmatrix} \tag{3.28}$$

where

$$\mathbf{H}(\tau) = \Xi + \frac{1}{\sigma^2} - \Omega^T\mathbf{H}^{-1}(\tau-1)\Omega \tag{3.29}$$

$$\Xi = \mathbf{S}_1^{-1} + \mathbf{A}^T\mathbf{Q}^{-1}\mathbf{A} \tag{3.30}$$

$$\Omega = -\mathbf{Q}^{-1}\mathbf{A} \tag{3.31}$$

$$\mathbf{G}(\tau) = -\begin{bmatrix} \mathbf{H}^{-1}(\tau-1)\Omega \\ \mathbf{G}(\tau-1)\mathbf{H}^{-1}(\tau-1)\Omega \end{bmatrix} \tag{3.32}$$

with initial conditions $\mathbf{G}(2) = -\boldsymbol{\beta}(1)\Omega$, $\mathbf{H}(2) = \Xi + \frac{1}{\sigma^2}\mathbf{I} - \Omega^T\boldsymbol{\beta}(1)\Omega$ and $\boldsymbol{\beta}(1) = \left[\mathbf{S}_1^{-1} + \frac{1}{\sigma^2}\mathbf{I}\right]^{-1}$. This computation requires the inverse of $\boldsymbol{\Gamma}(\tau)$ and $\left[\mathbf{S}_\tau - \frac{1}{\sigma^2}\boldsymbol{\Upsilon}(\tau)\Delta(\tau)\boldsymbol{\Upsilon}^T(\tau)\right]$, both $n \times n$ matrices. The other matrices requiring inversion are $\mathbf{H}(\tau)$ and $\left[\mathbf{S}_1^{-1} + \frac{1}{\sigma^2}\mathbf{I}\right]$, also of size $n \times n$.

The trace term has recursion

$$\operatorname{tr}\left[\boldsymbol{\Phi}^{-1}(\tau)\boldsymbol{\Phi}_i(\tau)\right] = \omega_i(\tau) - \operatorname{tr}[\boldsymbol{\beta}(\tau)\boldsymbol{\Psi}_i(\tau)] \tag{3.33}$$

where

$$\omega_i(\tau) = \frac{1}{\sigma^2}\operatorname{tr}[\mathbf{S}_{i,\tau}] + m\frac{\sigma_i^2}{\sigma^2} - \frac{\sigma_i^2}{\sigma^4}\operatorname{tr}[\mathbf{H}^{-1}(\tau)] - \frac{\sigma_i^2}{\sigma^4}\operatorname{tr}[\mathbf{H}^{-1}(\tau)\mathbf{G}^T(\tau)\mathbf{G}(\tau)] + \omega_i(\tau-1)$$

$$\boldsymbol{\Psi}_i(\tau) = \begin{bmatrix} \boldsymbol{\Psi}_i(\tau-1) & \xi_i^T(\tau)\mathbf{T}_i^T \\ \mathbf{T}_i\xi_i(\tau) & \frac{1}{\sigma^4}\mathbf{T}_i\mathbf{S}_{i,\tau}\mathbf{T}_i^T \end{bmatrix} \tag{3.34}$$

$$\xi_i(\tau) = \frac{1}{\sigma^4}\begin{bmatrix} \mathbf{A}\xi_i(\tau-1) & \mathbf{A}\mathbf{S}_{i,\tau-1}\mathbf{T}_i^T \end{bmatrix}, \tag{3.35}$$

with $\mathbf{T}_i = \mathbf{C}^T\mathbf{C}_i$, and initial conditions, $\omega_i(1) = \frac{1}{\sigma^2}\operatorname{tr}[\mathbf{S}_{i,1}] + m\frac{\sigma_i^2}{\sigma^2} - \frac{\sigma_i^2}{\sigma^4}\operatorname{tr}\left[\mathbf{S}_1^{-1} + \frac{1}{\sigma^2}\mathbf{I}\right]$, $\boldsymbol{\Psi}_i(2) = \frac{1}{\sigma^4}\mathbf{T}_i\mathbf{S}_{i,1}\mathbf{T}_i^T$, $\xi_i(1) = \frac{1}{\sigma^4}\mathbf{A}\mathbf{S}_{i,1}\mathbf{T}_i^T$. Note that, because $\boldsymbol{\beta}(\tau)$ and $\boldsymbol{\Psi}_i(\tau)$ are symmetric matrices of equal size, the trace of their product is simply the sum of the entries of the Hadamard product. Finally, the determinant of $\boldsymbol{\Phi}(\tau)$ is given by

$$\log|\boldsymbol{\Phi}(\tau)| = \sum_{k=1}^{n}\log\left(\frac{\lambda^{(k)}}{\sigma^2} + 1\right) + m\log\sigma^2 \tag{3.36}$$

where $\lambda^{(k)}$ is the k^{th} eigenvalue of $\boldsymbol{\Sigma}(\tau)$. This reduces the problem of computing the determinant of an $m \times m$ covariance matrix to that of computing the n eigenvalues of $\boldsymbol{\Sigma}(\tau)$. The determinant of $\boldsymbol{\Phi}_i(\tau)$ can be computed in a similar manner.

3.5.5 Recursive Evaluation of KL Divergence

Direct computation of the KL divergence between image sequences containing τ frames is intractable since the covariance matrices Φ_i are $m\tau \times m\tau$, where m is the number of pixels per frame. Using several matrix identities, it is possible to rewrite the terms of the KL divergence in a recursive form that is computationally efficient and only requires storing $n\tau \times n\tau$ matrices, where n is the dimension of the state space ($n \ll m$). The recursions are derived in [34] and, for completeness, reproduced here without derivation. We only consider the case where the image noise is indpendently distributed, i.e., where the covariance matrices $\mathbf{R}$ of the noise term $\mathbf{w}_t$ in (3.1) are diagonal, $\mathbf{R}_i = \sigma_i^2\mathbf{I}, i \in \{0, 1\}$.

Denoting matrices (and vectors) of (3.21) at time τ as $\mathbf{A}_\tau$. For simplicity, we refer to the image at time τ as $\mathbf{y}$, the state at time τ as $\mathbf{x}$, the sequence of preceding $\tau - 1$ images as $\mathbf{Y}$, and the sequence of preceding states as $\mathbf{X}$. The means γ_i and μ_i and covariances $\mathbf{\Phi}_i$, and $\mathbf{\Sigma}_i$, $i \in \{0, 1\}$, of (3.10) and (3.16) are defined recursively as

$$\gamma_1^\tau = \begin{bmatrix} \gamma_1^{\tau-1} \\ \gamma_{1\mathbf{y}} \end{bmatrix}, \mu_1^\tau = \begin{bmatrix} \mu_1^{\tau-1} \\ \mu_{1\mathbf{x}} \end{bmatrix}, \mathbf{\Phi}_1^\tau = \begin{bmatrix} \mathbf{\Phi}_1^{\tau-1} & \mathbf{\Phi}_{1\mathbf{Yy}} \\ \mathbf{\Phi}_{1\mathbf{yY}} & \mathbf{\Phi}_{1\mathbf{yy}} \end{bmatrix}, \mathbf{\Sigma}_1^\tau = \begin{bmatrix} \mathbf{\Sigma}_1^{\tau-1} & \mathbf{\Sigma}_{1\mathbf{Xx}} \\ \mathbf{\Sigma}_{1\mathbf{xX}} & \mathbf{\Sigma}_{1\mathbf{xx}} \end{bmatrix} \quad (3.37)$$

$$\gamma_2^\tau = \begin{bmatrix} \gamma_2^{\tau-1} \\ \gamma_{2y} \end{bmatrix}, \quad \Phi_2^\tau = \begin{bmatrix} \phi_2^{\tau-1} & \phi_{2Yy} \\ \phi_{2yY} & \phi_{2yy} \end{bmatrix} \quad (3.38)$$

Similarly, we can define $\mu_{1x}, \mu_{1X}, \Sigma_{1XX}, \Sigma_{1Xx}, \Sigma_{1xX}, \Sigma_{1xx}$ for the probability of a state sequence under p_1, and likewise for p_2.

Mahalanobis Distance Term

For the Mahalanobis distance, we have the following recursion,

$$\|\gamma_1\tau - \gamma_2\tau\|^2_{\Phi_2\tau} = \left\|\gamma_1^{\tau-1} - \gamma_2^{\tau-1}\right\|^2_{\Phi_2^{\tau-1}} + \|z\|^2_{\Phi_2} \quad (3.39)$$

where

$$\|z\|^2_{\hat{\Phi}_2} = \frac{1}{\sigma_2^2}\|z\|^2 - \frac{1}{\sigma_2^4}\|z\|^2_{(\mathbf{C}_2\hat{\gamma}_2^{-1}\mathbf{C}_2^T)^{-1}} \tag{3.40}$$

$$z = \frac{1}{\sigma_2^2}\mathbf{C}_2\Sigma_{2xX}\Delta_2(\mathbb{T}\mu_1^{\tau-1} - \mu_2^{\tau-1}) - \gamma_{1\mathbf{y}} + \gamma_{2\mathbf{y}} \tag{3.41}$$

$$\Delta_2 = \mathbf{I} - \frac{1}{\sigma_2^2}(\boldsymbol{\beta}_2^{\tau})^{-1} \tag{3.42}$$

$$\boldsymbol{\beta}_2^{\tau} = (\Sigma_2^{\tau})^{-1} + \frac{1}{\sigma_2^2}\mathbf{I} \tag{3.43}$$

$$\hat{\gamma}_2 = \gamma_2^{-1} + \frac{1}{\sigma_2^2}\mathbf{I} \tag{3.44}$$

$$\gamma_2 = \Sigma_{2xx} - \frac{1}{\sigma_2^2}\Sigma_{2xX}\Delta_2\Sigma_{2Xx} \tag{3.45}$$

and $\mathbb{T} = \mathbb{C}_2^T\mathbb{C}_1$ is a $n(\tau-1)\times n(\tau-1)$ block diagonal matrix with $\mathbf{T} = \mathbf{C_2^T C_1}$ at each of its diagonal entries. The computation of the distance requires the inverse of γ_2 and $\hat{\gamma}_2$, both $n \times n$ matrices, and $\boldsymbol{\beta}_2$, an $n(\tau-1)\times n(\tau-1)$ matrix. The inverse of $\boldsymbol{\beta}_2\tau$ can be computed recursively with.

$$(\boldsymbol{\beta}_2\tau)^{-1} = \begin{bmatrix} V_\tau^{-1} & V_\tau^{-1}U_\tau^T \\ U_\tau V_\tau^{-1} & (\boldsymbol{\beta}_2^{\tau-1})^{-1} + U_\tau V_\tau^{-1}U_\tau^T \end{bmatrix} \tag{3.46}$$

where

$$V_\tau = \Xi + \frac{1}{\sigma_2^2} - \boldsymbol{\Omega}^T V_{\tau-1}^{-1}\boldsymbol{\Omega} \tag{3.47}$$

$$\Xi = \mathbf{Q}_2^{-1} + \mathbf{A}^T\mathbf{Q}_2^{-1}\mathbf{A}_2 \tag{3.48}$$

$$\boldsymbol{\Omega} = -\mathbf{Q}_2^{-1}\mathbf{A}_2 \tag{3.49}$$

$$U_\tau = -\begin{bmatrix} V_{\tau-1}^{-1}\boldsymbol{\Omega} \\ U_{\tau-1}V_{\tau-1}^{-1}\boldsymbol{\Omega} \end{bmatrix} \tag{3.50}$$

$$\tag{3.51}$$

with initial conditions $U_2 = -(\boldsymbol{\beta}^1)^{-1}\boldsymbol{\Omega}$, $V_2 = \Xi + \frac{1}{\sigma^2}\mathbf{I} - \boldsymbol{\Omega}^T(\boldsymbol{\beta}^1)^{-1}\boldsymbol{\Omega}$ and $\boldsymbol{\beta}^1 = Q + \frac{1}{\sigma^2}\mathbf{I}$. The only matrices requiring inversion are V_τ and $\boldsymbol{\beta}^1$, both $n \times n$ matrices.

Trace term

The trace term is

$$\mathrm{tr}\left[(\Phi_2\tau)^{-1}\Phi_1\right] = \alpha_\tau - \mathrm{tr}[(\boldsymbol{\beta}_2\tau)^{-1}\Psi_\tau] \tag{3.52}$$

where

$$\alpha_\tau = \frac{1}{\sigma_2^2}\text{tr}[\Sigma_{1xx}] + m\frac{\sigma_1^2}{\sigma_2^2} - \frac{\sigma_1^2}{\sigma_2^4}\text{tr}[V_\tau^{-1}] - \frac{\sigma_1^2}{\sigma_2^4}\text{tr}[V_\tau^{-1}U_\tau^T U_\tau] + \alpha_{\tau-1} \tag{3.53}$$

$$\Psi_\tau = \begin{bmatrix} \Psi^{\tau-1} & \frac{1}{\sigma_2^4}\mathbb{T}\Sigma_{1Xx}\mathbf{T}^T \\ \frac{1}{\sigma_2^4}\mathbf{T}\Sigma_{1xX}\mathbb{T}^T & \frac{1}{\sigma_2^4}\mathbf{T}\Sigma_{1xx}\mathbf{T}^T \end{bmatrix} \tag{3.54}$$

$$\tag{3.55}$$

Note that $\boldsymbol{\beta}_2\tau$ and Ψ_τ are symmetric matrices with the same size, thus the trace of their product is simply the sum of the entries of the Hadamard product.

Determinant Term

Finally, the determinants of $\Phi_i^\tau, i \in \{0, 1\}$ are given by

$$\log\left|\Phi_i^\tau\right| = \sum_{k=1}^{n} \log\left(\frac{\lambda_i^{(k)}}{\sigma_i^2} + 1\right) + m\log\sigma_i^2 \tag{3.56}$$

where $\lambda_i^{(k)}$ is the k^{th} eigenvalue of $\boldsymbol{\Sigma}_i^\tau$. This reduces the problem of computing the determinant of a $m \times m$ covariance matrix to that of computing the n eigenvalues of $\boldsymbol{\Sigma}$.

3.5.6 Background subtraction algorithm

Background pixels are identified by computing the saliency measure $S(l)$ at each location l. Center and surround windows are centered at the location, and a collection of spatio-temporal patches extracted from each window. DT parameters are then learned from the center, surround, and total windows, to obtain the densities $p_{\boldsymbol{Y}|C(l)}(\mathbf{y}(\tau)|1)$, $p_{\boldsymbol{Y}|C(l)}(\mathbf{y}(\tau)|0)$, and $p_{\boldsymbol{Y}}(\mathbf{y}(\tau))$, respectively. $S(l)$ is finally computed with (2.3), using the recursive implementation of (3.21) given by (3.22)-(3.36). The procedure is summarized in Algorithm 1, and illustrated in Figure 3.5. All locations whose saliency is below a threshold are assigned to the background.

3.6 Experimental evaluation

To evaluate background subtraction performance, Algorithm 1 was tested on 18 sequences collected on the web. Frames from some of these sequences are shown in

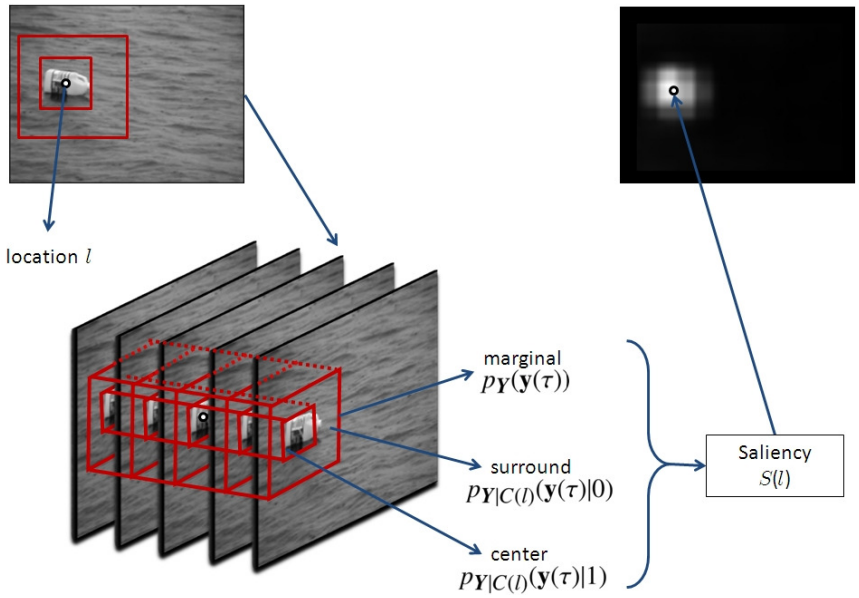


Figure 3.5: Illustration of the center and surround windows used to compute the saliency of location l. Conditional distributions are learned from the center and surround window, while the marginal distribution is learned from the total window. The saliency measure $S(l)$ is finally computed with (2.3).

panel (a) of Figures 3.6 - 3.11. In all cases, the background is highly dynamic, consisting of water, smoke, fire or even a flock of birds. In addition, most sequences were shot with significant camera motion. Figure 3.6 (a), presents frames from a sequence which depicts two people skiing in a heavy snowfall. The sequence of Figure 3.7 (a) shows a surfer riding a low frequency sweeping wave, which is interspersed with high frequency components due to turbulent wakes (due to the surfer and the crest of the sweeping wave), creating significant challenges for background subtraction. A pair of cyclists ride through a grassy plain in Figure 3.8 (a), birds walk against a background of wave crests in Figure 3.9 (a), a helicopter flies amidst heavy smoke in Figure 3.10 (a), and a boat moves through water, against a background of flying birds, in Figure 3.11 (a).

3.6.1 Comparison to previous methods

To compare the performance of the proposed algorithm (denoted in short as Disc-Sal) with existing methods, we selected four representatives of the current state of the art in background subtraction - the modified Gaussian mixture model (GMM) of [196, 2],

Algorithm 1 Computing Discriminant Center Surround Motion Saliency

1: **Input:** Given video $\mathcal{V}$ indexed by location vector $l \in L \subset \mathcal{R}^3$, state-space dimension n, center window size n_c, patch size n_p, temporal window τ.
2: **for** $l \in L$ **do**
3: Identify center $\mathcal{W}_l^1$ and surround $\mathcal{W}_l^0$.
4: List all overlapping patches of size $n_p \times n_p \times \tau$ in $\mathcal{W}_l^1$ and $\mathcal{W}_l^0$
5: From the patches learn dynamic texture parameters for center $\mathbf{\Theta}_1(l)$, surround $\mathbf{\Theta}_0(l)$ and the total $\mathbf{\Theta}(l)$ using (3.2)-(3.9).
6: Compute the mutual information, $S(l)$, between class-conditional and total densities (2.3), using the recursive implementation of (3.21) given by (3.22)-(3.36).
7: **end for**
8: **Output:** Saliency map for $S(l), l \in L$

the non-parametric kernel density estimator (KDE) of [51], the linear dynamical model of Monnet et al. [118], and the "surprise" model proposed by Itti and Baldi [86, 85]. The original implementation of Monnet et al. [118] is not publicly available, and the algorithm requires explicit training with background frames. Since no training data was available for the sequences considered, we implemented an adaptive version, where the auto-regressive model parameters were estimated from the 20 frames preceding the location under consideration. The higher adaptiveness of this version allows for a fairer comparison to saliency-based background subtraction.

The sequences were converted to grayscale, and at each pixel location, the center window occupied 16×16 pixels and spanned 11 frames - 5 past frames, the current frame, and 5 frames in the future ($n_c = 16, \tau = 11$). A causal version of Algorithm 1 (denoted DiscSal-Causal) was also implemented, by considering only the current and 10 past frames. In both cases, the surround window was set to 6 times the size of the center (i.e $96 \times 96 \times 11$). DTs with a 10-dimensional state space, patch dimension $n_p = 8$, and temporal dimension $\tau = 11$, were learned using overlapping $8 \times 8 \times 11$ patches from the center and surround windows. Saliency maps obtained with DiscSal, DiscSal-Causal, Surprise, KDE, Monnet, and GMM are shown in panels (b)-(f), respectively, of Figures 3.6-3.11 (since the results for DiscSal and the causal version, DiscSal-Causal, were very similar we omit the latter). Videos of the maps obtained for all sequences are avail-

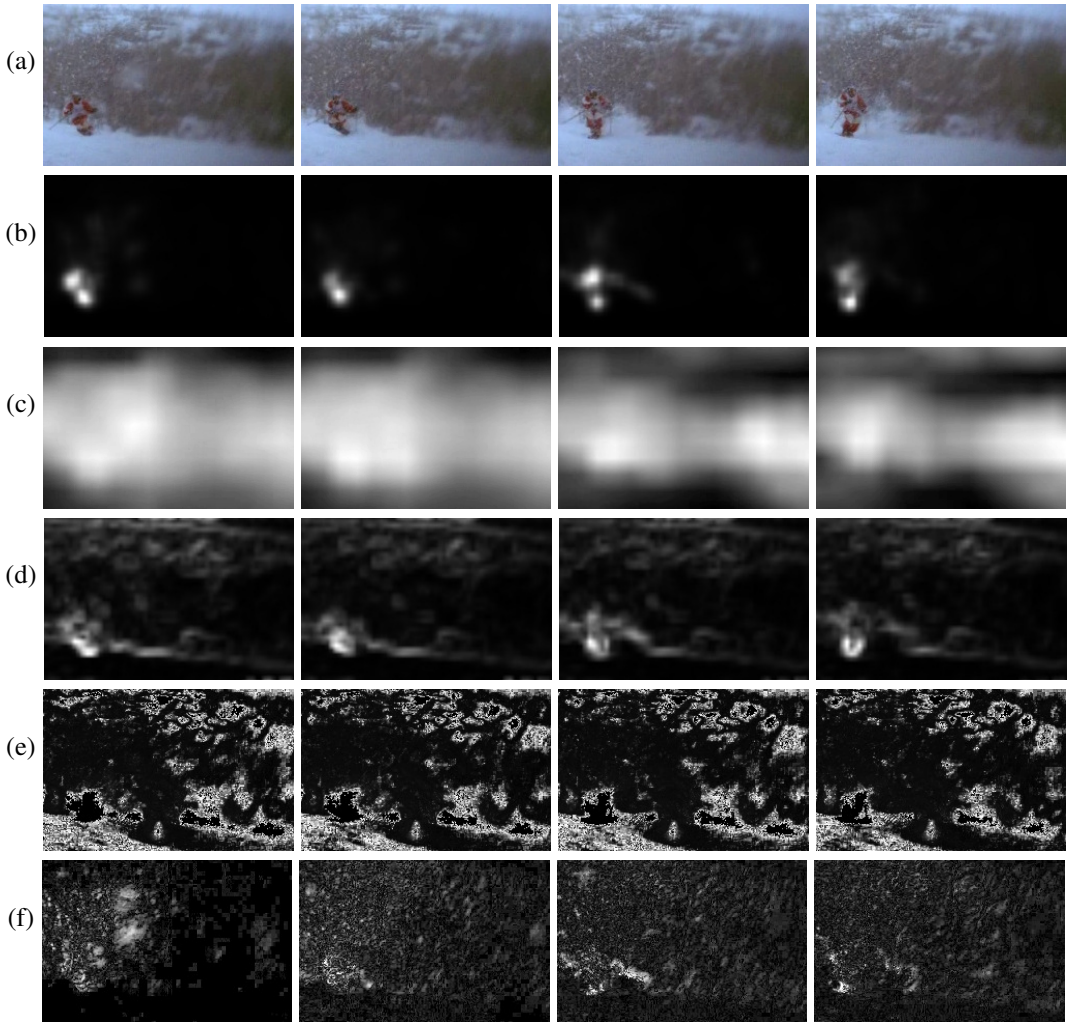


Figure 3.6: Results on skiing: (a) original (b) DiscSal (c) surprise (d) Monnet et al. (e) KDE (f) GMM.

able in [3]. The proposed algorithms clearly outperform all other methods, detecting the foreground motion and almost entirely ignoring the complex moving background. For all other methods, foreground detection is very noisy, and does not adapt well to the fast background dynamics. In result, the saliency maps contain substantial energy in background regions, sometimes missing the foreground objects completely.

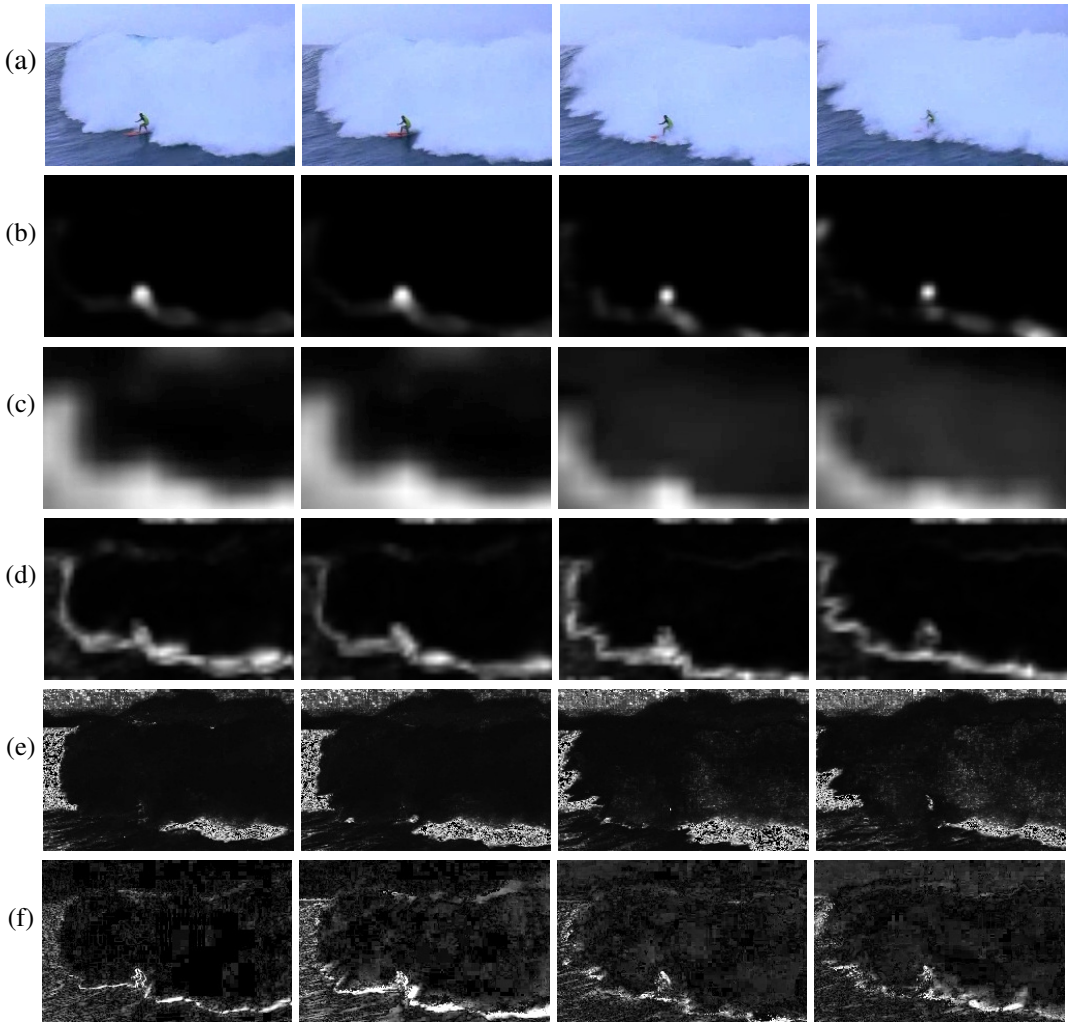


Figure 3.7: Results on surf: (a) original (b) DiscSal (c) surprise (d) Monnet et al. (e) KDE (f) GMM.

3.6.2 Quantitative analysis

To enable a quantitative analysis, all sequences were manually annotated with segmentation groundtruth for the objects of interest. The saliency maps were then thresholded at a large number of values, and the false alarm (α) and detection rate (β) computed. The resulting receiver operating characteristic (ROC) curves are shown in Figure 3.12. It is clear that the proposed algorithm achieves better performance than all others. The equal error rate (EER), defined as the error at which false alarm equals miss rate ($\alpha = 1 - \beta$), was also computed for all methods. Table 3.1 shows the EERs

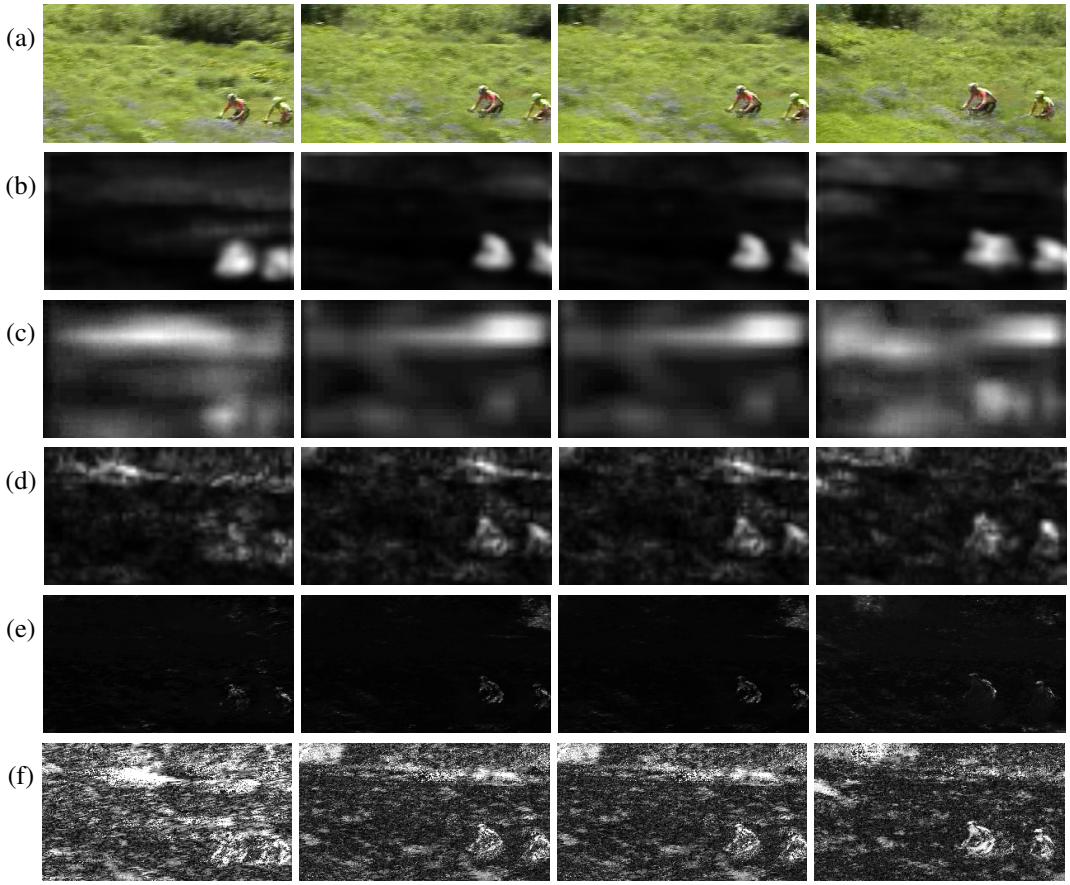


Figure 3.8: Results on cyclists: (a) original (b) DiscSal (c) surprise (d) Monnet et al. (e) KDE (f) GMM.

of the various methods (DiscSal, DiscSal-Causal, Surprise, KDE, Monnet, and GMM, referred to in the table as DS, DS(C), Su, KDE, Mo, and GMM, respectively) measured on all sequences, as well as the average over the sequence set. The proposed methods outperformed all others, achieving average EERs of 7.6% (DiscSal) and 9.3% (DiscSal-Causal), versus 16% for the closest competitor (the method of Monnet et al. [118]).

3.6.3 Sensitivity analysis

The size of the center window, n_c is the only free parameter with a significant impact on the performance of the proposed background subtraction algorithm. It determines the scale at which the saliency operation is computed. In all results shown in the previous section, it was set to $n_c = 16$, making the center 16×16 pixels spa-

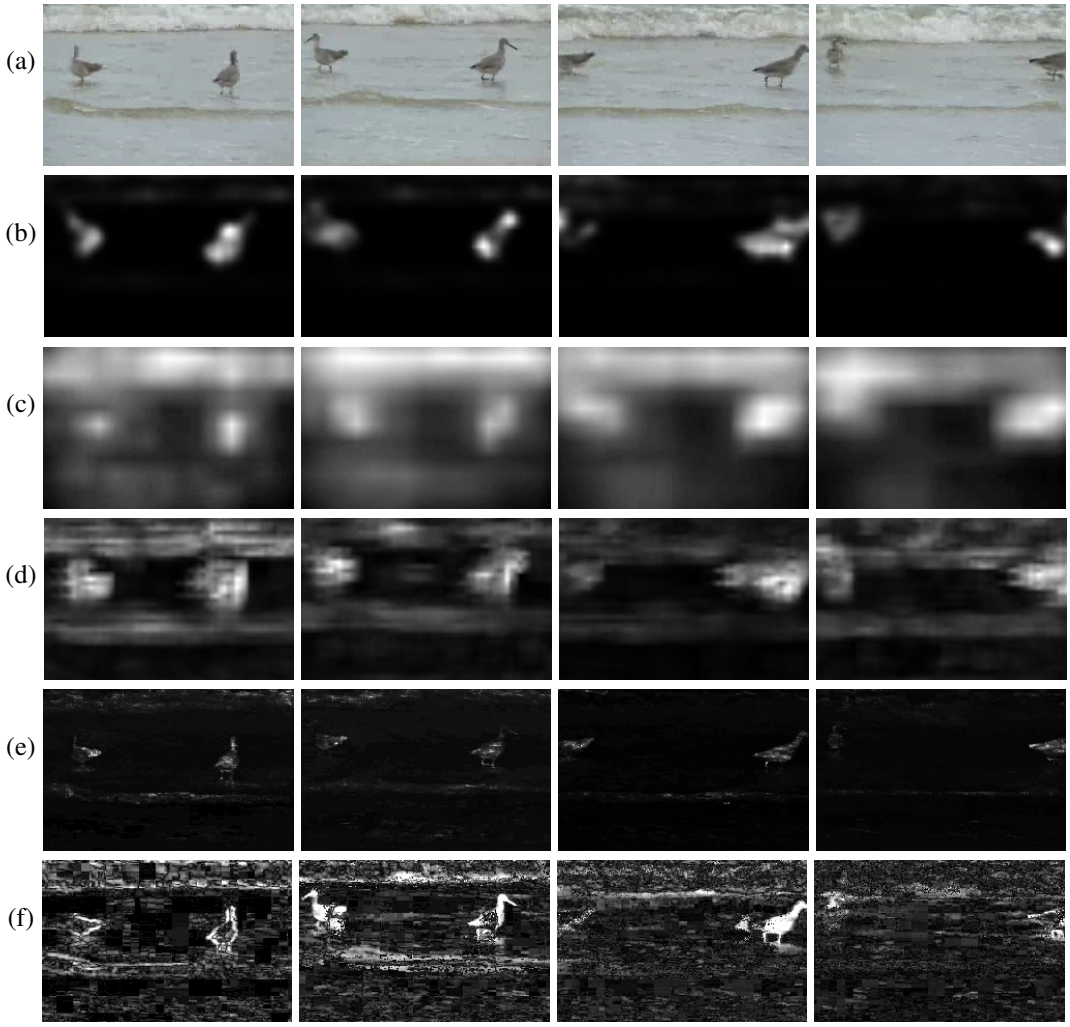


Figure 3.9: Results on birds: (a) original (b) DiscSal (c) surprise (d) Monnet et al. (e) KDE (f) GMM.

tially. To evaluate the impact of this parameter on saliency performance, we selected the 'birds' and 'cycle jump' sequences, varying n_c in the range $[8, 64]$. The sequences have 156×242 pixels, and the average size of the foreground object (across all frames in the sequence) is 30×40 pixels for 'birds' and 60×100 for 'cycle jump'.

A plot of EER as a function of scale is shown in Figure 3.13. The algorithm is fairly robust to variations in the scale parameter, achieving low error rates over a broad range of n_c values. Also, the lowest error rates are obtained at scales that match the size of the foreground object. This suggests that a standard multi-scale implementation can be used to automatically select the best value. Such schemes may also be used to

Figure 3.10: Results on helicopter: (a) original (b) DiscSal (c) surprise (d) Monnet et al. (e) KDE (f) GMM.

estimate the size of the salient object.

3.7 Discussion

In this work, we have proposed an algorithm for background subtraction based on center-surround saliency. The new algorithm is inspired by biological vision, namely the psychophysics of motion-based perceptual grouping, and extends a discriminant formulation of center-surround saliency previously proposed for static imagery [63, 65]. This extension is based on the representation of video with dynamic texture models,

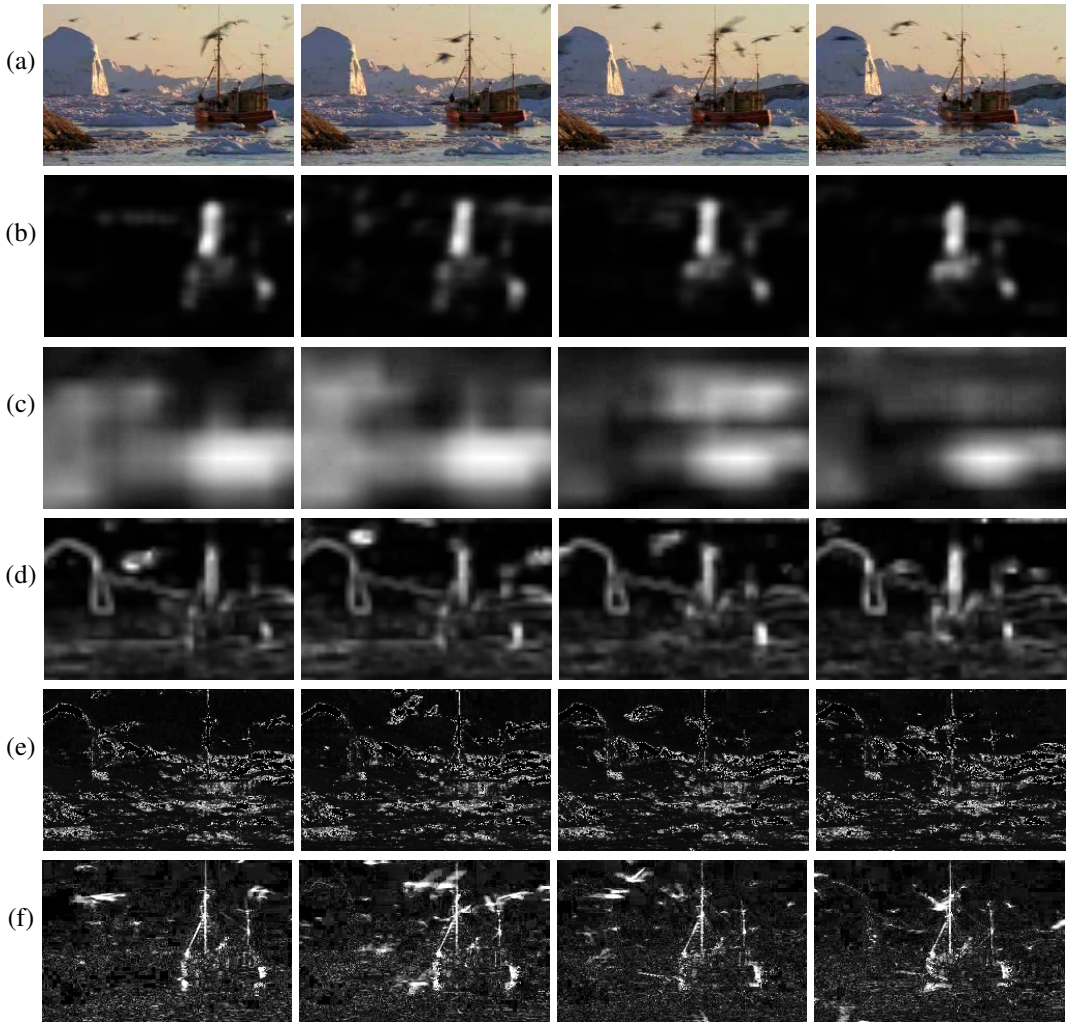


Figure 3.11: Results on flock: (a) original (b) DiscSal (c) surprise (d) Monnet et al. (e) KDE (f) GMM.

and is applicable to dynamic scenes. The algorithm combines spatial and temporal components of saliency in a principled manner, and is completely unsupervised. The combination of the discriminant center-surround saliency framework with the modeling power of dynamic textures leads to a robust and versatile background subtraction, which is successful even for scenes with highly dynamic backgrounds and a moving camera. Experimental evaluation on challenging sequences with complex backgrounds (snow, smoke, fire, water and flocks of birds) shows that the proposed algorithm substantially outperforms various state of the art background subtraction techniques. Quantitatively, the average error rates of the new algorithm are almost half that of the best competitor.

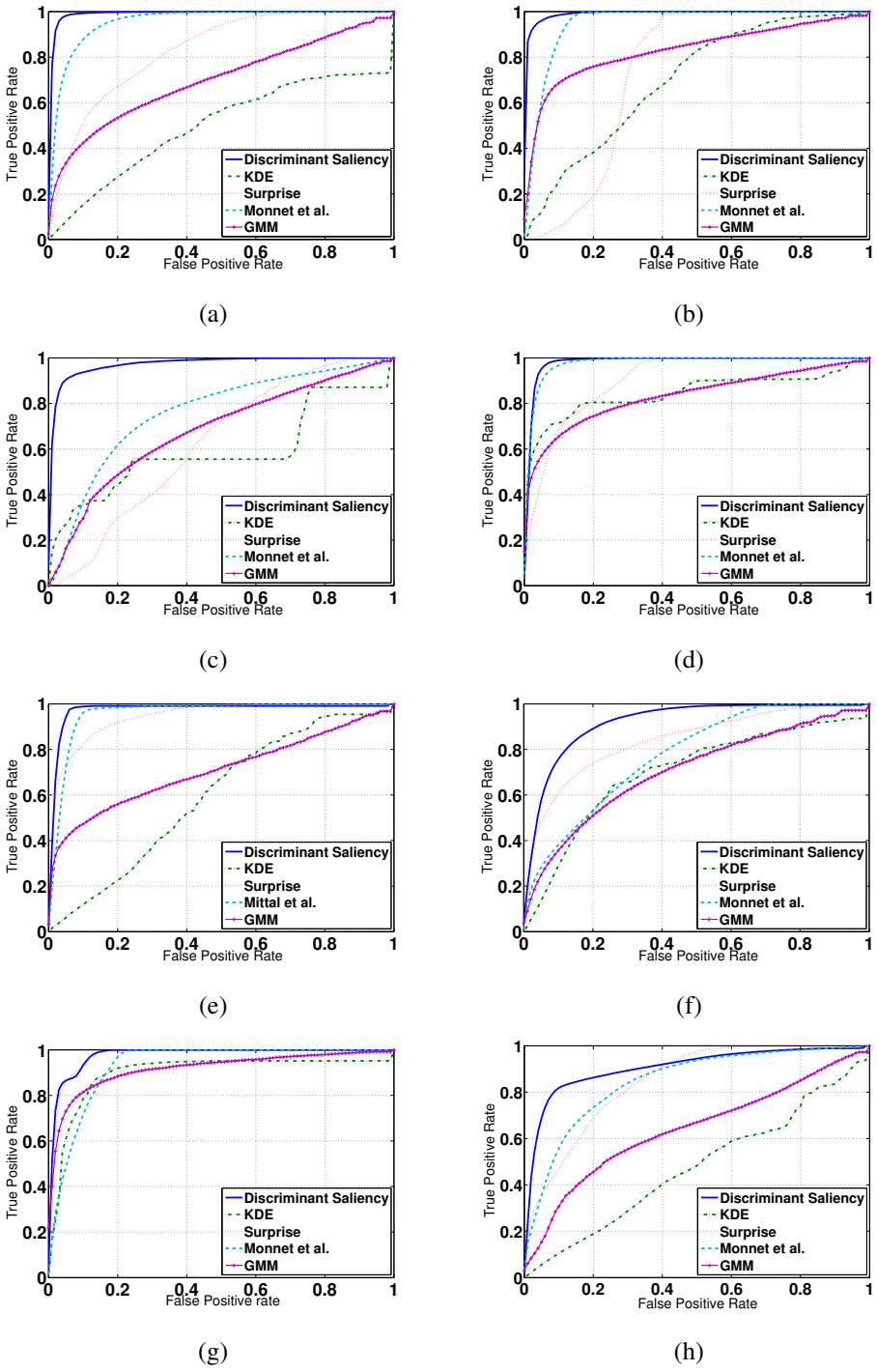
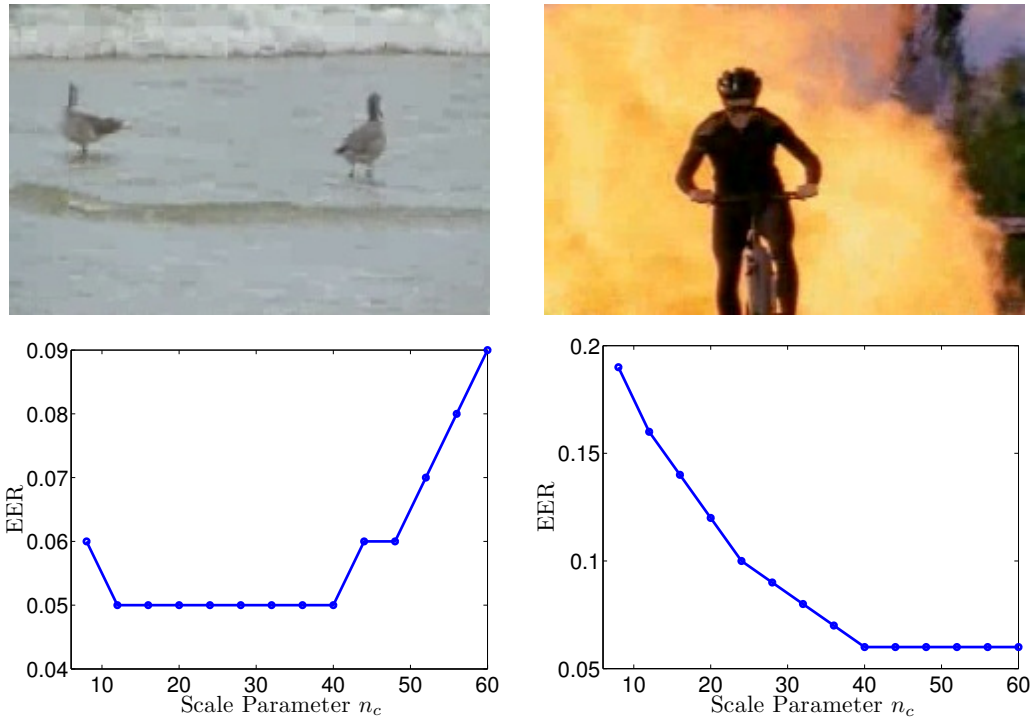


Figure 3.12: Performance of background subtraction algorithms on: (a) skiing (b) surf (c) cyclists (d) birds (e) helicopter (f) flock (g) boats (h) cycle jump

Table 3.1: Equal Error Rates for different saliency models. The average over all sequences is shown in the last row.

	DS	DS(C)	Su	KDE	Mo	GMM
skiing	**3%**	4%	26%	46%	11%	36%
surf	**4%**	5%	30%	36%	10%	23%
cyclists	**8%**	19%	41%	44%	28%	36%
birds	**5%**	9%	19%	20%	7%	23%
chopper	**5%**	**5%**	13%	43%	8%	35%
flock	**15%**	17%	23%	33%	31%	34%
boat	**9%**	11%	**9%**	13%	15%	15%
jump	**15%**	**15%**	25%	51%	23%	39%
surfers	**7%**	8%	24%	25%	10%	35%
bottle	**2%**	3%	5%	38%	17%	25%
hockey	**24%**	27%	28%	35%	29%	39%
land	**3%**	5%	31%	54%	16%	40%
zodiac	**1%**	**1%**	19%	20%	3%	40%
peds	**7%**	**7%**	37%	17%	11%	11%
traffic	**3%**	4%	46%	39%	9%	34%
freeway	**6%**	10%	43%	21%	31%	25%
ocean	**11%**	**11%**	42%	19%	**11%**	30%
rain	**3%**	6%	10%	23%	17%	15%
Avg	**7.6%**	9.3%	26.2%	33.1%	16%	29.7%

It is interesting to compare the performance of the different algorithms in light of their saliency representation. There are at least two significant differences between the previous methods and that now proposed. First, the GMM, KDE, and "surprise" models lack a sophisticated unified representation for the spatial and temporal components of saliency. For complex dynamic scenes where local variation in the background, either spatially or temporally, is significant, this leads to many false-positives. In this respect, the dynamic texture representation is a significant asset. Second, both the method of Monnet et al. and the GMM/KDE approaches, rely uniquely on models of the back-



(a) (b)

Figure 3.13: Effect of scale parameter n_c on EER for : (a) birds : mean EER = 5.86%, standard deviation = 1.29%; (b) cycle jump: mean EER = 9.36%, standard deviation = 4.31%; while the rates are low for all scales, a preference towards scales of the order of the object size is observable.

ground, treating foreground objects as outliers. Outlier detection can be a difficult problem, and is frequently more difficult than the problem of discriminating between two classes. Once again, this increased difficulty is exacerbated for highly dynamic scenes, where it is difficult to account for the large variability of background pixels with a single model. In this respect, the discriminant nature of the proposed saliency framework is a significant asset. Overall, both the discriminant formulation and the unified spatio-temporal representation seem to be necessary for good performance. This can be seen from the relative error rates of the various techniques, as shown in Table 3.1. The algorithm now proposed (DS) exhibits both properties and performs best. Methods that exhibit only one property ("surprise" discriminates between prior and posterior distributions, Monnet relies on a spatio-temporal representation similar to that of DS) achieve the next best levels of performance. Finally, methods that lack the two properties (GMM,

and KDE) perform quite poorly.

The proposed saliency detector can be applied to video processing tasks such as frame rate up-conversion [91] and in other computer vision tasks, such as tracking, activity recognition, and surveillance. The close parallels, previously shown to exist, between static discriminant saliency [65] and the neurophysiology of the early stages of the human visual system, also suggest that the proposed saliency detector is biologically plausible. We investigate this in the following chapter.

3.8 Acknowledgments

We thank Prof. Ahmed Elgammal for providing the code for non-parametric background subtraction using kernel density estimates from [51], Prof. Stan Sclaroff for providing some of the test sequences, and Antoni Chan, and Dashan Gao for useful discussions.

The text of Chapter 3, in full, is based on the material as it appears in: V. Mahadevan, and N. Vasconcelos, "Spatiotemporal saliency in dynamic scenes", IEEE Transactions on Pattern Analysis and Machine Intelligence, 32(1), pp. 171-177, 2010; V. Mahadevan, and N. Vasconcelos, "Background Subtraction in Highly Dynamic Scenes", IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 1-6, 2008. The dissertation author was a primary researcher and an author of the cited material.

Chapter 4

Biological Plausibility of Motion Saliency

4.1 Introduction

An important property of human saliency is its ubiquity: saliency mechanisms have been observed for various cues, including orientation, color, texture and motion [167, 123]. It has also been suggested that orientation and motion saliency could be encoded by similar mechanisms [126, 90]. In this Chapter we verify the hypothesis that the discriminant saliency detector for motion stimuli is biologically plausible by providing evidence of its ability to predict human psychophysics.

4.2 Biologically plausible motion saliency detector

We first derive discriminant saliency detectors for motion stimuli, using a biologically plausible approach to compute motion information from video sequences. For this we adopt the spatiotemporal filtering approach of [8], and [73] in place of the dynamic textures used in Chapter 3. Spatiotemporal filtering is a biologically plausible mechanism for motion estimation, and has been shown to comply with the physiology and psychophysics of the early stages of the visual cortex [8]. Since spatiotemporal orientation is equivalent to velocity, a set of 3-D Gabor (spatiotemporal) filters, tuned

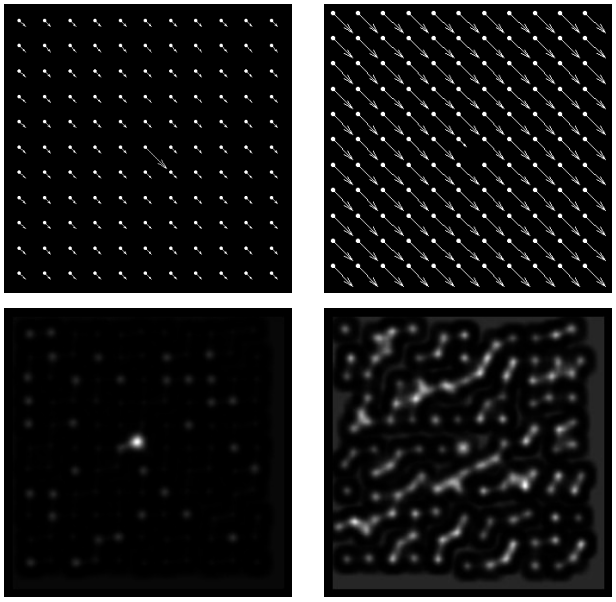
to a specific orientation in space and time, is used to extract the motion energy associated with different velocities. The algorithmic implementation of the spatiotemporal filters used in this work was based on the separable spatiotemporal filters of [73]. We considered only one spatial scale, and the spatial frequency of each Gabor filter was fixed to 0.25 cycles/pixel. Three temporal scales (temporal frequencies of 0, ±0.25 cycles/frame) and 4 spatial orientations ($0, \pi/4, \pi/2$ and $3\pi/4$) were used, in a total of 12 filters. The standard deviation of the spatial Gaussian was set to 1, and that of the temporal Gaussian to 2. This set of filter parameters were chosen for simplicity, we have not experimented thoroughly with them. We have also only considered the intensity of the input video frames, and all color information was discarded. These intensity maps were convolved with the 12 spatiotemporal filters, to produce the feature maps used by the saliency algorithm. Saliency was then computed as in the Chapter 2, using (2.1) and (2.3).

4.2.1 Consistency with psychophysics of motion perception

To evaluate the compliance of the discriminant saliency detector with the psychophysics of human motion saliency [126, 90], we start with some qualitative observations[1]. [90] showed that search asymmetries also hold for moving stimuli. For example, searching for a fast-moving target among slowly-moving distractors is easier than the reverse. We applied the motion-based discriminant saliency detector to a set of sequences used to demonstrate the asymmetries of motion pop-out [90], with the results illustrated in Figure 4.1. The figure presents quiver plots of the motion stimuli, under the two conditions, and one frame of the resulting discriminant saliency map. The conspicuous saliency peak at the target in Figure 4.1 (a) shows a strong pop-out effect when the target speed is greater than that of the distractors. No noticeable pop-out effect is observed in Figure 4.1 (b), where the distractor speed is greater than that of the target. This shows that the discriminant saliency detector can replicate the asymmetries of motion saliency.

As was the case for static stimuli in the work of [65], we complemented this qualitative observation with a quantitative analysis of the saliency predictions made by the discriminant detector. [126] found that human saliency responses to motion are very

[1] All motion stimuli sequences in the experiments were generated using the Psychtoolbox [24].



(a) (b)

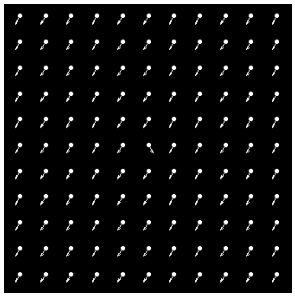
Figure 4.1: Discriminant saliency detector output for (a) a fast-moving target among slowly-moving distractors, and (b) a slowly-moving target among fast-moving distractors. Top row shows quiver plots of the stimuli (the direction of motion is specified by the arrow whose length indicates the speed), and bottom row the corresponding saliency maps.

similar to those observed for orientation: the perception of saliency of moving targets increases nonlinearly with motion contrast, and shows significant saturation and threshold effects. To test the compliance of discriminant saliency with this nonlinearity, we applied it to the motion displays of [126]. An example is shown in plot (a) of Figure 4.2, where (b) shows a plot of the human saliency data, reproduced from the original figure of [126], and (c) presents the predictions made by discriminant saliency. The two plots are very similar, both exhibiting threshold and saturation effects.

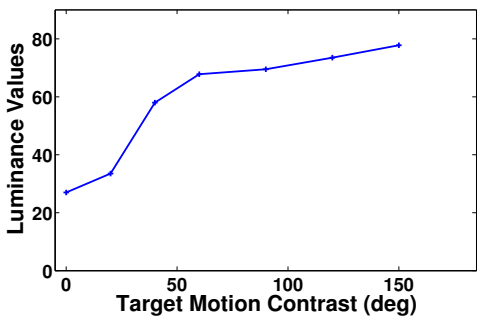
4.3 Acknowledgments

We thank Prof. David Heeger for providing the code for motion energy computation using spatiotemporal filters [73].

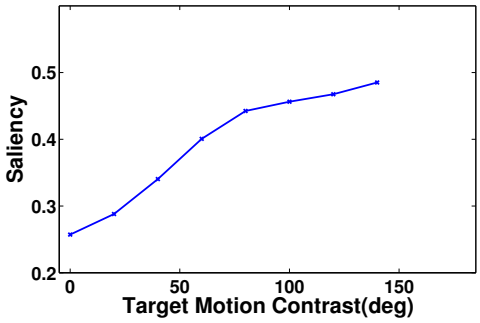
The text of Chapter 4, in full, is based on the material as it appears in: D. Gao, V. Mahadevan, and N. Vasconcelos, “On the plausibility of the discriminant center-



(a)



(b)



(c)

Figure 4.2: The nonlinearity of human saliency responses to motion contrast (reproduced from Figure 9 of Nothdurft, 1993) (b) is replicated by discriminant saliency (c). A quiver plot of one instance of the motion display used in the experiment (with background contrast (bg)=0, target contrast (tg)=60) is illustrated in (a). The direction of motion is specified by the arrow, whose length indicates the speed.

surround hypothesis for visual saliency", Journal of Vision, 8(7):13, 1-18, June 2008. The dissertation author was a primary researcher and an author of the cited material.

Chapter 5

Tracking

5.1 Introduction

In the biological world, object tracking is closely related to the task of fixating objects of interest. The goal is to keep an object on the fovea of the observer, even when either or both are moving. This is achieved with a combination of overt and covert eye movements, and underlies the mechanisms for identification of moving objects [133]. Due to the evolutionary advantage of solving these tasks accurately, it is not surprising that biological vision has evolved extremely efficient tracking mechanisms, in terms of accuracy, robustness, and speed. It has been hypothesized that the effectiveness of these mechanisms, even under the most adverse conditions, involving clutter, low-light etc., is a consequence of the availability of robust saliency mechanisms, that cause pre-attentive pop-out of certain locations of the visual field [133]. These salient locations become the *focus of attention* (FoA) for the post-attentive stages of visual processing, where top-down feedback from higher level cortical layers is used to solve problems such as object recognition or visual search [183].

In this Chapter, we make a connection between tracking and saliency, by postulating that *tracking is simply a manifestation of the continuous computation of saliency over time*. More precisely, we frame tracking as a byproduct of the center-surround saliency mechanisms that are prevalent in biological vision [65, 33]. This is done with recourse to the computational of visual saliency, denoted *discriminant saliency* formulation discussed in Chapter 2.

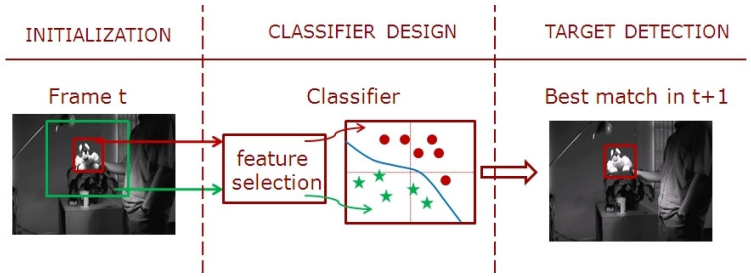


Figure 5.1: Overview of discriminant tracking. Tracking iterates between two main steps: classifier design and target detection.

Object tracking is a classical problem in computer vision, and a pre-requisite for many of its important applications, such as surveillance, activity or behavior recognition and video retrieval. Decades of research on this topic have produced a diverse set of approaches and a rich collection of tracking algorithms [190]. Many of the these are based on *appearance modeling*. They learn (and maintain) a model of target appearance, which is used to locate the target as time evolves [83, 41, 17, 92]. The main limitation of these methods is that they rely uniquely on models of object appearance, and do not take the background into account. This limits tracking accuracy when backgrounds are cluttered, or targets have substantial amounts of geometric deformation, such as out-of-plane rotation. To address this limitation, various authors have noted that it is frequently easier to model the differences between target and background than to model the target itself. This has led to the idea of *discriminant tracking*, where the tracking problem is framed as one of continuous object detection, through incremental *target vs. background* classification [39, 12, 66]. Discriminant tracking has two main steps, which are illustrated in Figure 5.1. Given an initial target bounding box, say at time t, the first step consists of *classifier design*: a classifier is trained by selecting visual features that discriminate between target and background, and a decision rule is learned based on these features. In the second step, denoted *target detection*, the classifier is applied to every location of the visual field, so as to determine the most likely location of the target at time $t+1$. The target bounding box is moved to this location, and the process iterated.

This generic formulation has been used to design various trackers [39, 12, 66, 67, 13]. Although these efforts have demonstrated that discriminant tracking can achieve state-of-the-art performance in computer vision [66], this performance is still far from

that of the tracking mechanisms implemented by biological vision.

In this Chapter, we start by showing that *optimal* (in a decision theoretic sense) discriminant tracking can be implemented with a combination of operations that are well documented in the biological attention literature: *center-surround saliency* [88], a spatial *spotlight of attention* [137], and *feature-based attention* [168]. It is then shown that, under the discriminant saliency formulation, these operations are mapped into statistical operations such as *feature selection* or *target detection*. This enables the derivation of *optimal trackers* that can be implemented with *simple* and *highly efficient* computations, two important requirements for the practical feasibility of any tracker. The saliency formulation is next shown to also establish a *unified framework for classifier design, target detection, automatic tracker initialization, and scale adaptation*. While the steps of classifier design and target detection are addressed by all discriminant trackers in the literature, previous solutions cannot cope with the initialization and scale adaptation problems. Finally, it is shown that the proposed discriminant tracker outperforms a number of state-of-the-art tracking approaches in the literature.

5.2 Related work on object tracking

Many popular approaches to object tracking are based on *appearance modeling*. They learn and maintain a model of target appearance, which is used to locate the target as time evolves. Conditional density propagation [83] is one of the most popular methods in this class. Targets are represented by some type of visual features, e.g. their contours or deformable templates [194], and the temporal evolution of these features is modeled with a particle filter. Alternatively, target appearance is frequently represented by kernel weighted color histograms, which are combined with the mean shift procedure to identify the most likely position of the target in the next frame [41]. Representations of the target and/or background with probabilistic models, e.g. a mixture of Gaussian (MoG) models, have also been proposed [159, 68]. Equally popular are subspace methods, which maintain a low-dimensional representation of target appearance [17, 77]. Recently, there has been an interest in making these representations adaptive, by updating subspaces incrementally, using online principal component analysis [145]. More so-

phisticated appearance models include a combination of short term descriptors and long term stable representations [92], specialized representations tailored to specific entities, such as people [141], or multiple image patch representations, such as "FragTrack" [7].

Appearance based trackers have limited accuracy when backgrounds are cluttered, or targets have substantial amounts of geometric deformation, such as out-of-plane rotation. *Discriminant trackers* frequently achieve better performance in these scenarios [66] by framing tracking as incremental *target vs. background* classification [39, 12]. Collins et al. [39] rely on a feature set composed of histograms of filter responses to the R,G,B channels of the visual stimulus. Discriminant features are selected with a variant of the Fisher discriminant, and the classifier is implemented with a likelihood-based decision rule. Fisher discriminants are also used to classify foreground from background in [107] and [121]. The "ensemble tracking" method of Avidan [12] uses a combination of histograms of oriented gradients [46] and R,G,B pixel values as features. A set ("ensemble") of weak hyperplane classifiers are trained to separate target from background, and combined into a decision rule, using AdaBoost [58]. Grabner et al. [66] have proposed an alternative ensemble tracker, based on online boosting. This maintains a set of weak learners that are updated at every time step. More recently, online boosting has been combined with a semi-supervised update of the weak learners to increase tracker robustness [67]. A multiple instance learning (MIL) based approach has also been proposed in [13], to minimize the ensemble tracker sensitivity to outliers due to misalignment of the target bounding box.

The robustness of biological tracking mechanisms has inspired computer vision researchers to augment conventional trackers with FoA mechanisms. For instance, Toyama and Hager [164] proposed an incremental FoA procedure to combine multiple trackers, leading to increased robustness. Nevertheless, there has been little work aimed at deriving a principled understanding of what computational mechanisms could be used by biological vision to solve the tracking problem, how these mechanisms relate to the state-of-the-art algorithms from computer vision, and how these connections could be exploited to achieve increased computer vision performance. In this work, we present a formulation of tracking that addresses these questions.

5.3 Discriminant tracking

The central hypothesis of this work is that discriminant tracking can be implemented with a combination of bottom-up and top-down saliency detection. In this section, we build on this hypothesis to propose a saliency-based discriminant tracker.

5.3.1 The connection to saliency

We start by relating discriminant tracking to saliency. Given an initial target location at time t, l^*, the first step of discriminant tracking is to design a target/background classifier. The target and background hypotheses are defined by the feature responses in a window centered at l^*, the *target window,* and a surrounding annular *background window*. Hence, like bottom-up saliency, discriminant tracking requires the computation of the discriminant power of each feature in $\boldsymbol{Y}$ with respect to a *center-surround discrimination* problem. The main difference is that, while bottom-up saliency performs the computation at *each* location of the visual field, discriminant tracking only requires it at location l^*. This is equivalent to computing bottom-up saliency after application of a *spatial focus of attention* mechanism tuned to the target location. Given a measure of how discriminant each feature is for target/background discrimination at time t, the next step is to find the target in the next frame, i.e. at time $t + 1$. This is formulated as a target detection problem. It requires the selection of the most discriminant features in $\boldsymbol{Y}$, and a decision rule for target detection. Since the discriminant power of each feature is already known, feature selection reduces to suppression of non-discriminant features and enhancement of discriminant ones. This type of manipulation is exactly the function of a *feature-based attention* mechanism. Finally, target detection can be implemented with a top-down saliency measure trained from the feature responses in the target and background windows at time t. The position of the target at time $t + 1$ is determined by a search for the location of largest saliency within a neighborhood of the target position at time t. This restriction of the search space reduces the computation needed to identify the new target location, by ignoring regions peripheral to the current focus of attention. It is consistent with the "center bias" observed in the human visual system, where a saccade to a new fixation location is biased to be close to the current

center of view [161, 169].

The overall process is illustrated in Figure 5.2. The display in a) shows two disks (one red, one brown) moving against a background of green distractors. Assume that the red disk is the target and that the feature set $\mathcal{Y}$ consists of a number of color detectors. At time t, the spatial focus of attention mechanism narrows the field of view to the neighborhood of the target, as shown in b). This makes the target salient. Computation of center-surround saliency, as in c) finds the red color to be the most discriminant feature. Training a top-down saliency measure for target/background classification in this area produces a detector of red disks. For simplicity, we assume that this a threshold on the red channel of the visual stimulus. Target detection at time $t + 1$ starts with the application of feature-based attention, which strengthens the red channel and inhibits all others. This is illustrated in d) and e) where we present the display at time $t + 1$, and its projection on the selected feature, i.e. its red color channel. Note how the feature-based manipulation of attention eliminates much of the clutter in the scene. In fact, only the red disk elicits a strong response after feature selection. Further application of the top-down saliency detector (red threshold classifier), followed by a search for maximum saliency within a neighborhood of the previous target location, leads to the identification of the red disk, as shown in f).

5.3.2 The core tracking procedure

The discussion of the previous section suggests that discriminant tracking can be implemented with discriminant saliency measures. Starting with the target location l^* at time t, and the associated target ($\mathcal{W}^1_{l^*}$) and background ($\mathcal{W}^0_{l^*}$) windows, the tracker is implemented as follows.

- **Learning:** at time t, estimate the probability distributions $p_{\mathbf{Y}|C(l)}(\mathbf{y}|i), i \in \{0, 1\}$ using the feature responses in $\mathcal{W}^i_{l^*}$, as training sample, and the distribution $p_{\mathbf{Y}}(\mathbf{y})$ from the responses in $\mathcal{W}_{l^*} = \mathcal{W}^0_{l^*} \cup \mathcal{W}^1_{l^*}$.

- **Feature selection:** Among the N available features, select the subset of $K < N$ that maximizes the saliency measure of (2.3).

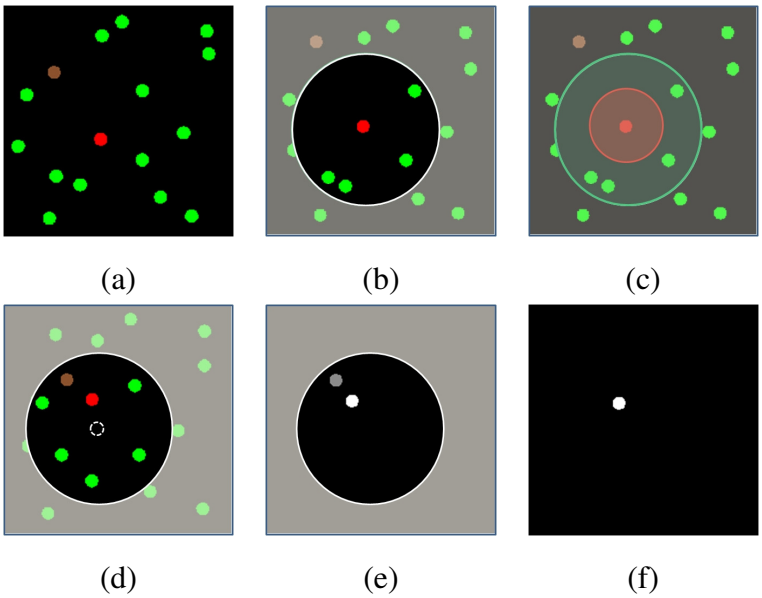


Figure 5.2: Illustration of saliency-based tracking. (a) two disks, one red and one brown are salient amongst green distractors; (b) defining the red disk as the target, at time t, focuses spatial attention on it; (c) computing center surround saliency at this location leads to the selection of the feature "red" as the most salient; (d) the position of the disks at time $t + 1$, shown with the focus of attention from time t; (e) feature based attention suppresses all but the red feature channel, which has non-zero response only at the locations of the red and brown disks; (f) the location of the target has the largest saliency inside the focus of attention.

- **Classification:** using these K features compute, at time $t+1$, the top-down saliency of each location l of the visual field, using the saliency measure of (2.6). Move the target/background windows to the location of largest saliency within a neighborhood of l^*, and iterate the process.

While optimal in theory, this implementation has a number of practical limitations. First, the saliency measures of (2.3) and (2.6) require the evaluation of the joint probability distribution of the features in $\boldsymbol{Y}$. This is too complex for most applications of saliency and infeasible for tracking, where there is a premium on computational efficiency. Various simplifications can be achieved by restricting the features to bandpass filters, and exploiting the statistical regularities of the responses of such features to natural images. However, a classifier built from bandpass features may not have the robustness necessary to track complex objects subject to non-planar motion. This type of robustness usually requires more abstract features. Finally, the classifier should operate across multiple scales, so as to enable scale adaptation as the distance between objects and camera varies. These issues are addressed in the remainder of this section.

5.3.3 Salient Feature Selection

Feature selection is naturally implemented under discriminant saliency, since the saliency measure is itself a measure of discrimination. In fact, extremely efficient implementations are possible when the features belong to the class of bandpass filters. Assuming this to be the case, let the feature space $\mathcal{Y}$ have dimension N, and denote $\boldsymbol{Y} = (Y_1, \ldots, Y_N)$. *Salient feature selection* involves the identification of the subset of $K < N$ features that maximizes discrimination between target and background. One possibility to accomplish this is to define $\boldsymbol{Y}_{1,k} = (Y_1, \ldots, Y_k)$, and expand the mutual information of (2.1) into [170]:

$$I(\boldsymbol{Y}; C) = \sum_k I(Y_k; C) + \sum_k [I(Y_k; \boldsymbol{Y}_{1,k-1}|C) - I(Y_k; \boldsymbol{Y}_{1,k-1})] \tag{5.1}$$

where

$$I(\boldsymbol{Y}; C|\mathbf{Z}) = \sum_i \int P_{\boldsymbol{Y},C,\mathbf{Z}}(\mathbf{y}, i, \mathbf{z}) \log \frac{P_{\boldsymbol{Y},C|\mathbf{Z}}(\mathbf{y}, i|\mathbf{z})}{p_{\boldsymbol{Y}|\mathbf{Z}}(\mathbf{y}|\mathbf{z}) p_{C|\mathbf{Z}}(i|\mathbf{z})} d\mathbf{y} d\mathbf{z} \tag{5.2}$$

is the conditional mutual information between $\boldsymbol{Y}$ and C given the observation of $\mathbf{Z}$. In (5.1), the term $I(Y_k; C)$ is the marginal mutual information (MMI) between the k^{th} feature and the class label. It measures how discriminant the k^{th} feature is individually. The terms $I(Y_k; \boldsymbol{Y}_{1,k-1}|C) - I(Y_k; \boldsymbol{Y}_{1,k-1})$ quantify the discriminant information contained in feature dependencies between the k^{th} feature and the set of $k-1$ previously selected features [170]. This decomposition allows a substantial simplification of the mutual information, by exploiting a well known property of band-pass features extracted from natural images: that such features exhibit *consistent* patterns of dependence across an extremely wide range of natural image classes [28, 80]. This implies that the dependencies between features carry little information about the class from which the features are extracted, allowing the approximation of (5.1) by

$$I(\boldsymbol{Y}; C) \approx \sum_{k=1}^{N} I(Y_k; C). \tag{5.3}$$

As noted in Section 2, the mutual information $I(Y_k, C)$ measures the extent to which feature Y_k discriminates between target and background classes. However, a large mutual information does not imply that the feature is characteristic of the target. In fact, a feature that is totally absent from the target but prevalent in the background is highly discriminant for target/background classification. In the tracking context, it is usually undesirable to rely on such features, since the background can vary drastically as the target, the camera, or both, move. For example, the target can move from an area of the scene where the background is highly textured (e.g. vegetation) to an area where is has virtually no texture (e.g. a white wall). A tracker that relies on features characteristic of the background texture to detect the target can lose the latter as it moves into the textureless regions of the scene. Hence, features that are discriminant but absent from the target can lead to unstable tracking, and should be discarded. For bandpass features, whose responses to natural images have zero mean and probability density functions that decay with the distance to the origin, the detection of features that are expressed in the target is fairly straightforward. It suffices to select features that have larger variance under the target class than under the background class. Since the feature responses have zero mean, this can be written as

$$E_{Y_k|C}[y_k^2|1] > E_{Y_k}[y_k^2|0]. \tag{5.4}$$

This condition can be combined with (5.3) to obtain a very efficient salient feature selection mechanism. Since the mutual information is always non-negative, the selection of the optimal subset of K ($K < N$) salient features reduces to 1) ordering the N features by decreasing MMI, $I(Y_k, C)$, 2) discarding features that do not satisfy the variance condition of (5.4) and 3) selecting the first K. This is denoted feature selection by maximum marginal diversity in [171].

5.3.4 Efficient computation of saliency measures

In addition to efficient feature selection, the combination of (5.3) and the statistics of bandpass responses to natural images also simplifies the discriminant saliency measures. This follows from the well known observation that the probability distribution of feature responses of a bandpass feature, to natural images, follows a generalized Gaussian distribution (GGD) [80]

$$P_Y(y; \alpha, \beta) = \frac{\beta}{2\alpha\Gamma(1/\beta)} \exp\left\{-\left(\frac{|y|}{\alpha}\right)^{\beta}\right\}, \tag{5.5}$$

where $\Gamma(z) = \int_0^\infty e^{-t}t^{z-1}\mathrm{d}t$, $t > 0$, is the Gamma function, α a *scale* parameter, and β a *shape* parameter. The β parameter controls the rate of decay of the GGD, from the peak value (e.g. Laplacian when $\beta = 1$ or Gaussian when $\beta = 2$). It has been shown that $\beta \in (0.5, 0.8)$ provides a good fit to large corpora of natural images [157]. We found $\beta = 0.7$ to work best and we adopt this parameter value throughout this work.

Given β, the only parameter that remains to be learned is the scale α. This can be done by the method of moments [153], which exploits the fact that the scale $\alpha_{k,i}$ of the response of feature Y_k under class $C = i$ can be derived from

$$\alpha_{k,i} = \sqrt{\frac{\sigma_{k,i}^2 \Gamma(\frac{1}{\beta})}{\Gamma(\frac{3}{\beta})}} \tag{5.6}$$

with

$$\sigma_{k,i}^2 = E_{Y_k|C}[y_k^2|i] \approx \frac{1}{n} \sum_{j|y_k^j \in \mathcal{D}_i} \left(y_k^j\right)^2, \tag{5.7}$$

where $\sigma_{k,i}^2$ is the variance of Y_k under the class $C = i$, $\mathcal{D}_i = \{y_k^1, \ldots, y_k^n\}$ a training sample from this class, and we have used the fact that the responses of bandpass filters have zero

mean. In summary, given a sample of feature responses from the target and background windows, the estimation of the scale parameters is trivial.

We next note that, for bottom-up saliency, the approximation of (5.3) reduces (2.3) to

$$S(l) \;=\; \sum_k I_l(Y_k; C) \tag{5.8}$$

where, $I_l(Y_k; C) = \sum_{i=0}^{1} p_{C(l)}(i) KL[p_{Y_k|C(l)}(y_k|i) \| p_{Y_k}(y_k)]$. Combining this with the KL divergence between two GGDs [47]

$$\mathrm{KL}[P_Y(y; \alpha_i, \beta) \| P_Y(y; \alpha, \beta)] = \log\left(\frac{\alpha}{\alpha_i}\right) + \frac{1}{\beta}\left[\left(\frac{\alpha_i}{\alpha}\right)^{\beta} - 1\right], \tag{5.9}$$

leads to the *simplified bottom-up saliency measure*

$$S(l) \;=\; \sum_k \sum_{i=0}^{1} \pi_i \left(\log\left(\frac{\alpha_k}{\alpha_{k,i}}\right) + \frac{1}{\beta}\left[\left(\frac{\alpha_{k,i}}{\alpha_k}\right)^{\beta} - 1\right]\right), \tag{5.10}$$

where $\pi_i = P_C(i)$ is the prior for class i, and $\alpha_k, \alpha_{k,i}$ the scale parameters of $p_{Y_k}(y_k), p_{Y_k|C(l)}(y_k|i)$.

With respect to top down saliency, (2.6) reduces to

$$S(l) = \sum_k S_k(l), \qquad S_k(l) = \begin{cases} I(C; Y_k = y_k(l)) & \text{if } l \in \mathbf{S}_k \\ 0, & \text{otherwise.} \end{cases} \tag{5.11}$$

$$\mathbf{S}_k = \left\{ l \,\middle|\, \frac{P_{C,Y_k}(1, y_k(l))}{P_C(1) P_{Y_k}(y_k(l))} > \frac{P_{C,Y_k}(0, y_k(l))}{P_C(0) P_{Y_k}(y_k(l))} \right\}. \tag{5.12}$$

We next note that [65], when $p_{Y_k|C}(y_k|i), i \in \{0, 1\}$ are GGDs with scale parameters $\alpha_{k,i}$,

$$I(C; Y_k = y_k) = s[g_k(y_k)] \log \frac{s[g_k(y_k)]}{\pi_1} + s[-g_k(y_k)] \log \frac{s[-g_k(y_k)]}{\pi_0}, \tag{5.13}$$

with $s(y) = (1 + e^{-y})^{-1}$ a sigmoid function, $\pi_i = P_C(i)$, and

$$g_k(y) = \xi_k |y|^{\beta} - T_k, \quad \xi_k = \frac{1}{\alpha_{k,0}^{\beta}} - \frac{1}{\alpha_{k,1}^{\beta}}, \quad T_k = \log \frac{\alpha_{k,1} \pi_0}{\alpha_{k,0} \pi_1}. \tag{5.14}$$

Furthermore, it follows from (5.14) and (5.6) that the variance condition of (5.4) is equivalent to $\xi_k > 0$. Under this condition, the sets $\mathbf{S}_k$ can be simplified into

$$\mathbf{S}_k = \{ l \,|\, |y_k(l)| > t_k \} \quad \text{with} \quad t_k = \left(\frac{1}{\xi_k} \log \frac{\alpha_{k,1}}{\alpha_{k,0}}\right)^{\frac{1}{\beta}} \tag{5.15}$$

Using this in (8.1) leads to the *simplified top-down saliency measure*

$$S_k(l) = \begin{cases} \sum_{i=0}^{1} h_i[\xi_k |y_k(l)|^\beta - T_k] & \text{if } |y_k(l)| > t_k \\ 0, & \text{otherwise,} \end{cases} \tag{5.16}$$

with $h_i(x) = s\{(-1)^{1-i}x\} \log\left\{\frac{1}{\pi_i} s\{(-1)^{1-i}x\}\right\}$. The form of (5.16) suggests the interpretation of salient features as matched filters for the detection of visual attributes of the target class. This is due to the constraint $|y_k(l)| > t_k$, which only assigns saliency to the regions where the k^{th} feature response has large magnitude. These are regions where the visual stimulus resembles the feature.

In summary, for bandpass features, both salient feature selection and saliency detection are quite simple. Given a sample of responses from feature Y_k in the target and background windows, the parameters $\alpha_{k,i}$, ξ_k, T_k, and t_k are estimated with (5.6), (5.14), and (5.15). Features Y_k for which $\xi_k \leq 0$ are then discarded. The remaining are ordered by decreasing mutual information $I(Y_k, C)$, using (5.8) and (5.9), and the top K selected. Saliency detection is then performed with these features, using (8.1) and (5.16). The simplicity of all these operations is crucial for discriminant tracking, where they have to be repeated at each time step.

5.3.5 Spatial Importance Maps

The implementation of a discriminant tracker requires trade-offs between detector robustness, computational complexity, and adaptivity. Typically, robustness requires decision functions with many features, and learned from a large number of examples. Such functions are difficult to learn and adapt. Adaptation is particularly challenging, since both the feature subset added at a given time step, and the examples from which it is learned tend to be overwhelmed by those of the previous steps. The more robust a classifier becomes, the more difficult it is to adapt to variations in the statistics of the two classes. However, adaptation is crucial for tracking, where the difficulty is exactly to track objects as they *change appearance*, due to variations in lighting, pose, background, etc. The saliency-based discriminant tracker of the previous section is highly adaptive, since learning is reinitialized at each frame. The price is that, due to limited training data and computation available, it can only use a small number of simple

features. Hence, as an object detector, it is not very robust.

One of its major limitations is that no positional information is stored for the filter responses. As a result, the saliency assessments of (5.16) do not require spatial consistency of feature responses. For example, it is indifferent if a feature only has large response in the top or bottom half of the target window. Since salient features are usually not expressed in the entire target window, this can lead to noisy saliency maps for target detection. An obvious improvement is to define a feature for each combination of bandpass filter *and* location within the target window, as is popular in face detection [175]. This is, however, infeasible for tracking, due to the extensive amounts of computation and training data required. A better alternative is to learn a second layer of features, that model *configurations of feature responses*. This is inspired by recent work in HMAX networks [151]. These are biologically inspired object recognition networks, composed of two layers. The first layer can be seen as a (weak) object detector, based on simple bandpass features (Gabor functions) such as those used in this work. The second is an equivalent classifier, but uses more complex features. These are obtained by randomly sampling the responses of the first layer to objects in the target class, and can be interpreted as representative templates of first layer response. In fact, the first layer of the HMAX network can be expanded to perform top-down saliency detection [69], in which case the second layer filters are *saliency templates*. These summarize the saliency configurations that appear during training, providing a rough characterization of object shape. In this way, the addition of the second HMAX layer increases the robustness of the saliency detector implemented by the first [69].

While the training complexity of a full HMAX network is too large for tracking, the idea of accounting for positional information through the inclusion of saliency templates can still be used. In fact, there is a very natural template to use at time step $t + 1$: the map of saliency responses, within the target window $\mathcal{W}^1_{l^*}$, of each salient feature at time t. This is denoted *the spatial importance map*, and computed as

$$\mathcal{T}_k(l) = \frac{< S_k(l) >_t}{\sum_{l \in \mathcal{W}^1_{l^*}} < S_k(l) >_t}, \quad l \in \mathcal{W}^1_{l^*} \tag{5.17}$$

where $< S_k(l) >_t$ is a local average over 4×4 pixels of the k^{th} saliency response at time t. The proposed normalization guarantees that $\mathcal{T}_k(l)$ sums to 1, giving it the interpretation of a weighting function that emphasizes regions of strong feature response. Since

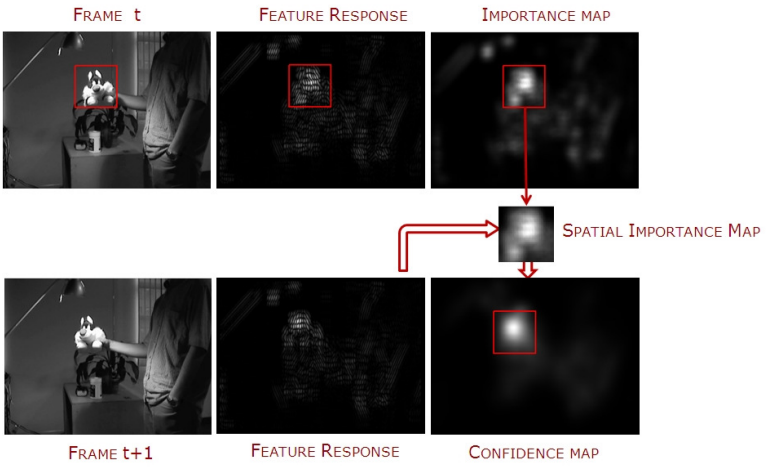


Figure 5.3: Spatial importance maps. For each selected feature, a saliency template is stored at time t. At time $t+1$, the saliency of (5.16) is correlated with this template, to enforce spatial consistency of the saliency detection over time.

1) salient features are discriminant for target/background classification, and 2) bandpass features respond to image landmarks, such as edges, corners or texture, these are regions of landmarks that distinguish target from background. In summary, the spatial importance map models the spatial configuration of a set of distinctive target landmarks. This is illustrated in Figure 5.4(a).

The consistency of the saliency patterns of (5.16), at times t and $t+1$, can be verified by computing the cross-correlation between the saliency map S_k at time $t+1$ and the spatial importance map $\mathcal{T}_k$ learned at time t,

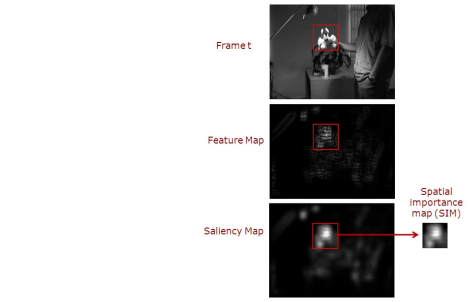
$$R_k(l) = \left\langle S_k|_{\mathcal{W}_l^1}, \mathcal{T}_k \right\rangle, \tag{5.18}$$

where $S_k|_{\mathcal{W}_l^1}$ is the restriction of S_k to the target window $\mathcal{W}_l^1$, and $< \cdot, \cdot >$ a dot-product. The final saliency measure for the set of K feature responses is

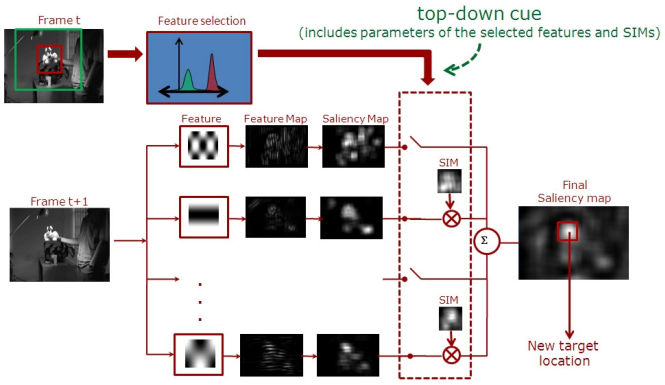
$$S_T(l) = \sum_{k=1}^{K} R_k(l). \tag{5.19}$$

Its computation is illustrated in Figure 5.4(b).

The location l_{t+1}^* of largest saliency, within a neighborhood $\mathcal{W}_{l_t^*}^s$ of the last known target position l_t^*, is selected as the new position of the target at time $t+1$. The feature



(a)



(b)

Figure 5.4: (a) Spatial importance map (SIM) : for each feature, a saliency template of the target is stored at time t. (b) Target localization at $t+1$: for each selected feature, a top-down saliency map is computed with (5.16), and then correlated with the SIM from time t using (5.18). These saliency maps are combined to produce the overall saliency map, the maximum of which is taken to be the new location of the target.

statistics of target and background windows are updated in an online manner, using

$$\sigma^2_{k,i}(t+1) = \begin{cases} \frac{1-\lambda}{n} \sum_{j|y^j_k \in \mathcal{D}_i} \left(y^j_k\right)^2 + \lambda \sigma^2_{k,i}(t), & \text{if } t > 0 \\ \frac{1}{n} \sum_{j|y^j_k \in \mathcal{D}_i} \left(y^j_k\right)^2 & \text{if } t = 0 \end{cases} \tag{5.20}$$

where $\mathcal{D}_i$ is the sample of examples collected from class i at time $t+1$, and λ a decay factor. These statistics are then used for target detection at time $t+2$, and the procedure is iterated.

5.3.6 Scale Adaptive Tracking

Target scale can vary significantly as targets move towards or away from the camera. Trackers that do not adapt to these variations end up relying on a target window

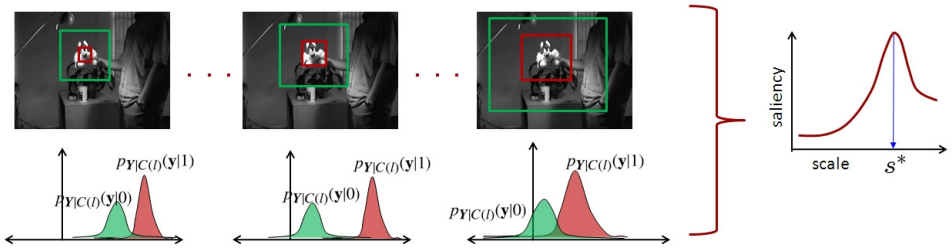


Figure 5.5: Saliency-based scale adaptation. The mutual information between the selected salient features and the class label is evaluated over a scale space. The scale at which saliency peaks is chosen as the optimal tracker scale. This is the scale of largest discrimination between target and background.

that either 1) includes background (when the target shrinks) or 2) excludes foreground (when it grows), and can easily drift. This has motivated a number of scale adaptive extensions of tracking algorithms, ranging from the combination of tracking and scale space representations [25] to specific enhancements applicable only to some trackers, e.g. mean shift [40, 15, 189]. However, scale adaptivity has received little attention in the discriminant tracking literature. Saliency-based tracking offers a natural solution to this problem, since scale and saliency are strongly related [93]. In fact, scale adaptation can be achieved as a *byproduct* of discriminant center-surround saliency: the scale of the target is simply that of the center-surround operator that maximizes target/background discrimination. To determine this scale, at a given target location, it suffices to search over a discrete scale space $s \in (s_{min}, s_{max})$ of target and background window sizes. For each s, the GGD parameters $\alpha^s_{k,i}, \alpha^s_k$ are computed from the feature responses in the target and background windows. This can be done efficiently through the use of integral images [175]. For each feature k, an integral image of the second moment of feature responses $\mathcal{I}_k(l) = \sum_{j \preceq l} \left(y^j_k\right)^2$, where $j \preceq l$ if location j is not below or to the right of location l, is first computed. The variance estimate of (5.20) within a window $\mathcal{D}_i$ of scale s determined by bottom-right, upper-right, bottom-left and upper-left coordinates $l^s_{br}, l^s_{ur}, l^s_{bl}, l^s_{ul}$ is then

$$(\sigma^2_{k,i})^s = \frac{1}{n}[\mathcal{I}(l^s_{br}) - \mathcal{I}(l^s_{ur}) - \mathcal{I}(l^s_{bl}) + \mathcal{I}(l^s_{ul})], \tag{5.21}$$

where n is the number of pixels in $\mathcal{D}_i$. The GGD parameters are finally estimated with (5.20) and (5.6).

Given a set of estimates of the GGD parameters $\alpha^s_{k,i}, \alpha^s_k$, at all window sizes $s \in (s_{min}, s_{max})$, the optimal scale is that at which the center-surround saliency measure peaks:

$$\begin{aligned} s^* &= \underset{s:s\in\mathcal{S}_p}{\text{argmax}} \sum_k I_s(Y_k, C) \\ I_s(Y_k, C) &= \sum_{i=0}^{1} \pi_i \left(\log\left(\frac{\alpha^s_k}{\alpha^s_{k,i}}\right) + \frac{1}{\beta}\left[\left(\frac{\alpha^s_{k,i}}{\alpha^s_k}\right)^\beta - 1\right]\right). \\ \mathcal{S}_p &= \left\{ s : \frac{\partial(\sum_k I_s(Y_k, C))}{\partial s} = 0, \frac{\partial^2(\sum_k I_s(Y_k, C))}{\partial s^2} < 0 \right\} \end{aligned} \tag{5.22}$$

As illustrated in Figure 5.5, this is the scale at which the discrimination between target and background is largest.

5.3.7 Features

Discriminant tracking can be implemented with any set of bandpass features. In this work, we rely on a combination of discrete cosine transform (DCT) filters to account for spatial information and 3D spatiotemporal Gabor filters to account for motion. DCT features are computed by representing each frame as a Gaussian pyramid and convolving each layer of the pyramid with 8×8 DCT basis functions. The spatiotemporal features are based on the 3D Gabor filters of [8, 73].

$$g(x, y, t) = \frac{1}{\sqrt{2}\pi^{\frac{3}{2}}\sigma_x\sigma_y\sigma_t} \sin(2\pi\omega_{x_0}x + 2\pi\omega_{y_0}y + 2\pi\omega_{t_0}t) e^{-(\frac{x^2}{2\sigma_x^2} + \frac{y^2}{2\sigma_y^2} + \frac{t^2}{2\sigma_t^2})} \tag{5.23}$$

where $\omega_{x_0}, \omega_{y_0}$ is a spatial frequency, ω_{t_0} a temporal frequency, and the 3D Gaussian envelope has standard deviations $\sigma_x, \sigma_y, \sigma_t$. This type of filtering is biologically plausible and has been shown to comply with the physiology and psychophysics of the early stages of the visual cortex [8]. Filters tuned to a single spatial frequency of 0.25 cycles/pixel and temporal frequencies of 0 cycles/frames (stationary objects) and ±0.25 cycles/frames (objects moving to the left or right) were chosen, for a total of 3 motion based filters.

It should be noted that, while the discriminant tracker does not require explicit modeling of target dynamics (e.g. through Kalman or particle filtering [83]), the inclusion of spatiotemporal features guarantees their *implicit* modeling. For example, if a

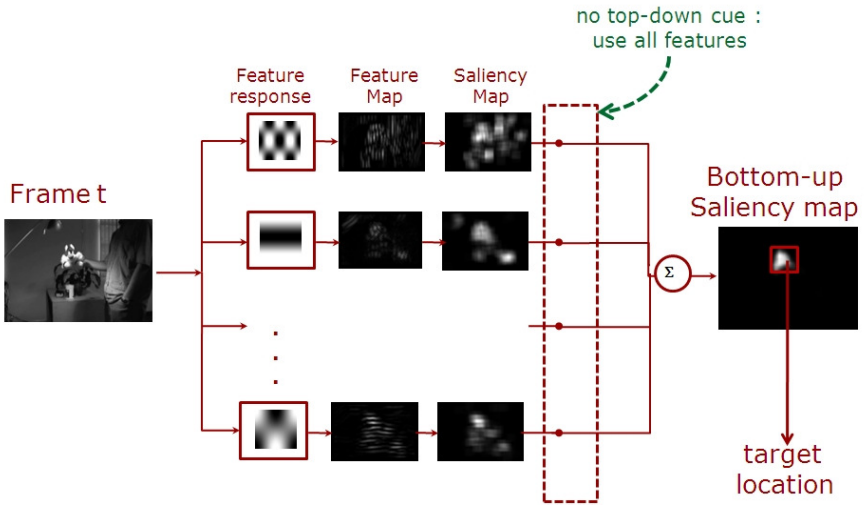


Figure 5.6: Target initialization. A saliency map is computed for each feature, according to (5.10). Feature saliency maps are combined to produce the overall saliency map, the maximum of which is taken to be the initial location of the target.

target is moving to the right at time t, the associated spatiotemporal filter is likely to be discriminant at that time. The selection of this filter as a salient feature implies that locations of right-moving objects are more likely to be declared salient at time $t+1$. Hence, the tracker has some ability to *predict* the dynamics of the target. This ability obviously increases with the addition of spatiotemporal filters to the feature set. The limited set used in this work is mostly due to the desire to guarantee low complexity. The implicit modeling of target dynamics is further reinforced by the restriction of the target search to the window $\mathcal{W}^s_{l^*}$. This assumes that targets do not instantaneously jump beyond the region of the focus of attention, i.e. that target motion is smooth.

5.3.8 Automatic tracker initialization

Most tracking algorithms assume a known initial target location l^* and bounding box $\mathcal{W}^1_{l^*}$ [39, 12]. However, these are not available in most tracking applications. While many initialization strategies, such as background subtraction and blob or motion detection, have been proposed [39], they are mostly heuristic. A more principled approach, based on bootstrapping a weak and generic target model for automatic initialization, was proposed in [166]. However, it requires a pre-specified target model, and some degree

of supervision to adapt it to different scenes. Saliency-based tracking provides a more natural solution to the initialization problem: to declare as targets the locations of largest bottom-up saliency. This is implemented by evaluating (5.10) at all locations of the visual field, and finding the most salient (or the set of most salient locations if multiple objects are to be tracked). If desired, the search can also be performed over target scales, i.e.

$$(l^*, s^*) = \operatorname*{argmax}_{l,s} \sum_k I_{l,s}(Y_k, C) \tag{5.24}$$

where,

$$I_{l,s}(Y_k, C) = \sum_{i=0}^{1} \pi_i \left(\log\left(\frac{\alpha_k^{l,s}}{\alpha_{k,i}^{l,s}}\right) + \frac{1}{\beta}\left[\left(\frac{\alpha_{k,i}^{l,s}}{\alpha_k^{l,s}}\right)^{\beta} - 1\right]\right). \tag{5.25}$$

where the parameters $\alpha_{k,i}^{l,s}, \alpha_k^{l,s}$ are learned from the windows associated with a center-surround operator of scale s centered at location l. As before, these parameters can be computed efficiently with resort to integral images. Overall, the initialization procedure finds the regions whose motion and appearance is most distinct from those of the surrounding background.

The use of (5.24) has a number of appealing properties. First, it can be seen as an optimal (in the discriminant sense) form of background subtraction. In fact, it is a simplification of a state-of-the-art formulation of background subtraction that performs well even on highly dynamic backgrounds [111]. The proposed simplification sacrifices the ability to model complex dynamics for the sake of computational tractability. Second, while the use of spatiotemporal features enables it to account for both target appearance and motion, it is robust to camera motion. This follows from the fact that only motion different from that of the background can be declared salient. For example, an object followed by a panning camera is considered salient. Third, it reduces initialization to a special case of discriminant tracking. In the absence of prior information about which features are discriminant for target detection, the tracker simply uses all of them. This unification of tracker initialization and operation is not possible for most previous trackers.

5.4 Experiments and Results

The performance of the proposed discriminant saliency tracker (DST) was evaluated with an extensive set of experiments. We next report the results of this evaluation.

5.4.1 Comparison to previous trackers

The saliency based tracker was compared to four trackers in the literature: three discriminant trackers, the MILTracker of [13], the method of Collins et al. [39], and the ensemble tracker of [12], and the incremental visual tracker (IVT) of [145]. The latter represents the state of the art in appearance-based tracking. Software for the MILTracker and IVT was obtained from the authors' webpages. Since no implementations are publicly available for the Collins and ensemble trackers, these algorithms were implemented according to the descriptions in [39, 12].

The performance of all five methods was evaluated against manual groundtruth. The tracking error for a frame at time t was defined using the overlap measure of [53] as the normalized lack of overlap between the groundtruth target bounding box, G^t, and that produced by the tracker, B^t. Performance is evaluated by the average tracking error over a sequence of T frames,

$$\epsilon = \frac{1}{T}\sum_t \left(1 - \frac{\sum_{ij} G^t_{i,j} B^t_{i,j}}{\sum_{ij} G^t_{i,j} + \sum_{ij} B^t_{i,j} - \sum_{ij} G^t_{i,j} B^t_{i,j}}\right). \tag{5.26}$$

where the error $\epsilon = 0$ for perfectly correct tracking, while for complete loss of tracking, $\epsilon = 1$.

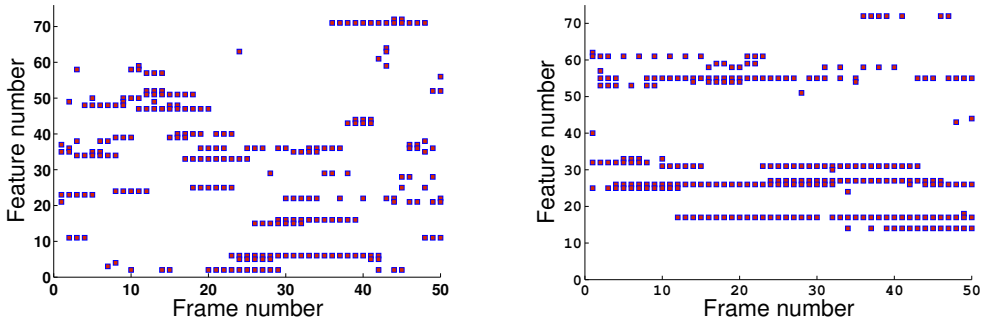
The test video sequences were selected from diverse sources (e.g previous works, standard databases, and the web). All sequences include challenging tracking scenarios, such as varying illumination, complete object rotation, or change in perspective. For instance, the "motinas_toni_change_ill" sequence of [109] shows a person turning by 360°, in extremely low light (Figure 5.8(a)), while the "athlete" sequence includes extreme variations of appearance due to occlusion and strong video compression artifacts (Figure 5.8 (b)). The "skater" sequence (Figure 5.8 (d)), and "CAVIAR" sequence (from [1]) have severe partial occlusions. To increase the difficulty of the tracking task,

all sequences were converted to grayscale. To account for this, the Collins tracker was implemented with DCT features, instead of the R,G,B color features proposed in [39]. All five algorithms were manually initialized with target bounding box in the first frame. The background bounding box was assumed to have an edge 4 times larger than the corresponding edge of the target box.

The saliency based tracker used a two-level Gaussian pyramid, leading to a total of $N = 3 + 64 \times 2 = 131$ features (8×8 DCT features per level plus three spatiotemporal Gabor features). The number of selected salient features, K, is a tunable parameter. To understand its impact on tracking performance, it was varied in the range $[1, 29]$, for two representative sequences. Good performance was obtained for any $K \geq 3$, albeit tracking accuracy improved with the number of features, at the expense of increased computation. To guarantee a realistic balance between tracking performance and computation, K was set to 5 in all subsequent experiments. Figure 5.7 shows the 5 selected features for the first 50 frames in two representative sequences. The plot shows that the same or very similar features are selected in successive frames, and the set of selected features is fairly stable over time. The search neighborhood, $\mathcal{W}^s_{l^*}$, was set to a rectangular region centered at the current target position l^* with size twice that of the object bounding box.

To explicitly understand the contribution of two of the components of the saliency based tracker - a) the spatial importance map (SIM) and (b) spatio-temporal features, we created four variants of the tracker depending on which of the two components where included. These are termed "Sal" (only using the saliency measure of (8.1) and the spatial DCT features), "Sal+SIM" (saliency and SIM), "Sal+ST" (saliency with spatio-temporal Gabor features included along with DCT features, but no SIM) and "Sal+SIM+ST" (saliency including spatio-temporal features and the SIM). These four variants were also included in the experimental evaluation.

Table 5.1 presents the errors measured on a set of 13 sequences. First, the results show that spatio-temporal features help improve tracking performance, as seen from lower average error rate for "Sal+ST" over the baseline version "Sal". As discussed in Section 5.3.7, these features provide the tracker with some *implicit* ability to account for the motion of the target, which are an important discriminatory cue for most objects.



(a) (b)

Figure 5.7: Features selected in the first 50 frames on (a) "karlsruhe" and (b) "sylvester". The spatial features are numbered from 1 to 64, and correspond to the zig-zag scanning order of the DCT basis functions, while the three spatio-temporal features are numbered from 70 to 72.

Additionally, it is seen that the spatial importance map ("Sal+SIM") also leads to a significant improvement in tracking accuracy. Finally, it is clear that the effects of SIM and spatio-temporal features are complementary and the version with both components, performs the best on average. In the following discussion, we refer to this version as the discriminant saliency tracker (DST).

Overall, DST or its variants are the top performers on 8 sequences. Among the remaning methods, MIL is the best performer, with lowest error rates on three sequences. More importantly, the tracking error of DST is very close to that of the best performing method in most of the sequences where it does not produce the best results. On the other hand, in the sequences where it is the top performer, the error of DST tends to be substantially smaller than those of all other methods. This is captured by fact that the average error, across sequences, of DST is about 64% that of the next best method (MIL). Alternatively, it can be seen from the fact that, while DST never loses track, this happens for all other methods in four of the sequences ("ram", "skater", "motinas", and "athlete").

Figure 5.8 illustrates the tracking results on four of the sequences considered. The qualitative performance of IVT and the ensemble tracker is quite poor, as these methods lose the target in most scenes. Somewhat better performance is achieved by the Collins and MIL trackers. However, these methods lose the target when it under-

Table 5.1: Average tracking error of the five trackers compared. 0 indicates perfect tracking, 1 complete lack of overlap between groundtruth and target bounding box produced by the tracker.

Sequence	IVT	Collins	Ensemble	MIL	Sal	Sal+SIM	Sal+ST	DST
coke11	0.97	0.76	0.71	0.68	**0.62**	0.68	0.63	0.68
tiger2	0.80	0.78	0.88	**0.38**	0.64	0.77	0.78	0.44
karls	0.64	0.47	0.93	**0.29**	0.52	0.30	0.53	0.31
dtneu	0.93	0.27	0.96	0.49	0.21	0.21	**0.15**	0.26
plushtoy	**0.11**	0.37	0.38	0.17	0.16	0.26	0.21	0.25
ram	0.77	0.86	0.87	0.64	0.36	0.77	**0.33**	**0.33**
ballroom	0.62	0.38	0.70	**0.34**	0.39	0.46	0.38	0.44
roadcrossing	0.51	0.74	0.83	0.46	0.87	0.78	0.77	**0.45**
motinas	0.60	0.47	0.73	0.61	0.95	**0.22**	0.92	0.24
athlete	0.98	0.78	0.94	0.92	0.75	0.41	0.75	**0.37**
skater	0.94	0.49	0.62	0.93	0.47	0.33	0.36	**0.30**
CAVIAR	0.34	0.56	0.96	0.48	**0.29**	0.33	0.73	0.31
seq10	**0.03**	0.99	0.94	0.08	0.95	0.89	0.14	0.14
average	0.63	0.61	0.80	0.50	0.55	0.49	0.51	**0.35**

goes extreme appearance variations, due to partial occlusions, illumination changes or rotation. On the other hand, DST tracks the targets successfully in all sequences. The results on "seq10", a very long sequence used in [67], show that DST is also able to track over long durations reliably without drifting (Figure 5.8(c)).

Overall, it is clear that DST has the best performance. Videos of all tracking results are available from [5].

5.4.2 Scale Adaptive Tracking

To test scale adaptivity, the performance of the DST was evaluated on various sequences of widely varying target size. The comparison was restricted to IVT, since no scale adaptive extensions are available for the other methods. The initial position and size of the target were manually specified, since IVT has no ability for automatic initialization. Examples of the tracking results are shown in Figures 5.9. Note that these sequences are challenging in many ways. Besides wide scale variability, the target can change appearance quite dramatically due to a 360° rotation, and non-rigid motion (on

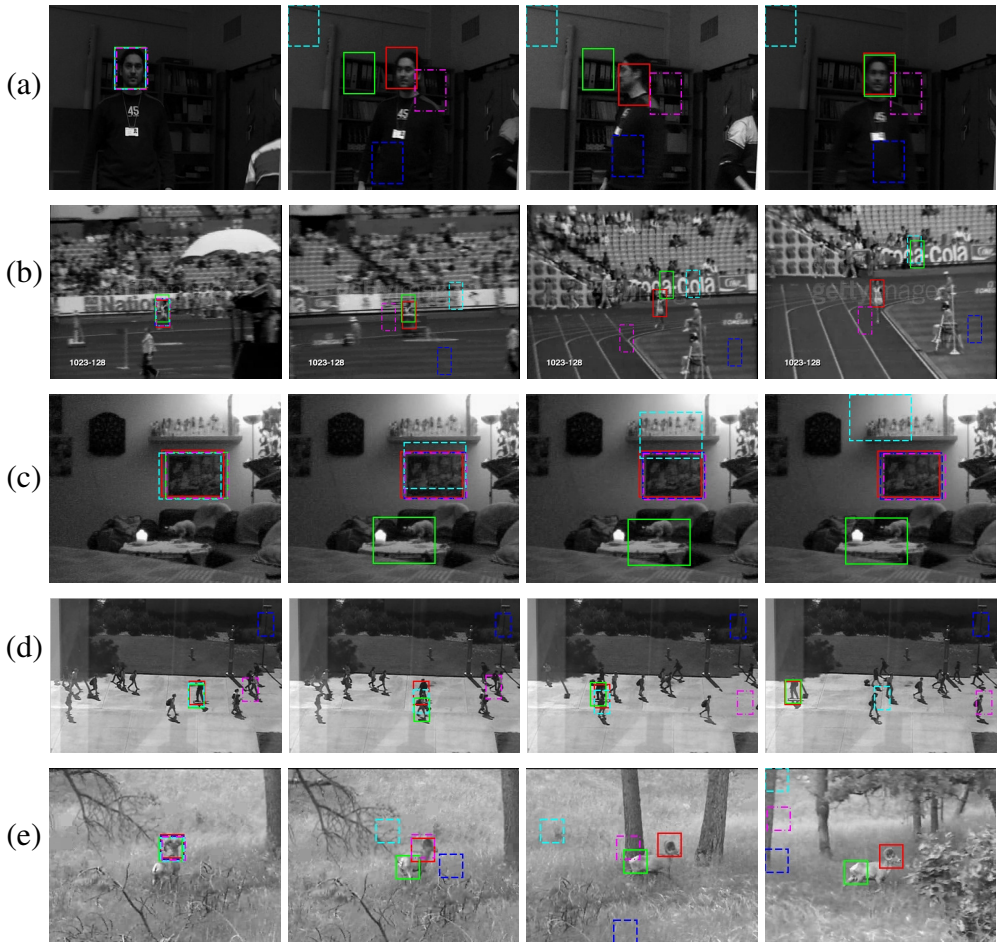


Figure 5.8: Tracking results on a) "motinas_toni_change_ill" [109] - the person turns around and the illumination changes drastically, b) 'athlete"- a person running inside a stadium. The video is very noisy and the target appearance changes widely, c) "seq10" - extremely long video sequence used in [67] to test for drifting, (d) "skater" on a pedestrian walkway - the target undergoes partial occlusions on multiple occasions and (e) "ram" walking in the woods. Target locations: DST - thick red box, Collins - thick green box, ensemble - cyan dashed box, IVT - blue dashed box, MIL - magenta dashed box.

"gravel" the subject turns, picks a rock, and throws it in the water), as well as perspective effects (on "dirtbike" the motorcycle approaches the camera from the left and leaves to the right), and the background varies substantially (sky, then sand dunes, then strongly shaded background on "dirtbike"). While IVT loses track in both cases, DST is able maintain track throughout the sequences, accurately tracking the target position. This robustness is due to the continuous updating of the features used to represent both target and background, and the discriminant nature of the tracker. Panels on the extreme right of Figure 5.9 present plots of the variation of target scale over time. It is clear that DST is able to handle a wide variability of target scales, while IVT loses track ("gravel") or dwindles into an infinitesimal target box ("dirtbike"). Table 5.2 summarizes the errors measured on these and two other sequences, confirming the superior performance of DST. Videos of the sequences are again available from [5].

Table 5.2: Comparison of average tracking error of IVT and DST when target scale varies widely

Name	IVT	DST
dirtbike	0.86	**0.33**
speedboat	0.45	**0.38**
gravel	0.76	**0.44**
baseball	0.96	**0.44**
average	0.76	**0.40**

Table 5.3: Comparison of tracking errors for the DST using automatic and manual tracker initialization

Name	Auto Init	Manual Init
dirtbike	**0.33**	**0.33**
surfer	0.33	**0.32**
dog	0.38	**0.37**
skiing	**0.27**	0.28

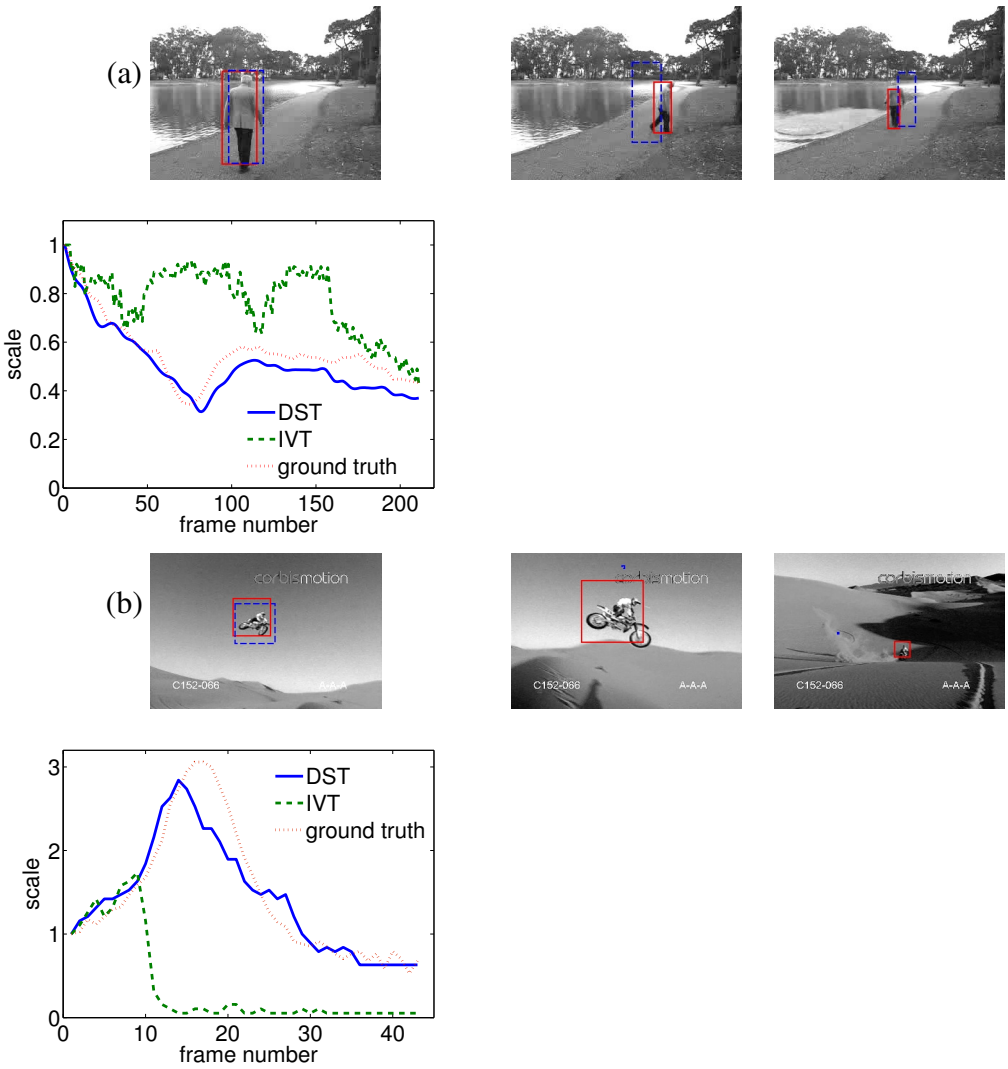


Figure 5.9: Scale adaptive tracking on (a) "gravel" and (b) "dirtbike". Target locations: DST - red box, IVT - dashed blue box. Plots of target scale, expressed as the ratio of target size at a frame to size in the initial frame for the respective sequences are shown below the frames.

5.4.3 Automatic Initialization

Finally we performed a set of experiments designed to evaluate automatic tracker initialization using DST. Since none of the other methods have this capability, no comparison was performed for these sequences. Examples of DST results are shown in Figure 5.10. The tracker uses the bottom-up discriminant saliency procedure of Section 5.3.8 to identify the object to track. The region of maximal saliency is then input to the scale adaptive DST algorithm, which tracks the target through the remaining frames. The leftmost column of Figure 5.10 shows the bottom-up saliency map, and the columns on the right show a few of the subsequent frames (target bounding box shown in red). The tracker initializes the target correctly, and tracks it through substantial variations of scale and pose (note the 3D rotation in "dog").

Table 5.3 presents the error measures obtained for these sequences. The error of DST with automatic initialization is compared to that obtained when the tracker is manually initialized with the groundtruth target bounding box. There is no substantive difference. Overall, these results demonstrate the ability of the DST to perform robust target initialization and accurate scale adaptive tracking, in scenes with complex motion. Videos of all sequences are available in [5].

5.5 Connections to other discriminant trackers

At an abstract level, the proposed DST is similar to previous discriminant trackers [39, 12, 13]. Like the DST, these are center-surround discriminators, equating target to center and background to surround. In fact, they rely on classifier design and target detection operations that are similar in spirit to those of DST. There are, nevertheless, differences of detail that significantly affect tracking performance. As summarized in Table 5.4, these report to the features used, the method employed for their selection, and the confidence measure used for target detection.

All discriminant trackers include a classifier design stage which begins with feature extraction. This maps the space of image pixels into (a presumably lower dimensional) feature space $\mathcal{Y}$, where it is easier to discriminate target from background. The transformation can be linear or non-linear. Collins et al. [39] use linear combinations of

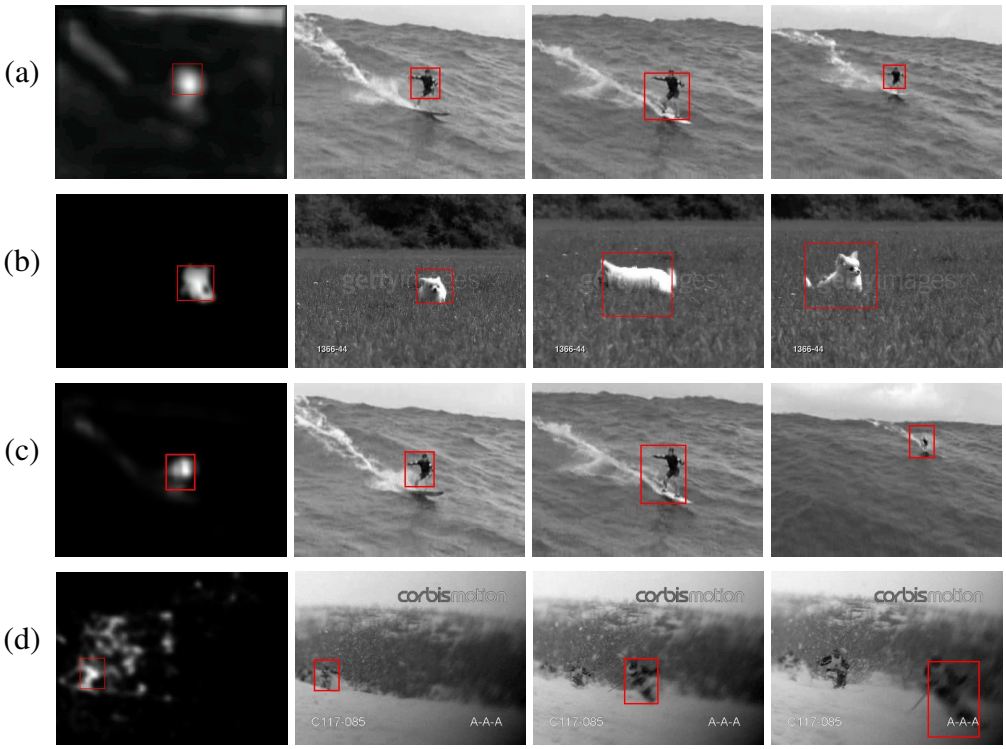


Figure 5.10: Automatic initialization and tracking. Bottom-up saliency map used to initialize the tracker is shown on the left column. Target bounding boxes are shown in red. a) "surfer" b) "dog" (c) "surfer" and (d) "skiing". Target locations in subsequent frames are shown in red.

Table 5.4: Connections between the four discriminant trackers in terms of the components used.

Method	**Features**	**Feature selection**	**Confidence Measure for Target Detection**
Collins	RGB features	Fisher discriminant like variance ratio	Log-likelihood ratio
Ensemble	RGB+HoG features	Boosting of hyperplane classifiers	Margin from boundary
MIL	Haar features	Boosting of decision stumps	Posterior target class probability
DST	DCT+Gabor filters	Salient features selected by MMI	Information measure

R,G,B pixel values, ensemble tracking [12] complements R,G,B values with histograms of oriented gradients [46], and MIL [13] relies on Haar wavelets. DST relies on a combination of DCT and spatiotemporal filters. In our experience, feature selection in tracking is not different from feature selection in any other learning problem. More features will generally improve performance, at the price of higher computational complexity. While the gains tend to saturate after a relatively rich set of features is available, many feature sets can be used. Gradient histograms, Haar wavelets, and DCT filters are all good examples. We caution, however, against the standard practice of relying solely on color histograms [41]. While color is a sufficiently discriminant cue for many sequences, it can artificially inflate the effectiveness of the tracker. For example, it is not difficult to track an extremely complex object that rotates in 3D, if it is the only red blob in the scene. Our choice of grayscale sequences was intended to minimize these types of effects.

First, our results show that it is important for the feature set to include a combination of spatial and spatiotemoporal features. As discussed in Section 5.3.7, these features provide the tracker with some *implicit* ability to predict the dynamics of the target, which are an important discriminatory cue for most objects. As an illustration, consider the "roadcrossing" sequence of Figure 5.11. Panels (a) and (b) show tracking

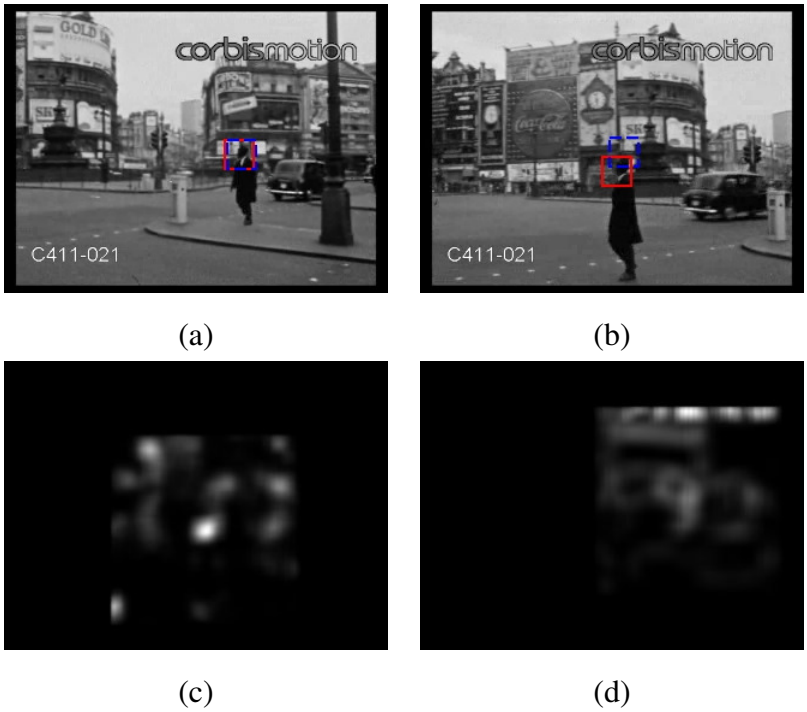


Figure 5.11: (a) and (b) tracking on "roadcrossing". Target locations: full DST - red box, DST with spatial features only - blue box. The average tracking error is 0.1 for the former and 0.51 for the latter. (c) and (d) are associated saliency maps for the frame in (b). (c) shows full DST, and (d) DST with spatial features alone.

results with (red box) and without (blue box) spatiotemporal features. The substantial amount of background clutter in this scene hampers the learning of a tracking model from spatial features alone, causing the tracker to fail. The addition of spatiotemporal features provides a discriminant cue (different motion of pedestrian and background) that is much easier to model and detect, even though both target and background are moving (panning camera). This can be confirmed by inspecting the saliency maps associated with the frame where the spatial tracker starts to lose track. It is clear that the addition of spatiotemporal features enables the discrimination, despite the background clutter (Figure 5.11(c)), but this is not possible in their absence (Figure 5.11(d)).

With respect to feature selection and classifier design, all discriminant trackers analyze the feature set for target/background discrimination. Collins et al. [39] first compute histograms of filter responses on the R,G,B channels of both target and background, and construct a log likelihood ratio between the two class histograms, considering this

a new non-linear feature. Feature discrimination is evaluated by a Fisher discriminant-like *variance ratio* that measures how tightly clustered the log-likelihood ratios are for the two classes. This is equivalent to transforming the features into a non-linear space and learning a linear classifier in that space. It is optimal, in the minimum probability of error sense, only when the classes are *Gaussian* and have equal covariance, after the feature transformation. Overall, the tracker suffers from the fact that this discrimination measure is somewhat heuristic. The distribution of log-likelihood ratios is hard to characterize [37], the assumption of unimodality (Gaussianity) does not hold in general (i.e. for all features), and is especially troubling when there is background clutter. There is even less evidence in support of the assumption of equal class variance. These observations could account for the limited effectiveness of the tracker.

The ensemble tracker [12] relies on a set ("ensemble") of weak hyperplane classifiers to separate target from background. Each weak learner implements a threshold on a linear combination of the original features. A simpler approach is used by the MIL tracker [13], where each weak learner is a decision stump, i.e. a threshold on one of the original features. Both trackers rely on the classification error rate as measure of discrimination for feature selection. While this is a close approximation to the mutual information used by the DST [171], the feature selection procedure is quite different: both the ensemble and MIL trackers rely on boosting (AdaBoost and MILBoost respectively). Boosting has a number of disadvantages for tracking. The first is that it is too sensitive to outliers. This a major limitation, since the target and background classes of a tracking problem are rarely exclusive. On the contrary, a certain amount of background is usually covered by the target window and vice-versa. When the learning procedure lacks robustness, the resulting outliers can easily bias the decision rule, and the tracker tends to drift. As this happens, the number of outliers increases, and the target and background classes become gradually more ill-defined, until tracking is simply lost. The sensitivity of boosting to outliers is well known in machine learning, where a number of extensions have been proposed to address the problem [115]. This is indeed the difference between the ensemble and MIL trackers. The latter implements boosting under the MIL formalism, exactly to decrease outlier sensitivity. The excessive outlier sensitivity of the ensemble tracker justifies the fact that it has the weakest performance

of all methods tested. By minimizing this problem, MIL achieves significantly better results.

A second limitation of boosting is that it tends not to perform well with the amounts of data and computation available to tracking problems. This is not a limitation of boosting per se, but of boosting the weak learners commonly used in vision applications. Since the decision function to approximate can be fairly complex, the space has non-trivial dimensionality, and the weak learners are very simple (e.g. a simple feature threshold for decision stumps), a large number of weak learners are needed to produce a good classifier. This is infeasible for tracking, where 1) not enough computation is available to perform a large number of boosting iterations between two video frames, and 2) not enough training data is available to constrain the learning of a large combination of weak learners. The second problem could be minimized by extending the training set, e.g. by collecting target samples over a large number of frames, but this would increase the difficulty of the first problem. Furthermore, the tracker would have great difficulty in adapting to object variability. Instead, boosting-based trackers try to solve the problem with a small number of learners and a limited training set (a few frames, at most). This produces a classifier with little generalization ability, which does not perform well when there are large variations of appearance, due to effects such as the rotating objects or the noisy athlete sequence of Figure 5.8.

The results above show that the proposed DST, and associated MMI feature selection, have a better trade-off between complexity and generalization. A similar observation has been reported for image classification, where mutual information based feature selection has been shown to outperform boosting based methods [56]. For tracking, this translates into a better trade-off between robustness and adaptivity of the decision rule, and justifies the superior performance of DST even when the outlier sensitivity of boosting is minimized (i.e. in comparison to MIL).

5.5.1 Target detection

Given a set of discriminant features, all discriminant trackers use some measure of classification confidence to detect the target in the next video frame. Collins [39] suggests a decision rule based on the log-likelihood ratio between the target and back-

ground hypotheses. This is not fundamentally different from the information measure of (2.5). The confidence measure of the ensemble tracker is the classification margin of the boosted ensemble. This is a measure of the level of belief in the classification result, and is directly analogous to the saliency measure (2.1). For the MIL tracker, the confidence measure is the posterior target probability given the feature responses at each image location. This has an exact correspondence to the information measure of (2.5), which is simply a monotonic non-linear function of the posterior target probability. In summary, the confidence measures used by the Collins, ensemble, and MIL trackers are not fundamentally different from the saliency measure of the DST.

5.6 Conclusion

In this work, we have shown that discriminant tracking follows naturally from the discriminant formulation of visual saliency. In particular, optimal tracking (in the decision-theoretic sense) can be implemented with a combination of bottom-up center-surround discriminant saliency and spatial attention for learning, feature-based attention for feature selection, and top-down saliency for target detection. This was exploited to construct a simple and computationally efficient framework for tracking, which is consistent with what is known about the attentional mechanisms of biological vision, and provides a unified solution to the problems of classifier design, target detection, automatic tracker initialization, and scale adaptation. Experimental comparison with previous trackers shows that the proposed discriminant saliency tracker is significantly more robust. An implementation of this tracker in C, without any optimization, currently runs at $\sim$ 1.5 frames per second (fps), on a standard PC without special hardware. On the same machine, the running times of other discriminant trackers are comparable ($\sim$ 4fps for MIL and $\sim$ 3fps for the Collins tracker).

Among its shortcomings, DST does not explicitly retain target features that appear in the previous frames. Therefore, it cannot handle prolonged partial or complete occlusions. Also, as the approach depends on finding features that can discriminate the target from the background, DST is not suitable when there are objects very similar to the target in the background or for tracking large targets with inadequate backgrounds.

Finally, DST has been designed for tracking single targets. To track multiple targets, DST has to be augmented with additional modules such as an identity management scheme.

5.7 Acknowledgments

We thank Dr. Boris Babenko for providing the code for the MIL tracker [13], and Dr. Helmut Grabner for video sequences used in [67].

The text of Chapter 5, in full, is based on the material as it appears in: V. Mahadevan, and N. Vasconcelos, "Biologically inspired object tracking using center-surround saliency mechanisms", in review, IEEE Transactions on Pattern Analysis and Machine Intelligence; V. Mahadevan, and N. Vasconcelos, "Saliency Based Discriminant Tracking", IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 1007-1013, 2009. The dissertation author was a primary researcher and an author of the cited material.

Algorithm 2 Tracking Using Discriminant Saliency

Input: Initial target location l^*, $t = 0$, initial frame $\mathcal{I}_0$ (M pixel locations), feature set $Y_k, k \in \{1, \ldots, N\}$, and a target number of features K.

while Next frame exists **do**

Learning and Feature Selection:

Extract target $\mathcal{W}^1_{l^*}$ and surround $\mathcal{W}^0_{l^*}$ windows from $\mathcal{I}_t$.

for $k = \{1, \ldots, N\}$ **do**

From the responses of Y_k in $\mathcal{W}^i_{l^*}$, $i \in \{0, 1\}$, estimate the variances $\sigma^2_{k,i}$ using (5.20), and the scale parameters $\alpha_{k,i}$, using (5.6).

From the responses of Y_k in $\mathcal{W}^1_{l^*} \cup \mathcal{W}^0_{l^*}$, estimate the variances σ^2_k using (5.20), and the scale parameters α_k, using (5.6).

Compute the parameters ξ_k, T_k, and t_k, using (8.6) and (5.15) .

Compute $I(Y_k, C)$, using (5.9).

Compute spatial importance map $\mathcal{T}_k$ using (5.17).

end for

Output: return the K features with $\xi_k > 0$ and largest $I(Y_k, C)$, the corresponding parameters ξ_k, T_k, t_k, and spatial importance maps $\mathcal{T}_k$.

Set $t = t + 1$.

Target detection:

for $k = \{1, \ldots, K\}$ **do**

for $m = \{1, \ldots, M\}$ **do**

Compute the response y_m of Y_k, and the saliency values $S_k(l_m)$, at location l_m of pixel m of $\mathcal{I}_t$, using (5.16).

Compute $R_k(l_m)$, using (5.18).

end for

end for

for $m = \{1, \ldots, M\}$ **do**

Compute the saliency $S_T(l_m)$, at l_m, with (5.19).

end for

Output: Set $l^* = \text{argmax}_{l_m} S_T(l_m), \{m \in \mathcal{W}^s_{l^*}\}$.

end while

Chapter 6

The Saliency hypothesis for tracking

6.1 Introduction

Biological vision systems have evolved sophisticated tracking mechanisms, capable of tracking complex objects, undergoing complex motion, in challenging environments, e.g. cluttered scenes in low-light. These mechanisms have been an area of active research in both neurophysiology [45, 148] and psychophysics [130], where research has been devoted to the study of object tracking by humans [140]. This effort has produced several models of multi-object tracking, that account for the experimental evidence from human psychometric data [130]. Prominent among these are the FINST model of Pylyshyn [140], and the object file model of Kahnemman et al [94]. However, these models are not quantitative, and only explain the psychophysics of tracking simple stimuli, such as dots or bars. They do not specify a set of computations for the implementation of a general purpose tracking algorithm, and it is unclear how they could be applied to natural scenes. While some computational models for multiple object tracking (MOT) such as the oscillatory neural network model of Kazanovich et al. [97], and the particle filter based model of Vul et al. [177], have been proposed, there have been no attempts to demonstrate their applicability to real video scenes.

In the previous chapter, we noted that the best results among tracking algorithms are obtained for *discriminant trackers* which object tracking as incremental target/background classification [110, 39, 12, 66]. These train a classifier to distinguish target from background at each frame. This classifier is then used to determine the loca-

tion of the target in the next frame. Target and background are extracted at this location, the classifier updated, and the process iterated.

The explicit modeling of the visual information that differentiates target from background, i.e. discriminant features of object appearance and/or motion, leads to a much more accurate determination of the object support than what is possible with classical predictive tracking, e.g. modeling of target dynamics with Kalman filtering or condensation [83]. This improves the tracker accuracy and minimizes the contamination of the target model. The superior performance of discriminant trackers is consistent with what is known about biological tracking. For example, there is clear evidence that human perception of target correspondences relies much more on appearance features than on the prediction of object dynamics [54], and that motion extrapolation is not used in object tracking [98].

The discriminant saliency based tracker we proposed in Chapter 5 has, in fact, postulated a connection between saliency, one of the core processes of early biological vision, and discriminant tracking: that the ability to track objects is a side-effect of the saliency mechanisms that are known to guide the deployment of attention. More precisely,we have hypothesized that *tracking is a simple consequence of object-based tuning, over time, of the mechanisms used by the attentional system to implement top-down saliency*. We refer to this as the *saliency hypothesis for tracking*. Under this hypothesis, in Chapter 5 we proposed a tracker based on the *discriminant saliency* principle of [65]. This is a principle for bottom-up center-surround saliency, which poses saliency as discrimination between a target (center) and a null (surround) hypothesis. Center-surround discriminant saliency has previously been shown to predict various psychophysical traits of human saliency and visual search performance [60]. The extension proposed in Chapter 5, to the tracking problem, endows discriminant saliency with a top-down feature selection mechanism. This mechanism enhances features that respond strongly to the target and weakly to the background, transforming the saliency operation from a search for locations where center is distinct from the surround, to a search for locations where target is present in the center but not in the surround. Chapter 5 has shown that this tracker has state-of-the-art performance on a number of tracking benchmarks from the computer vision literature.

Besides improved tracking algorithms for computer vision, the potential connection between saliency and tracking is interesting in multiple ways. First, it could unify the study of the two processes. While there are universally accepted protocols for the study of attention, e.g. extensive saliency psychophysics [129] and a large visual search literature [184], the study of visual tracking is much less developed. If the processes are indeed related, variations of the protocols used to study attention could be applied to tracking. Second, it could lead to unified computational models for the neural circuits that solve the two tasks. This would place more stringent conditions on the models, which would have to explain data from both saliency and tracking, and provide a stronger evolutionary justification for models that meet such conditions. Third, it could provide novel justifications for both the neurophysiology and the behavior of various neural circuits of early vision, e.g. the need for a combination of circuits that modulate saliency spatially (by implementing a spotlight of attention [138]) and by attribute (by enhancing/suppressing the responses of entire feature channels [116]).

In the next chapters we discuss our contributions along these three areas. We first present results of several psychophysics experiments on the dependence between target saliency and human tracking performance, and demonstrate that the saliency based tracker of Chapter 5 is compliant with this data. These experiments build on well understood properties of saliency, such as pop-out effects, to show that tracking requires discrimination between target and background using a center-surround mechanism. In addition, we characterize the dependence of tracking performance on the extent of discrimination by gradually varying feature contrast between target and distractors in the tracking tasks. The results show that both tracking performance and saliency show highly similar patterns of dependency on feature contrast. This provides strong evidence for the proposed connection between saliency and tracking. Second, we show that the hypothesis has biological support by mapping the saliency-based tracker into a network compliant with the widely accepted neurophysiological models of neurons in area V1 [30] and the middle temporal area (MT) [156], and with the emerging view of attentional control in the lateral intra-parietal area (LIP) [16]. This mapping extends the substantial connections between discriminant saliency and the standard model that have already been shown [65]. In particular, we show that all information required

for optimal (in a decision-theoretic sense) feature selection can be obtained by divisive normalization, across feature channels, of the responses of the saliency network. The resulting network is a biologically plausible optimal model for *both* saliency and tracking. Finally, we show that the tuning of top-down saliency associated with this feature selection mechanism explains the well-known phenomenon of *feature-based attention* [168]. In particular, we show that the tracking network replicates data from feature-based attention experiments with MT neurons. This provides a *functional* justification for feature-based attention (tracking) which complements the functional justification previously available for spatial attention (center-surround saliency) [89].

6.2 Tracking and attention

Various authors have suggested that, in the human visual system, 1) tracking is achieved by attentional mechanisms [32, 10] and 2) the underlying processes could be feature based [112]. For example, a target can be tracked as its appearance changes (in terms of features like orientation, spatial frequency or color), even when superimposed on a distractor [18]. Conversely, it has been shown that attentional tracking fails when the target features cannot be individuated [173, 31]. Additionally, experiments based on the "bouncing-streaming" [150] paradigm have shown that the perceived correspondence of an object in successive time-slices depends much more on the similarity of its featural attributes (shape, orientation, color, texture etc.) than on the predictability (smoothness) of the resulting trajectory [54]. This does not mean that object motion is irrelevant: it is known that targets can easily be tracked even when they have identical appearance to the distractors, as long as they are spatio-temporally distinguishable from the latter [140, 82]. It is thus believed that both spatial and spatio-temporal attributes of the target are attended to, for tracking purposes [18][1]. Overall, while there is plenty of evidence in support of the hypothesis that target features are important for the perception of object persistence, it is not clear which feature subset is actually selected within a tracking task. In fact, there have been no attempts, in the biological tracking literature, to understand the *computational principles* that underlie the selection of such features.

[1] In fact, it is thought that for the purposes of attention, stimulus location can be treated as just another feature [134].

Substantially more research has been devoted to the computational modeling of visual attention [27, 183, 88, 65]. A well studied attentional mechanism, thought to play a role in tracking [48], is *spatial attention*. It enables the visual system to direct attention towards a specific spatial neighborhood of the visual field, commanding what is often referred to as the spotlight of attention [50]. A second mechanism, whose connections to tracking have still not been explored, is *feature-based attention*. It enables the visual system to attend to specific visual features, such as direction of motion, orientation, and color [116], enhancing the responses to such features throughout the visual field.

It is also widely believed that stimulus saliency plays an important role in attention. For example, it is difficult not to attend to stimuli that "pop-out" [88]. Two types of saliency mechanisms are well known. Bottom-up saliency is entirely stimulus driven, and thought to result from a center-surround operation [33]. It identifies stimuli that are distinct from the surrounding background as salient, and facilitates the direction attention to the specific locations of such stimuli. It is thus primarily related to spatial attention. Top-down saliency is tunable for the detection of target features or objects [188]. It is primarily related to feature-based attention mechanisms, and reinforces (suppresses) the responses of salient (non-salient) features. This allows the rapid identification of stimuli in a target class, e.g. faces, by suppressing the responses of features that are not informative about that class.

Discriminant saliency reviewed in Chapter 2 is a generic saliency principle that equates saliency to discrimination between target and background classes [65]. It can be specialized to either bottom-up or top-down saliency. This has been exploited in Chapter 5 to propose a saliency-based approach to discriminant tracking. An object is initially declared salient by a bottom-up saliency detector, establishing a target to be tracked. The classifier stage of the discriminant tracker (cf. Figure 5.1) is achieved by identifying the most salient target features which are then used to design a top-down saliency detector tuned for target detection in the next time step. The object is then detected and the process iterated. Figure 5.4 illustrates the main steps involved in the saliency tracker. The resulting tracker has been shown to achieve state-of-the-art performance in various tracking benchmarks. While the feature selection operation of Chapter 5 (which itself is based on top-down saliency) closely follows what is known about the mechanisms

of feature-based attention, it is not clear if the computations proposed in Chapter 5 are biologically plausible. This motivates two inter-related questions: 1) is it likely that top-down tuning of discriminant saliency could drive tracking in biological visual systems as proposed in Chapter 5? and 2) Is top-down tuning of saliency a plausible model for feature-based attention?

In addition to the good experimental performance reported in Chapter 5, there are several compelling reasons to believe that top-down saliency could be the basis for tracking. The first is that the saliency and tracking tasks are not fundamentally different. Once a target is declared salient, it is likely to stay salient for some period of time. It appears sensible to use the computations already performed for saliency to keep track of where the object is. Hence, there is some evolutionary pressure for a common solution to the two problems. Second, the evidence from the tracking and saliency literatures suggests that both problems are effectively solved by a discriminant formulation, where the goal is to find locations where the center is different from the surround (background). The only difference is that, for tracking, the center must also contain the particular object to track. This suggests that tracking could be performed through top-down tuning of the mechanisms already in place for bottom-up saliency. Namely it would suffice to complement center-surround saliency with a feature-based attention mechanism that suppresses features not informative for target presence. In the succeding chapters we study the hypothesis that saliency is the basis for tracking. We seek evidence in three domains: 1) psychophysics support for the hypothesis 2) biologically plausible implementation of tracking by discriminant saliency, 3) neurophysiological support for this implementation.

We start by reporting on experiments investigating the connections between the psychophysics of tracking and saliency in the next Chapter.

6.3 Acknowledgments

The text of Chapter 6, in full, is based on the material as it appears in: V. Mahadevan, and N. Vasconcelos, “Biological plausibility of the saliency hypothesis for visual tracking”, in preparation. The dissertation author was a primary researcher and

an author of the cited material.

Chapter 7

Human Behavior Studies on Saliency and Tracking

7.1 Introduction

A frequently used approach to test biological plausibility of computational models involves comparison of the model predictions to human behavior (psychophysics) data on appropriate visual stimuli. Since there are no reports in the literature on psychophysics experiments studying the relation between attentional tracking of a single target and its saliency, we designed our own experiments. In the following paragraphs, we describe these experiments.

7.2 Experiment 1: Saliency affects tracking performance

7.2.1 Method

Participants Thirteen subjects with normal or corrected to normal vision participated in the study (age range 22-35, 4 female). The subjects for this and all subsequent experiments reported here provided informed consent and were compensated $8 per hour for their participation. The entire study was approved by the IRB at UCSD.

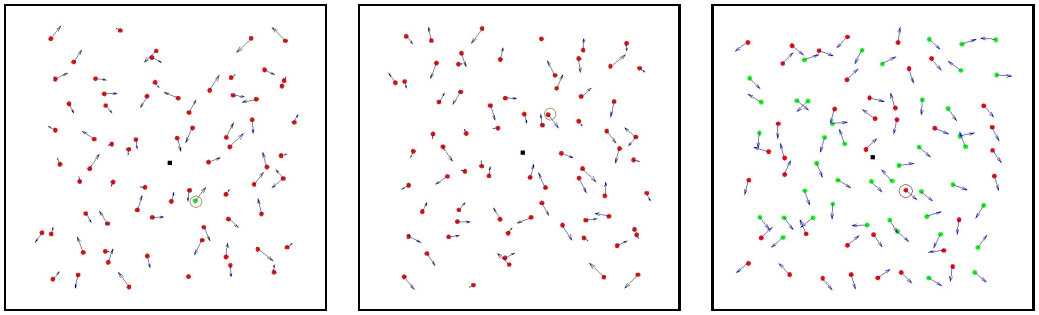
Apparatus The video stimuli were designed using the Psychtoolbox [24] with Matlab v7, running on a Windows XP PC. A 27 inch LCD monitor of size $47.5^\circ \times 30^\circ$ visual angle with resolution set to 1270×1068 pixels was used to present the stimuli, and the viewing distance was set at 57 cm. The same apparatus was used for all subsequent experiments in this work.

Stimuli The experimental setting was inspired by the tracking paradigm of Pylyshyn [140]. Subjects viewed displays containing a *green* target disk surrounded by 70 *red* distractor disks, identical in shape to the target, and a static fixation square.

Procedure At the start of each trial, the target disk was cued with a bounding box, in the first frame of the stimulus. The subjects were asked to track the target covertly, without moving their eyes from the fixation point. On a keystroke from the subject, all disks moved independently, with random motion, for around 7 seconds. Then, the disks stopped moving and the colors of three disks were switched to three new colors - cyan, magenta and blue. Of these, one was the target and the other two the spatially closest distractors. The subjects were asked to identify the target among the three highlighted disks. Participants performed 4 trials each, divided into 2 versions of 2 conditions.

Design The first version tested how tracking is affected by target saliency under two different conditions involving two types of displays. In the first, denoted *salient*, the target remained green throughout the presentation, changing randomly to one of the three highlight colors at the end of the 7 seconds. In the second, denoted *non-salient*, the target remained green for the first half of this period, switched to red for the remaining time, finally turning to a highlight color. While in the first condition the target is salient throughout the presentation, the second makes the target non-salient throughout the latter half of the trial. To eliminate potential effects of any other variables (e.g. target-distractor distances and motion patterns), the non-salient display was created by rotating each frame of a salient display by 90° (and changing the green disk to red in the second half of the presentation). Figure 7.1 shows typical frames from the two types of displays.

If tracking is driven by saliency, the rate of successful target tracking should be



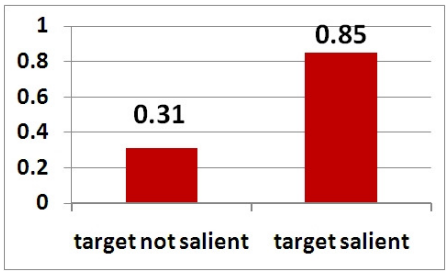
(a) (b) (c)

Figure 7.1: Displays used in the first psychophysics experiment. Subjects were asked to focus on the black fixation square in the center. The disks moved randomly, with velocity indicated by the arrows. (a) In the salient condition, target (shown circled) and distractors differ in color, (b) in the non-salient condition, both have the same color. The display of (b) was obtained by a counterclockwise rotation of that in (a) by 90°. (c) Display used in the "locally salient" condition. The target (shown circled) is *locally salient*, and seven nearest neighbors of the target are of a different color.

much higher for salient than for non-salient displays. This, however, could be explained as a side-effect of bottom-up saliency. Since the target is the only green disk in salient displays, it continuously popped-out. Hence, subjects could be not tracking at all, but simply acquiring the target at every time step. The second version of the experiment ruled out this hypothesis by using a different type of display for the *salient* condition. In this case, the target was a red disk, and its 7 nearest spatial neighbors were green. All other distractors were randomly assigned to either the red or green class. This eliminated the percept of pop-out. As before, the display for the non-salient condition was created by rotation and color switch of the target on the second half of the presentation. The video displays are available online at [4].

7.2.2 Results and Discussion

Figure 7.2 presents the rate of successful tracking in the two versions. In both cases, this rate was much higher in the salient than in the non-salient condition. In the latter, the tracking performance was almost at the chance level of $\frac{1}{3}$, suggesting complete tracking failure. Overall, tracking performance was vastly improved for salient targets even when they did not pop-out. In fact, the similarity of detection rates in

(a) 

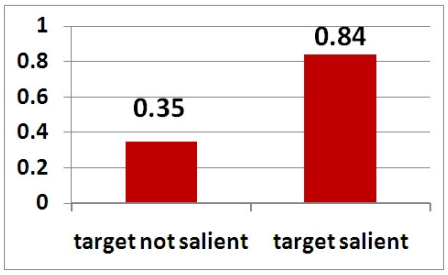
(b) 

Figure 7.2: Tracking success rate when targets are (a) globally salient (pop-out), and (b) locally saliency (do not pop-out).

the two experiments suggests rather than pop-out, it suffices for the target to be locally salient. This is consistent with the hypothesis that tracking is guided by a top-down center-surround saliency mechanism. Since all other parameters were equal under the two conditions, the difference in performance can only be attributed to the *saliency or discriminability* of the target. While this experiment used color as a discriminant cue, the same conclusions apply when other features are salient. For example, studies on multiple object tracking with identical targets and distractors have reported tracking failure when target and distractors are too close to each other [82]. This is easily explained in the discriminant framework: when target and distractors are identical, the target must be spatio-temporally salient (by its trajectory or position) in its neighborhood to be tracked accurately.

7.3 Experiment 2: Tracking performance and saliency as a function of feature contrast

The results of the first experiment show that tracking is related to saliency. While a salient target is tracked reliably, non-salient targets are difficult to track. Experiment 2 aimed to investigate the connection between the two phenomena in greater detail, namely to *quantify* how tracking reliability depends on target saliency. Since saliency is not an independent variable, it can only be controlled indirectly. This is usually done by manipulating *feature contrast* between the target and distractors. It is well known that when the target differs from distractors in terms of color, luminance, orientation or

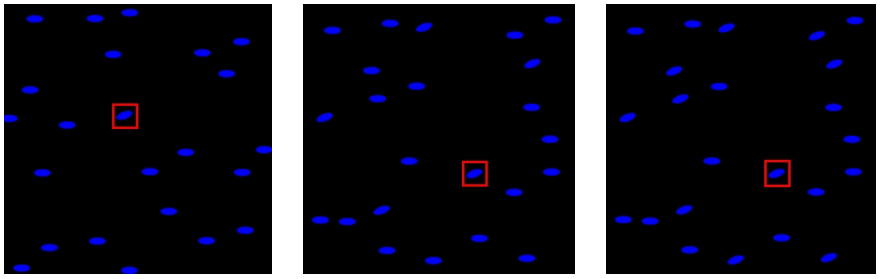
texture it can be perceived as salient [128, 124]. In particular, Nothdurft [126] *quantified* the dependence of saliency on orientation contrast in static displays. His work has shown that perceived target saliency increases with the orientation contrast between target and neighboring distractors. This increase is quite non-linear, exhibiting the threshold and saturation effects shown in Figure 7.2(c), where we present curves of saliency as a function of orientation contrast between target and distractors for three levels of distractor homogeneity. The relationship between tracking reliability and target saliency can thus be characterized by using orientation contrast as a proxy for target saliency, and measuring its effect on tracking performance. If saliency and tracking share common neural mechanisms, the variation of the two with orientation contrast should be identical. In particular, increasing orientation contrast between target and distractors should result in a non-linear increase of tracking reliability, similar to that observed for saliency by Nothdurft.

7.3.1 Method

Participants Twelve subjects (8 male and 4 female) in the age range 21-35 participated in the study.

Stimuli The experimental setting was adapted from the work of Makovski and Jiang [112]. The display on the monitor was of size $26^\circ \times 26^\circ$ (700×700 pixels) and consisted of 23 ellipses, all of color blue, against a black background. Each ellipse had a major axis of $\sim 0.56^\circ$ (15 pixels) and minor axis of $\sim 0.19^\circ$ (5 pixels). The orientation of the ellipses depended on the condition from which the trial was drawn.

Procedure At the start of a trial, one of the ellipses was designated as target (cued with a white bounding box). Subjects were asked to track the target covertly, while fixating on a white square at the center of the screen. On a keystroke, the ellipses started moving and continued to do so for $\sim$ 8-10 sec. At the end of the trial, all ellipses were completely occluded by larger white disks and the subjects asked to click on the disk corresponding to the target. Each subject performed 30 trials under 7 conditions, for a total of 210 trials. No feedback was given on the accuracy of their selection.



(a) (b) (c)

Figure 7.3: Typical frames of stimuli from the three versions in Experiment 2. (a) only the target had orientation different from the distractors (b) 4 of the distractors shared the orientation of the target and (c) 9 distractors in target orientation.

Design The seven conditions corresponded to different levels of orientation contrast between target and distractor ellipses. Distractor orientation, defined by the major axis of the distractor ellipses, was always 0°. Target orientation, determined by the major axis of the target ellipse, was selected from 7 values: 0°, 10°, 20°, 30°, 40°, 60° or 80°. This made orientation contrast equal to the target orientation. To keep all other variables (e.g. distance between items, motion patterns, distance from target to fixation square) identical, a trial was first created for one condition (target orientation 0°). The trials of all other conditions were obtained by applying a transformation to each frame of this video clip. This consisted of an affine transformation of the grid of ellipse centers, followed by the desired change in target orientation.

Versions To study the effect of distractor heterogeneity [126], three versions of the experiment were conducted with different numbers of ellipses in the target orientation. In the first version, only one ellipse (the actual target) was in target orientation. In this case, there was no distractor heterogeneity. In the second version, 18 of the 23 ellipses were in distractor orientation, and the remaining 5 in target orientation. One of the latter was the actual target. Finally, in the third version, 13 ellipses were in distractor and 10 in target orientation, for the largest degree of distractor heterogeneity. Frames from the three different versions, for target orientation of 40°, are shown in Figure 7.3.

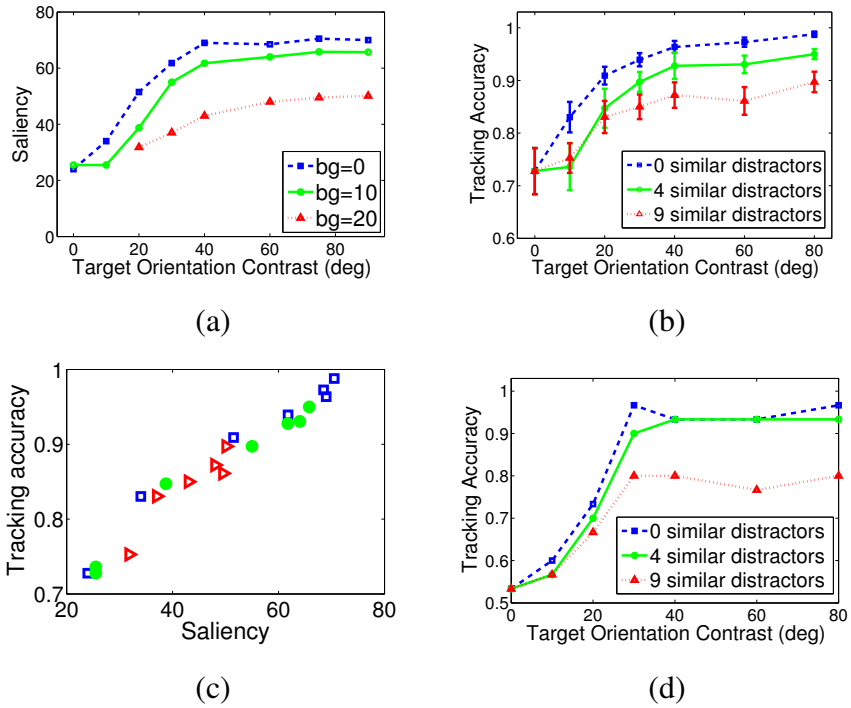


Figure 7.4: (a) saliency vs. orientation contrast (adapted from [126]) (b) human tracking success rate vs. orientation contrast. (c) scatter plot of saliency values from (a) vs tracking accuracy from (b), $r = 0.975$. (d) tracking success rate vs. orientation contrast for the discriminant tracker.

7.3.2 Results and Discussion

As shown in Figure 7.4(a), the curves of tracking accuracy vs. orientation contrast, obtained in all three versions of the experiment, were remarkably similar to the saliency vs. orientation contrast curves of Nothdurft. As is the case for saliency, 1) distinct threshold and saturation effects were observed for tracking, with tracking accuracy saturating when orientation contrast increases beyond 40°, and 2) increased distractor heterogeneity caused a decrease in tracking accuracy. The near perfect correlation ($r = 0.975$) between tracking accuracy and saliency is evident from the scatter plot of Figure 7.4(c). Each point in this plot corresponds to a different combination of heterogeneity and orientation contrast. In summary, tracking has a dependence on orientation contrast remarkably similar to that of saliency.

7.4 Experiment 3: Effect of background on tracking performance

The results of the Experiments 1 and 2 establish a strong connection between saliency and tracking, and provide evidence in favor of the saliency hypothesis. In relating saliency and tracking, the hypothesis proposes that tracking uses center-surround mechanisms to identify salient features that make the target distinct from their background. The involvement of a center-surround mechanism in tracking is consistent with the results of Experiment 2, where the tracking performance is seen to depend on distractor heterogeneity - if the surround were not involved in the tracking process, the performance would not depend on the number of distractors similar to the target in the surround and the three curves of Figure 7.4(b) would be identical.

To test the involvement of a center-surround mechanism in tracking further, we designed Experiment 3. In this experiment the distance between the target and the closest similar distractor (i.e. one with the same orientation as the target) is controlled so that a region of fixed radius around the target is devoid of any similar distractors. By varying this target-similar distractor distance (t_{tsd}), and observing the tracking performance, three possible scenarios can be evaluated :

- (a) a localized surround region is involved in the tracking process: in this case, when t_{tsd} is varied, there should be a distance, which we shall denote as $t_{critical}$, beyond which all similar distractors are outside the surround region relevant for tracking. So for large enough values of t_{tsd}, i.e. $t_{tsd} > t_{critical}$, distractor heterogeneity should not affect tracking performance.

- (b) the entire visual field is involved: if the entire visual field is involved, no such distance, $t_{critical}$, should exist and distractor heterogeneity should affect tracking performance for all values of t_{tsd}.

- (c) no surround region is included in the tracking process: in this case, the success rate of tracking should be identical in all versions regardless of the distractor heterogeneity

As the results of Experiment 2 already showed that conjecture (c) does not hold, Experiment 3 was designed to determine which among conjectures (a) and (b) holds.

7.4.1 Method

Participants 9 subjects (7 male and 2 female) in the age range 21-35 participated in the study.

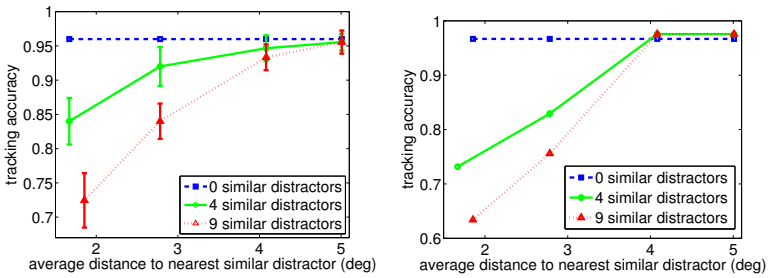
Stimuli and Procedure The experimental setting, stimuli and procedure were identical to those in Experiment 2.

Design The target orientation for all stimuli was fixed at 40°. Two versions of the experiment were conducted with different numbers of ellipses in the target orientation corresponding to two values of distractor heterogeneity. As in Experiment 2, in the first version, 18 of the 23 ellipses were in distractor orientation, and the remaining 5 in target orientation, one of the latter being the actual target. In the second version, 13 ellipses were in distractor and 10 in target orientation. In each version, the stimulus sequence could be in one of four conditions depending on the average value of t_{tsd}, i.e. the average, over all frames in the sequence, of the distance between the target and the nearest similar distractor. In each condition, the sequences were designed such that this quantity was in the range 1.67° to 5.01° (about 45 pixels to 135 pixels).

7.4.2 Results and Discussion

Figure 7.5(a) presents the rate of successful tracking in the two versions as a function of the average distance to nearest similar distractor. Also shown in the figure is the tracking accuracy for the version with no similar distractors at target orientation of 40° from Experiment 2. As there are no distractors similar to the target in this case, a flat line is used to denote the tracking accuracy over all values of the abscissa.

The results show that tracking performance improves as the average distance to nearest similar distractor increases under both versions with non-zero distractor heterogeneity. Further, for large enough value of the distance, tracking accuracy in the two versions are nearly the same as the one with no distractor heterogeneity. This shows that



(a) (b)

Figure 7.5: The effect of background on tracking performance. (a) Tracking accuracy of human subjects for two versions of distractor homogeneities are plotted as a function of the average target-similar distractor distance. Also shown, in blue, is the tracking accuracy for the version with no similar distractors at target orientation of 40° from Experiment 2. (b) model prediction for the same data using the saliency based model.

conjecture (a) holds, i.e. a localized surround region of limited size is involved in the tracking task, and $t_{critical} \approx 4°$. When the identical distractors are kept out of this region, adding more such distractors does not impact tracking performance.

In summary, the results of the human behavior studies show that there is strong evidence for the hypothesis that tracking performance is determined by the saliency of the target, and that tracking and saliency share common neural mechanisms based on center-surround discrimination. In the forthcoming Chapters, we reinforce this evidence by deriving a neurophysiologically plausible network that solves the two tasks. This network is based on the computational framework of discriminant saliency [65], which we briefly review next.

7.5 Acknowledgments

The text of Chapter 7, in full, is based on the material as it appears in: V. Mahadevan, and N. Vasconcelos, "Biological plausibility of the saliency hypothesis for visual tracking", in preparation. The dissertation author was a primary researcher and an author of the cited material.

Chapter 8

Biological plausible model for the tracking

8.1 Introduction

In this section, we study the question of whether saliency and tracking can be implemented with common neural mechanisms. We build on an existing saliency architecture and consider its extension to the tracking problem. This includes the identification of the mechanisms required to extend a saliency network to the tracking problem, and how these mechanisms can be implemented in a biologically plausible manner.

8.2 Discriminant saliency network for tracking

As reviewed in Chapter 2, discriminant saliency equates saliency to optimal decision-making between two classes of visual stimuli, with label $C \in \{0, 1\}$, $C = 1$ for stimuli in a *target* class, and $C = 0$ for stimuli in a *background* class. Saliency is defined in a center-surround manner where, at each location l, the target class is associated with stimuli within a target window $\mathcal{W}_l^1$, and the background class with stimuli in a surrounding background window $\mathcal{W}_l^0$. The saliency of the location l is then equated to the expected accuracy of the target/background classification, given the stimuli in the two windows. The stimuli are not observed directly, but through projection onto a set of

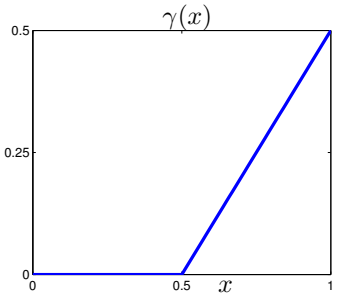


Figure 8.1: The non-linearity used in the saliency computation of (8.2). It thresholds the posterior at value 0.5.

n features, of responses $\mathbf{Z}(l) = (Z_1(l), \ldots, Z_n(l))$. Locations that can be classified with largest expected accuracy are denoted salient. The expected accuracy of classification is measured for each feature, and averaged across features, leading to another version of the overall saliency measure

$$S(l) = \frac{1}{n} \sum_k S_{Z_k}(l) \tag{8.1}$$

$$S_{Z_k}(l) = E_{Z_k(l)}\{\gamma[P_{C(l)|Z_k(l)}(1|z)]\}, \tag{8.2}$$

$$\gamma(x) = \begin{cases} x - \frac{1}{2} & x \geq \frac{1}{2} \\ 0 & x < \frac{1}{2} \end{cases} \tag{8.3}$$

where $\gamma(x)$ is a nonlinearity (shown in Figure 8.1) that thresholds the posterior probability, $P_{C(l)|Z_k(l)}(1|z)$, that the response of the k^{th} feature at l, $Z_k(l)$, was generated by the target class, $C(l) = 1$. Therefore, the saliency measure $S_{Z_k}(l)$ is the expected confidence with which the feature response $Z_k(l)$ belongs to the target class. The nonlinearity $\gamma(x)$ prevents locations declared as not belonging to the target class, by the Bayes decision rule $\left(P_{C(l)|Z_k(l)}(1|z) \leq \frac{1}{2}\right)$, from contributing to the saliency. This tunes the saliency measure to respond only to the presence of target stimuli, not to its absence.

The work of [65] has shown that the computation of (8.2) can be mapped to the standard neurophysiological model of area V1, when the features are Gabor-like and stimulus consists of static natural images. We briefly review the relevant findings here.

8.2.1 Mapping saliency computation to area V1

When the feature Z_k extracted from the target/background region is of bandpass nature, as is common in biological vision, the feature response follows a generalized Gaussian distribution (GGD) of scale *scale* α and *shape* β [80],

$$P_{Z_k}(z;\alpha,\beta) = \frac{\beta}{2\alpha\Gamma(1/\beta)} \exp\left\{-\left(\frac{|z|}{\alpha}\right)^{\beta}\right\}, \tag{8.4}$$

where $\Gamma(a) = \int_0^\infty e^{-t}t^{a-1}\mathrm{d}t$, $t > 0$.

In the discriminant saliency formulation for this case, the posterior probability of the target class is given by [65],

$$P_{C(l)|Z_k(l)}(1|z) = \sigma[g(z)], \tag{8.5}$$

where $\sigma(z) = (1+e^{-z})^{-1}$ is a sigmoid, and $g(z)$ the log-likelihood ratio between the two class-conditional GGDs,

$$g(z) = \log\frac{P_{Z_k(l)|C(l)}(z|1)}{P_{Z_k(l)|C(l)}(z|0)} = \frac{|z|^{\beta_0}}{\alpha_0^{\beta_0}} - \frac{|z|^{\beta_1}}{\alpha_1^{\beta_1}} + T, \qquad T = \log\frac{\alpha_0\beta_1\pi_1\Gamma(1/\beta_0)}{\alpha_1\beta_0\pi_0\Gamma(1/\beta_1)}. \tag{8.6}$$

The scale parameters for class 0 and class 1, α_0 and α_1 respectively, are estimated by the maximum a posteriori probability (MAP) method, with conjugate (Gamma) priors, according to

$$\alpha_c^{\beta_c} = \frac{1}{\kappa_c}\left(\nu_c + \sum_{l'\in\mathcal{W}_l^c} |z(l')|^{\beta_c}\right) \forall c \in \{0,1\}. \tag{8.7}$$

The shape parameters β_c, $\forall c \in \{0,1\}$ are quite consistent across image classes, and can be set to the value $\beta_c = 1$, which provides a good fit to natural images. Finally, replacing expectations by empirical averages, the bottom-up saliency for the feature Z_k can be written as:

$$\begin{aligned} S_{Z_k}(l) &= E_{Z_k(l)}\{\gamma[P_{C(l)|Z_k(l)}(1|z)]\} && (8.8)\\ &\approx \frac{1}{|\mathcal{W}_l|}\sum_{l'\in\mathcal{W}_l}\gamma\{\sigma(g[z(l')])\}, \qquad \mathcal{W}_l = \mathcal{W}_l^0 \cup \mathcal{W}_l^1 && (8.9)\end{aligned}$$

The computations of (8.6)-(8.9) can be mapped into a neural network that replicates the standard neurophysiological model of V1 neurons [30]. In particular, combining (8.6)

and (8.7), it follows that $g[z(l')]$ computes a differential divisive normalization of the feature response $z(l')$ by the responses of the feature Z_k in the neighborhoods $\mathcal{W}_l^c$. Hence, $\sigma(g[z(l')])$ implements the computations of V1 simple cells under the standard model: a sequence of linear filtering, rectification, divisive normalization, and output saturation. (8.9) then pools the outputs of simple cells in $\mathcal{W}_l$ after passing them through the non-linearity $\gamma(x)$. These are the computations performed by V1 complex cells, under the standard model. The mapping is illustrated in Figure 8.3.

In the case of motion processing and tracking, the feature set needs to include velocity tuned spatio-temporal features. These are processed in area MT of the visual pathway and it is known that neurons in area MT receive input from V1 complex cells [23]. Therefore, computational models for MT are usually constructed by combining the outputs of V1 complex cell afferents [156, 146]. As simple and complex cells in V1 are well modeled by the saliency network discussed above [65], a model for MT that can process velocity tuned features can be constructed by substituting the saliency network model in place of the standard model for V1 in the approach of Simoncelli and Heeger [156].

We next show how this saliency based model for MT can be constructed and also show that the output of such a model computes saliency of the velocity tuned features.

8.3 Model for an MT neuron and saliency for velocity tuned features

In the saliency based V1 model, as in [156], the linear filtering in the V1 simple cell stage is achieved using spatio-temporal Gabor features $Z_k(l)$ that are sensitive to motion. The output of the model V1 complex cell then computes bottom-up saliency for the corresponding spatio-temporal feature using (8.9), making it selective to the component of stimulus velocity orthogonal to the spatial orientation of the feature, but not truly direction selective. By combining the responses of a set of such features, a unit responding to velocity in a specific direction can be constructed using the approach of Heeger [73].

8.3.1 Computations of weights for the MT model

The weight w_{jk} used in (8.22) to compute the response of the k^{th} velocity tuned feature Y_k, from the j^{th} spatio-temporal Gabor feature Z_j can be evaluated using the Gabor energy approach of [73]. The feature response corresponding to Z_j is the output of the visual stimulus passed through a sine-phase three dimensional Gabor filter $\mathsf{g}_j(x, y, t)$ of the form:

$$\mathsf{g}_j(x, y, t) = \frac{1}{\sqrt{2}\pi^{\frac{3}{2}}\sigma_x\sigma_y\sigma_t} \sin(2\pi\omega_{x_j}x + 2\pi\omega_{y_j}y + 2\pi\omega_{t_j}t) e^{-\left(\frac{x^2}{2\sigma_x^2}+\frac{y^2}{2\sigma_y^2}+\frac{t^2}{2\sigma_t^2}\right)} \tag{8.10}$$

where $\omega_{x_j}, \omega_{y_j}$ is the spatial frequency, ω_{t_j} the temporal frequency, and the 3D Gaussian envelope has standard deviations $\sigma_x, \sigma_y, \sigma_t$.

The Fourier transform of this Gabor filter is given by [72]:

$$\begin{aligned}\mathcal{F}_{\mathsf{g}_j}(\omega_x, \omega_y, \omega_t) = \frac{i}{2}\{&e^{-2\pi^2\sigma_x^2(\omega_x-\omega_{x_j})^2} - e^{-2\pi^2\sigma_x^2(\omega_x+\omega_{x_j})^2} + \\ &e^{-2\pi^2\sigma_y^2(\omega_y-\omega_{y_j})^2} - e^{-2\pi^2\sigma_y^2(\omega_y+\omega_{y_j})^2} + \\ &e^{-2\pi^2\sigma_t^2(\omega_t-\omega_{t_j})^2} - e^{-2\pi^2\sigma_t^2(\omega_t+\omega_{t_j})^2}\}\end{aligned} \tag{8.11}$$

The energy of feature responses to the filter of (8.10) can be computed if its power spectral density (PSD) is known. As the filter is separable in its three dimensions, we first illustrate the computation of PSD of a 1-D Gabor filter from its Fourier response [72]:

$$\mathsf{g}(x) = \frac{1}{\sqrt{2}\pi^{\frac{3}{2}}\sigma_x} \sin(2\pi\omega_{x_j}x) e^{-\frac{x^2}{2\sigma_x^2}} \tag{8.12}$$

$$\mathcal{F}_{\mathsf{g}}(\omega_x) = \frac{i}{2}\{e^{-2\pi^2\sigma_x^2(\omega_x-\omega_{x_j})^2} - e^{-2\pi^2\sigma_x^2(\omega_x+\omega_{x_j})^2}\} \tag{8.13}$$

The PSD of the filter is then

$$\begin{aligned}|\mathcal{F}_{\mathsf{g}}(\omega_x)|^2 &= \frac{1}{4}\{e^{-4\pi^2\sigma_x^2(\omega_x-\omega_{x_j})^2} + e^{-4\pi^2\sigma_x^2(\omega_x+\omega_{x_j})^2} + 2e^{-2\pi^2\sigma_x^2(\omega_x-\omega_{x_j})^2-2\pi^2\sigma_x^2(\omega_x+\omega_{x_j})^2}\} \\ &= \frac{1}{4}\{e^{-4\pi^2\sigma_x^2(\omega_x-\omega_{x_j})^2} + e^{-4\pi^2\sigma_x^2(\omega_x+\omega_{x_j})^2} + e^{-4\pi^2\sigma_x^2(\omega_x^2+\omega_{x_j}^2)}\} \qquad (8.14) \\ &\approx \frac{1}{4}\{e^{-4\pi^2\sigma_x^2(\omega_x-\omega_{x_j})^2} + e^{-4\pi^2\sigma_x^2(\omega_x+\omega_{x_j})^2}\} \qquad (8.15)\end{aligned}$$

where the third term,

$$D(\omega_x) = e^{-4\pi^2\sigma_x^2(\omega_x^2+\omega_{x_j}^2)}, \tag{8.16}$$

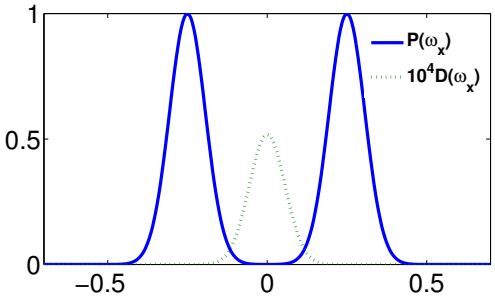


Figure 8.2: Approximation of the PSD of a sine phase Gabor filter in 1D. The thick blue curve shows the quantity in (8.15) for typical values of σ_x and ω_{x_j}, and the dotted curve shows 10^4 times the difference between the quantities in (8.15) and (8.14)

can be ignored because it is upper-bounded by $e^{-4\pi^2\sigma_x^2\omega_{x_j}^2}$, a quantity that is much smaller than 1. This is illustrated in Figure 8.2.

Similarly, the PSD of the 3D Gabor filter can be given by,

$$\begin{aligned} |\mathcal{F}_{\mathsf{g}_j}(\omega_x,\omega_y,\omega_t)|^2 &\approx \frac{1}{4}\{e^{-4\pi^2\sigma_x^2(\omega_x-\omega_{x_j})^2}+e^{-4\pi^2\sigma_x^2(\omega_x+\omega_{x_j})^2}+ \\ &\quad \frac{1}{4}\{e^{-4\pi^2\sigma_y^2(\omega_y-\omega_{y_j})^2}+e^{-4\pi^2\sigma_y^2(\omega_y+\omega_{y_j})^2}\}+ \\ &\quad \frac{1}{4}\{e^{-4\pi^2\sigma_t^2(\omega_t-\omega_{t_j})^2}+e^{-4\pi^2\sigma_t^2(\omega_t+\omega_{t_j})^2}\} \qquad (8.17) \\ &= P_j(\omega_x,\omega_y,\omega_t) \qquad (8.18) \end{aligned}$$

For a sinusoidal grating moving with a given velocity, $\bar{v}_k = v_{kx}\hat{e}_x + v_{ky}\hat{e}_y$, its energy in the frequency domain is contained in a plane defined by [181]:

$$v_{kx}\omega_x + v_{ky}\omega_y - \omega_t = 0 \qquad (8.19)$$

To construct a unit tuned to velocity $\bar{v}_k$ we compute a weighted combination of the outputs of a set of 3D Gabor filters following the approach of [156]. The weight assigned to each Gabor filter in the set is in proportion to the energy contained in the intersection between the PSD of the filter and the plane corresponding to $\bar{v}_k$. This can be computed as:

$$w_{jk}(\omega_{x_j},\omega_{y_j},\omega_{t_j},\bar{v}_k) \propto \qquad (8.20)$$

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} P_j(\omega_x,\omega_y,\omega_t)\, d\omega_x d\omega_y d\omega_t\Big|_{v_{kx}\omega_x+v_{ky}\omega_y-\omega_t=0} \qquad (8.21)$$

$$= \frac{1}{2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\left(e^{-2\pi^2\sigma_x^2(\omega_x-\omega_{x_j})^2}+e^{-2\pi^2\sigma_y^2(\omega_y-\omega_{y_j})^2}+e^{-2\pi^2\sigma_t^2(v_{kx}\omega_x+v_{ky}\omega_y-\omega_{t_j})^2}\right)d\omega_x d\omega_y$$

where the second step follows due to the symmetry between the two lobes of the PSD. The integral can be evaluated using the procedure outlined in [73].

In this work we use a total of 12 spatio-temporal filters, each with a center frequency $(\omega_{x_j}, \omega_{y_j}, \omega_{t_j})$, $j = 0 \ldots 11$. We consider 12 neurons each tuned to motion with constant speed in one of 12 different directions spread uniformly in $(0°, 360°)$, corresponding to velocities $\bar{v}_k$, $k = 0 \ldots 11$.

w_{jk} is then the weight assigned to the j^{th} filter for computing motion in the k^{th} direction.

Let $S_{Z_j}(l)$ be the model output at the V1 stage for the j^{th} spatio-temporal feature, Z_j, using (8.9). Then the output of a unit tuned to velocity $\bar{v}_k$, corresponding to feature Y_k, is given by:

$$S_k(l) = \sum_j w_{jk}(\bar{v}_k) S_{Z_j}(l) \tag{8.22}$$

where the weights $w_{jk}(\bar{v}_k)$ are computed using the approach of [73] (see Appendix 8.3.1).

This is equivalent to computing the saliency of *a complex spatio-temporal feature* Y_k, designed to respond to stimuli moving with a specific velocity $\bar{v}_k$, from a combination of simple spatio-temporal Gabor features:

$$Y_k(l) = \sum_j w_{jk}(\bar{v}_k) Z_j(l) \tag{8.23}$$

The expression for saliency of Y_k in (8.22) ignores the effect of dependencies between the simple spatio-temporal features, Z_j, which has been shown to be a reasonable approximation when the features are bandpass [170], as is the case for the Gabor features used in the construction of the model.

Finally, as in [156], the output of the unit is divisively normalized by the responses of other units:

$$S_k^{td}(l) = \frac{S_k(l)}{\sum_j S_j(l)} \tag{8.24}$$

Each model unit corresponding to a feature Y_k, tuned to velocity $\bar{v}_k$, can be thought of as being equivalent to a neuron in MT. The computed model output is analogous to the neuron's firing rate and responds maximally when the input stimulus moves

with $\bar{v}_k$, the velocity to which the feature is tuned. This velocity is referred to as the *preferred velocity* of the model neuron. The interpretation, given by (8.24), is that velocity selective tuning is a reflection of the neuron's *function* as a detector of salient motion configurations in a particular velocity channel. The resulting network is illustrated in Figure 8.3.

The model for MT proposed above is built from neurophysiologically plausible units, using the same architecture as [156]. So arguments for biological plausibility of the V1 stage [65] and of the architecture [156] extend to the proposed MT model. Further, by using a center-surround architecture in the V1 stage, the model accounts for the surround antagonism observed in MT neurons [160, 22], but not modeled by [156].

8.4 Neurophysiologically plausible feature selection

A key component of the saliency tracker of [110] is a feature selection procedure that continuously adapts the saliency measure of (8.2) to the target. The basic idea is to select, at each time step, the features in $\mathbf{Y}(l) = (Y_1(l), \ldots, Y_m(l))$ that best discriminate between target (center) and background. This changes the saliency from a bottom-up identification of locations where center and surround differ, to a top-down identification of locations containing the target in the center and background in the surround. However, the procedure of [110] (based on feature ranking) is not biologically plausible. To derive a biologically plausible feature selection mechanism, we replace the saliency measure of (8.1) with a feature-weighted extension

$$S(l) = \sum_k \alpha_k S_k(l), \qquad \sum_k \alpha_k = 1 \tag{8.25}$$

where α_k is the weight given to the saliency of the k^{th} feature channel. To determine these weights we need a biological measure of *feature saliency*. For this, we associate a binary variable F_k with each feature Y_k, such that $F_k = 1$ if and only if Y_k is the *most salient* feature of the target. We then assume that, given the knowledge of which feature is most salient, target presence at location l is independent of the remaining feature responses

$$P_{C(l)|\mathbf{Y}(l),F_k}(1|\mathbf{y}, 1) = 2\gamma[P_{C(l)|Y_k(l)}(1|y)], \tag{8.26}$$

where $\gamma(x)$ is the non-linearity of (8.2). This reflects a conservative strategy, where features cannot be considered salient unless they are individually discriminant for target presence.

Given the location l^* where the target has been detected, the posterior probability of feature saliency can then be computed by Bayes rule

$$P_{F_k|C(l^*)}(1|1) = \frac{P_{C(l^*)|F_k}(1|1)P_{F_k}(1)}{\sum_j P_{C(l^*)|F_j}(1|1)P_{F_j}(1)} \tag{8.27}$$

where

$$\begin{aligned} P_{C(l^*)|F_k}(1|1) &= \int P_{C(l^*)|\mathbf{Y}(l^*),F_k}(1|\mathbf{y},1)P_{\mathbf{Y}(l^*)|F_k}(\mathbf{y}|1)d\mathbf{y} && (8.28)\\ &= \int 2\gamma[P_{C(l^*)|Y_k(l^*)}(1|y)]P_{Y_k(l^*)}(y)dy \text{ (using (8.26))} && (8.29)\\ &= 2E_{Y_k(l^*)}\{\gamma[P_{C(l^*)|Y_k(l^*)}(1|y)]\} = 2S_k(l^*), && (8.30) \end{aligned}$$

and the last equality follows from (8.2). Hence,

$$P_{F_k|C(l^*)}(1|1) = \frac{S_k(l^*)P_{F_k}(1)}{\sum_j S_j(l^*)P_{F_j}(1)}. \tag{8.31}$$

These posterior probabilities serve as weights α_k in (8.25). Under reasonable assumptions of persistence of the dominant features in the target, this analysis can be extended over time, by denoting the state of F_k and l^* at time t by F_k^t and l_t^*, respectively, and the sequence of target locations till time t by $\mathbf{l}_t^* = (l_t^*, l_{t-\tau}^* \ldots l_0^*)$. Using Bayes rule, the posterior probability of feature F_k being the most salient feature can be written as,

$$P_{F_k^t|C(\mathbf{l}_t^*)}(1|\mathbf{1}) = P_{F_k^t|C(l_t^*),C(\mathbf{l}_{t-\tau}^*)}(1|1,\mathbf{1}) \propto P_{C(l_t^*)|F_k^t,C(\mathbf{l}_{t-\tau}^*)}(1|1,\mathbf{1})P_{F_k^t|C(\mathbf{l}_{t-\tau}^*)}(1|\mathbf{1}) \tag{8.32}$$

We do not assume an explicit motion model or motion extrapolation. However it is reasonable to assume that the probability of finding the target is uniformly distributed in a small neighborhood around $l_{t-\tau}^*$, and if the velocity of the target is not too high, l_t^* is in this neighborhood. We can then write,

$$P_{C(l_t^*)|F_k^t,C(\mathbf{l}_{t-\tau}^*)}(1|\mathbf{1}) \propto P_{C(l_t^*)|F_k^t}(1|1) \tag{8.33}$$

Using this in (8.32), we get,

$$P_{F_k^t|C(\mathbf{l}_t^*)}(1|\mathbf{1}) \propto P_{C(l_t^*)|F_k^t}(1|1)P_{F_k^t|C(\mathbf{l}_{t-\tau}^*)}(1|\mathbf{1}) \tag{8.34}$$

$$\propto P_{C(l_t^*)|F_k^t}(1|1)\sum_i P_{F_k^t,F_i^{t-\tau}|C(\mathbf{l}_{t-\tau}^*)}(1,1|\mathbf{1}) \tag{8.35}$$

$$\propto P_{C(l_t^*)|F_k^t}(1|1)\sum_i P_{F_k^t|F_i^{t-\tau},C(\mathbf{l}_{t-\tau}^*)}(1|\mathbf{1})P_{F_i^{t-\tau}|C(\mathbf{l}_{t-\tau}^*)}(1|\mathbf{1}) \tag{8.36}$$

The probabilities $P_{F_k^t|F_i^{t-\tau},C(\mathbf{l}_{t-\tau}^*)}(1|1,\mathbf{1})$ encode the likelihood of transition from state i to state k. Since dominant features of the target tend to stay dominant for some time in the neighborhood of the last known position of the target, we assume $P_{F_k^t|F_i^{t-\tau},C(\mathbf{l}_{t-\tau}^*)}(1|1,\mathbf{1}) = 1$ if $i = k$ and null otherwise (this is the likelihood of transition only using the information from the previous time step, it does not preclude new features from being selected if they become salient at t). Using this, and (8.30), in (8.36) we get the recursion,

$$P_{F_k^t|C(\mathbf{l}_t^*)}(1|\mathbf{1}) = \frac{S_k(l_t^*)P_{F_k^{t-\tau}|C(\mathbf{l}_{t-\tau}^*)}(1|\mathbf{1})}{\sum_j S_j(l_t^*)P_{F_j^{t-\tau}|C(\mathbf{l}_{t-\tau}^*)}(1|\mathbf{1})}. \tag{8.37}$$

Hence, the posterior probability of feature k being the most salient, at time t, is computed by divisively normalizing a modulated version of the bottom-up saliency of the feature response at the target location, by those of the remaining features. The saliency of each feature is modulated by the posterior probability of the feature being the most salient at time $t - \tau$. The posterior at time $t - \tau$ is fed back with a delay, to become the prior at time t. This *enhances the most salient features, suppressing the non-salient ones,* and is equivalent to applying a soft-thresholding to select only the dominant features.

Comparing (8.24) with (8.32), it can be seen that the output of the model for an MT neuron when is there no top-down feedback, given by (8.24), simply corresponds to the posterior probability of Z_k being the most salient feature when the prior feature probabilities are equal, i.e.

$$P_{F_k}(1) = \frac{1}{K} \tag{8.38}$$

This feature selection mechanism involving selective enhancement and suppression of features, and operating on the output of the MT stage bears a close resemblance to the phenomenon of feature-based attention [116]. Infact, the proposed approach to

feature selection has similarities with previous models of feature-based attention, which rely on a Bayesian formulation and include divisive normalization [142, 144, 102, 38].

8.4.1 Neurophysiological plausibility of feature selection

Neurophysiological studies have found evidence for the origin of attentional modulation in three distinct areas of the brain : the lateral intraparietal lobe (LIP) of the post-parietal cortex (PPC), the frontal-eye field (FEF) and the superior colliculus [187]. Among these, the LIP in particular is thought to compute a priority map that combines both bottom-up inputs and top-down signals, and the peak of this map response is used to guide visual attention [16]. Further, neuroanatomical studies have shown that LIP has cortico-cortical connections to area MT [103], and attentional control is thought to be fed-back from LIP to MT [147]. Therefore, area LIP is a plausible candidate for the region where feature selection is performed in our model. We next describe a mechanism by which LIP can achieve this.

The feature saliency maps from each feature from area MT are fed forward to the LIP which sums them up to form the final saliency map. As noted in [16], LIP can also include top-down input in forming the saliency map. But in the case of tracking, we assume that top-down modulation results only due to the salient attributes of the target (e.g features and position) seen in the previous time instances.

Once the saliency map has been computed, a maximum detector possibly based on integration using a Gaussian window, is applied to identify the peak of the saliency map. Spatial attention then shifts to this location, and feature weights are computed based on the features at the attended location, assigning higher weights to features that are present and attentuating those that are absent. These weights are then fed back to the MT area as feature-based attentional control. The delay in finding the peak of the saliency map and feeding back the attentional signal could account for the observed latency of $\approx$ 60ms between attentional modulation in areas LIP and MT [76]. In addition to feature weights, a retinotopic spatial attentional control is also fed back to areas MT and V1 to supress the regions that are not near the attended location.

In the next section, we put together the units discussed above to construct a neurophysiologically plausible version of the discriminant tracking algorithm of [110].

8.5 Neurophysiologically plausible discriminant tracker

A neurophysiologically plausible version of the discriminant tracker of [110] can be constructed with the discriminant saliency measure of (8.2), and the feature selection mechanism of (8.37). As in [110], in the absence of top-level information regarding the target, initialization can be treated as discrimination between the visual stimulus contained in a pair of *center* (target) and *surround* windows, at every location of the visual field. In this case, there is no explicit top-down guidance about the object to recognize, and the saliency of location l is measured by the saliency of *all* unmodulated feature responses. This consists of using the bottom-up saliency measure of (8.25) with $\alpha_k = P_{F_k^0}(1)$, where $P_{F_k^0}(1)$ is a uniform prior for feature selection, at time $t = 0$. The outputs of all features or neurons are then summed with equal weights to produce a final saliency map at the LIP. The peak of this map represents the location which is most distinct from its surround, based on the responses of the motion sensitive spatio-temporal features. Spatial attention then is shifted to the peak of this map.

Once the initial target location is attended, the feature selection mechanism modulates the saliency response of the individual feature channels, using the weights of (8.37). The final saliency value at that location also becomes the normalizing constant for the divisive normalization of (8.37). These feature weights are fed back to MT neurons, where each feature map is enhanced or attenuated depending on the corresponding feature weight given by (8.39). This *enhances the features that are salient for target detection, and suppresses the non-salient ones*. The LIP also feds back the retinotopic weight map corresponding to spatial attention, causing a suppression of feature responses in all areas other than a neighborhood of the current locus of attention.

Using these new feature weights and spatial weights, the modulated feature maps are then fed forward to LIP, where the updated saliency map is computed by simple summation. The top-down saliency of location l at time $t + \tau$ is then given by

$$S^{td}(l) \quad = \quad \sum_j S_j^{td}(l) = \sum_j S_j(l) P_{F_j^t|C(\mathbf{l}_t^*)}(1|\mathbf{1}). \tag{8.39}$$

where $S_j(l)$ is the modulated saliency response of the j^{th} feature.

The peak of this saliency map is again computed and spatial attention is shifted to that location at time $t + \tau_2$. As attentional signals have downweighted all but a neigh-

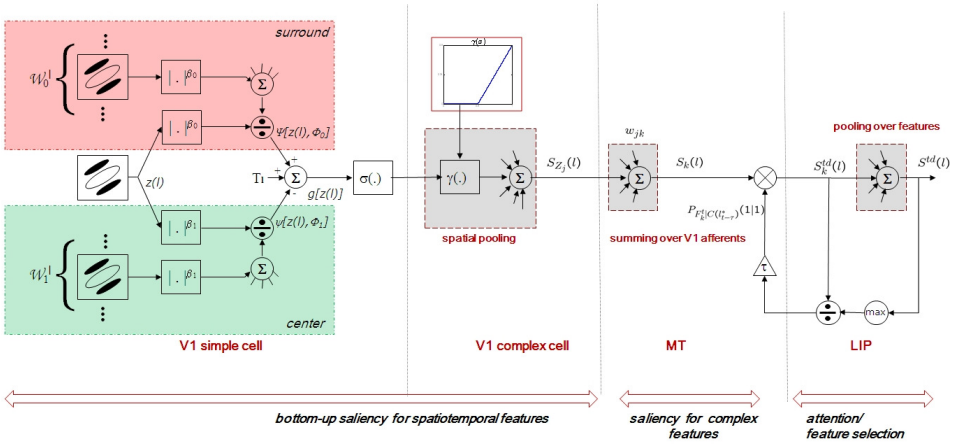


Figure 8.3: The network for tracking using feature selection. The discriminant saliency network of [65] is used to construct a model for an MT neuron. Feature selection, performed possibly in area LIP and fed-back to MT, is achieved by the modulation of the response of each feature channel by its saliency value after divisive normalization across features.

borhood of the last known target location l_t^*, and the feature-based attentional control has downweighted all but the features present in the target and discriminative with respect to the background, the peak of the new saliency map corresponds to the new position of the target. The process is iterated, so as to track the target over time, as in [110]. The entire tracking network is shown in Figure 8.3. The computation, in V1, of $S_j^{bu}(l)$ is implemented with the bottom-up network of [65]. V1 outputs are then linearly combined with weights w_{jk} (which are described in supplement [6]) to obtain the MT responses $S_k(l)$. The remaining operations, possibly in LIP, compute the probabilities of (8.37) and the top-down saliency map of (8.39).

8.6 Discussion

The saliency hypothesis for tracking has been shown to be neurophysiologically plausible, through construction of a tracking model that can be implemented with widely accepted models of cortical computation. Specifically, we have constructed a tracking model based on MT neurons and shown that saliency based tracking can be implemented with a feature selection mechanism akin to the well known phenomenon of

feature based attention in MT. In the next section we validate the biological plausibility of the top-down saliency network by comparing its predictions to psychophysics and neurophysiological data.

8.7 Acknowledgments

The text of Chapter 8, in full, is based on the material as it appears in: V. Mahadevan, and N. Vasconcelos, "Biological plausibility of the saliency hypothesis for visual tracking", in preparation. The dissertation author was a primary researcher and an author of the cited material.

Chapter 9

Model validation on psychophysics and neurophysiological Data

9.1 Introduction

In this Chapter we discuss the validation of the biological plausibility of the top-down saliency network of Figure 8.3 introduced in Chapter 8 by evaluating the tracking reliability of the extended network. This is done by comparing its performance to the human psychophysics of Chapter 7 and the behavior of its units to neurophysiological recordings.

9.2 Model Prediction for Human Behavior Experiments

Experiment 1 Model Prediction

We applied the network to the sequences used in the psychophysics experiment of Section 7.1. Representative frames of the result of tracking on the displays of the experiment are shown in Figure 9.1. The videos are available from [4]. The model replicates the trend observed in both psychophysics experiments, accurately tracking the target in the salient conditions, and losing track in the non-salient condition.

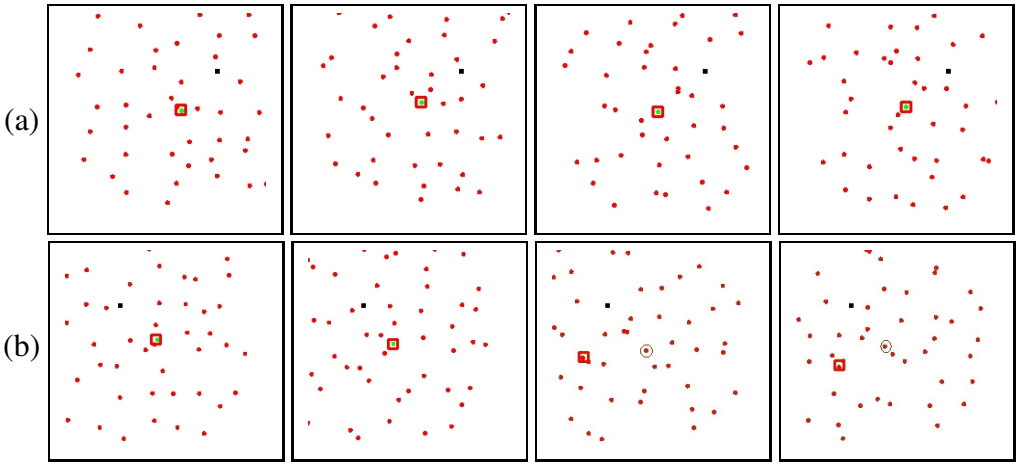


Figure 9.1: Tracking results for a) "salient" and b) "non-salient" conditions. Target detections by DST are marked with a thick red box. The target is tracked through all frames in (a), while tracking fails in (b) after target color changes from green to red. The actual target is shown circled. In both, only a portion of the display is shown.

Experiment 2 and 3 Model Predictions

The results of applying the top-down saliency network model of 8.5 to the stimuli in Experiments 2 and 3 are shown in Figures 7.4(b) and 7.5(b) respectively. It is seen that the model predictions accurately match the trend observed in all three versions of the Experiment 2. The model also predicts the effect of background seen in Experiment 3. The ability of the model to predict these human behavior trends reinforces its biological plausibility.

9.2.1 Comparison to human behavior data on tracking targets with distinct features

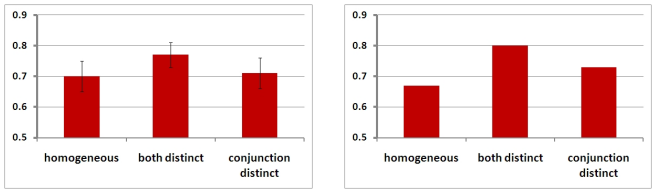
The results of the two experiments described above show a strong relationship between saliency and tracking and provide evidence for a top-down mechanism. The saliency hypothesis proposes a feature-based strategy for this top-down mechanism.An alternative hypothesis to explain this observation can be that an object representation of the target is learned and stored in memory and used subsequently to track the target. This is similar to the tracking-by-recognition paradigm in the computer vision literature [12, 13], where an object detector is trained at the current location of the target, and applied

in the next frame to detect the best location of the target.

Evidence for a feature-based strategy, as opposed to an object-based mechanism, for tracking can be found in recent psychophysics studies that have analyzed tracking performance in MOT with unique targets. Horowitz et al. [78] found that tracking performance improved when the objects were unique, but it was observed that object identities were not retained during the tracking task. In particular, it was found that if observers do recover lost targets during the course of a trial, they appear to recover only their locations, not their identities. This demonstrates that complete object information may not be stored while tracking, and argues against a object-recognition based model for tracking. In a related study, Makovksi and Jiang [112] explored the effect of target uniqueness further and showed that the enhancement in tracking performance for unique targets is feature-based. We describe their study below and illustrate how the results are consistent with the saliency hypothesis.

The experimental set-up of the study in [112] is similar to that of Experiment 2. The stimulus consisted of eight objects, four targets, and the rest distractors. In a typical trial, the targets and distractors moved independently for several seconds and subjects were asked to identify the targets at the end of the trial. Each trial could be from one of three conditions - "homogeneous", "unique" and "conjunction-distinct". In the "homogeneous" condition, all eight objects were identical. In the "unique" condition, the targets were of a unique color and orientation, and no distractor shared either of these features. A frame from each of these conditions is shown in Figure 9.4(a) and (b) respectively. In the "conjunction-distinct" condition (Figure 9.4(c)), the targets were unique when both features, i.e. color and orientation, were considered together. For each target, there was one distractor sharing a feature with that target. For instance, in (Figure 9.4(c)), the target shown inside a red box, is distinct from other objects by virtue of being the only object that is *blue and oriented horizontally*. However, there is a distractor of the same color, but with different orientation, and another distractor with the same orientation but in green.

The goal of the study was to investigate whether making the targets distinct conferred an advantage over the "homogeneous" condition in the tracking task. The results obtained in the study are shown in Figure 9.2(a). It was seen that while tracking



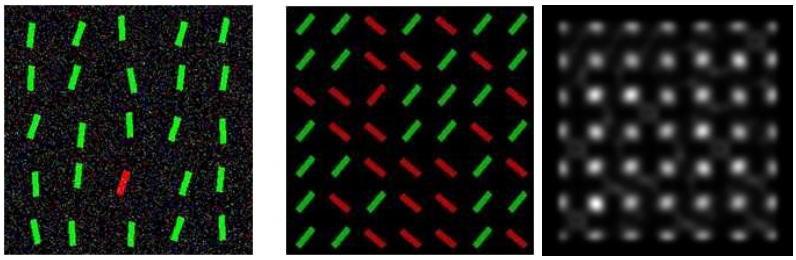
(a) (b)

Figure 9.2: Target tracking rate (a) from [112] (b) results obtained using the saliency based tracking model.

performance was enhanced in the "unique" condition, no improvement was seen in the "conjunction-distinct" condition. If color and orientation were *both* integrated with the location of the target, it would be perceived as being unique and the performance should have been comparable to the "unique" condition. However, this was not seen making it clear that the target features were not completely bound to their moving locations. Infact, the authors suggest that feature memory acts independently of the location processing.

The observations made by Makovski and Jiang are entirely compatible with properties of feature-based attention and the results of the study can be explained by the saliency hypothesis for tracking. To test this hypothesis, the saliency based tracker of Section 8.5 was extended to multi-object tracking by assuming that upto 4 independent objects can be tracked without any resource constraints. Around 100 clips for each of the three conditions - "homogeneous", "unique" and "conjunction distinct", were generated using Psychtoolbox [24] code provided by the authors [112]. For each clip, four independent saliency-based trackers were initialized and allowed to track till the end of the clip. Tracking was considered a success when all four target items were successfully tracked. The performance of the saliency model for each of the three conditions is shown in Figure 9.2(b). It is seen that the model replicates the trend reported in the original study (Figure 9.2(a)). The tracking rate in the "unique" condition is higher than in the "homogeneous" and "conjunction-distinct" conditions.

To understand the results in the context of the saliency hypothesis, we start by considering a well-studied phenomenon involving conjunctions of features in static displays of the type illustrated in Figure 9.3(b) [167]. The bar in the 3^{rd} row and 3^{rd} column of the display is the only item that is both red in color and tilted right. However, it does not pop-out, unlike the clear perception of pop-out in Figure 9.3(a) where the red bar is



(a) (b)

Figure 9.3: Saliency and feature conjunctions (from [60]) (a) the red bar is salient among the green distractors and "pops-out" (b) the bar in the 3^{rd} row and 3^{rd} column is different from other bars when both color and orientation are considered. There is no perception of pop-out. (c) the saliency map of (b) obtained using the discriminant center-surround approach of Section 8.2. The saliency value for the bar is not significantly different from other bars.

unique in terms of just one feature, i.e. color. This has been shown to be replicated using the discriminant center surround saliency model of Section 8.2 because it ignores the effect of dependencies between feature responses in discriminating between the target and background classes [60]. This leads to an overall saliency measure which is simply a sum of the saliency for each feature ((8.1)). As the target is not salient in either feature channel, color or orientation, its overall saliency value is not significantly different from other bars in the display. This is illustrated in Figure 9.3(c).

The feature selection strategy of the saliency hypothesis is based on the same principle of ignoring feature dependencies and combining the top-down tuned saliency maps of different features using (8.39). The top-down tuning enhances features that make the target salient - those features are predominantly present in the target and absent in the surround. However, as there are distractors sharing features with the target, these distractor locations are also enhanced in the corresponding saliency map, leading to multiple peaks in the overall saliency map. Hence, the target can be confused with the distractor leading to lower performance compared to the "unique" condition. This is illustrated using representative saliency maps from the three conditions shown in the bottom row of Figure 9.4. The saliency map for the "unique" condition has a single dominant saliency peak, while those of the other two conditions have several dominant locations. These dominant distractors could lead to errors when searching for the best

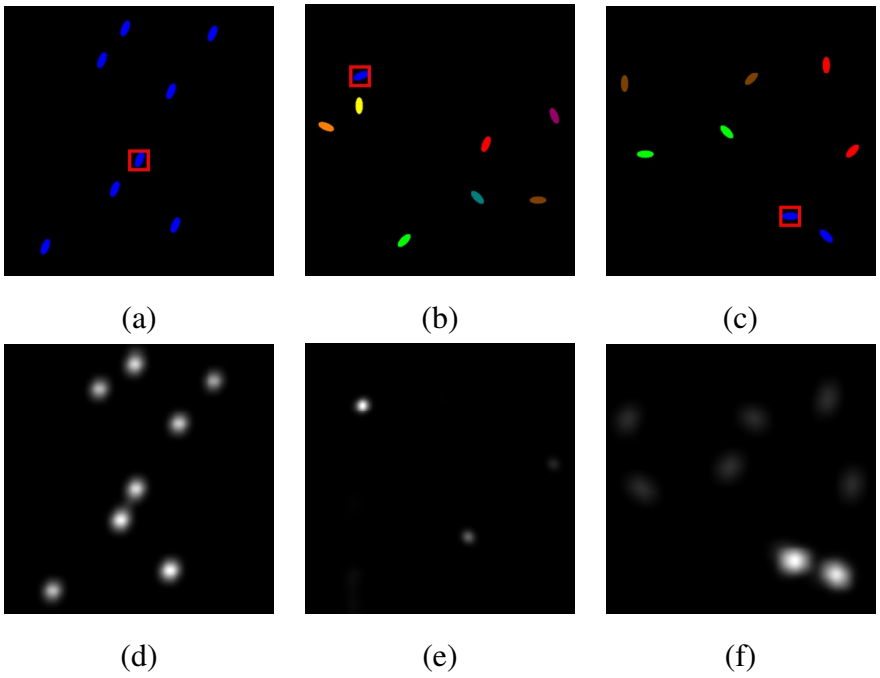


Figure 9.4: Top row shows frames from stimuli used in the experiment of [112]. Three conditions were tested (a) homogeneous (b) distinct and (c) conjunction distinct. The bottom row shows the confidence maps obtained by the saliency network for the frames in the top row. The confidence of the target as compared to distractors is highest in the case of (e) corresponding to the "unique" condition. In both other conditions (d) and (f) there are distractors that display high confidence, creating a possibility of tracking loss.

target location in the next frame.

9.3 Comparison of Model to Electrophysiological Recording Data

The selective enhancement and suppression of features with the mechanism of Section 8.4 bears a close resemblance to the phenomenon of feature-based attention [116]. As further validation of the biological plausibility of the tracking network, we measured its responses to random dot pattern (RDP) stimuli, and compared them to the responses reported in the literature from electrophysiological recordings of MT neurons. A hallmark property of these neurons is that feature based attention increases the gain of direction-selective neurons [168]. Trujillo and Treue showed that the mod-

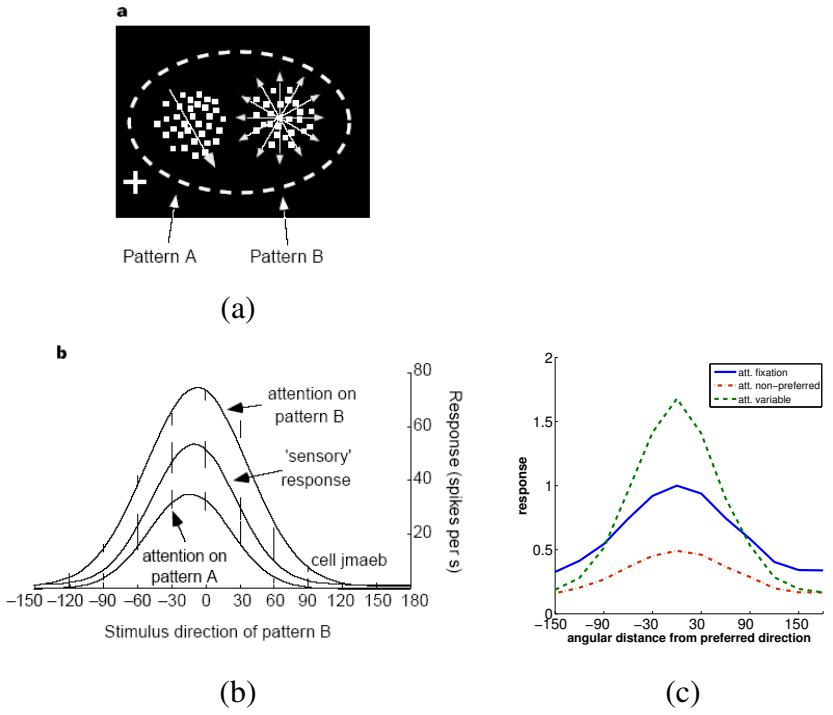


Figure 9.5: The multiplicative modulation of tuning curves. (a) and (b) are reproduced with permission from [168]. (c) Results obtained using the proposed saliency based network model. The enhancement of the response when Pattern B is attended, and attenuation when Pattern A is attended match the observed data in (b).

ulation is multiplicative [168]. In their experiment, they recorded the response from MT neurons in a macaque monkey when two RDPs were shown in the visual field located inside the receptive field (RF) of a neuron as illustrated in Figure 9.5. One of the RDPs (denoted Pattern A in the figure) always moved in the anti-preferred direction of the neuron. The second RDP (denoted Pattern B) moved in one of 12 possible directions. Recordings from the neuron were obtained under three conditions (i) attention to Pattern A (ii) attention to Pattern B, and (iii) attention to a task irrelevant fixation point, corresponding to the baseline sensory response. It was observed that, compared to the baseline in the third condition, the response of the neuron was enhanced in the first condition, and suppressed in the second.

In a subsequent experiment, Trujillo and Treue showed that the extent of modulation follows a monotonically decreasing function of the angular difference between the attended motion direction and the neuron's preferred direction [114]. To show this,

they used two RDPs, one of which was located inside the receptive field of the neuron whose response was being recorded, while the other RDP was located outside. The RDP inside the RF moved in the same direction as the one outside. There were two settings depending on the location the macaque was attending to. In the first setting, termed *attend fixation*, the monkey attended to a fixation point that was stationary, while in the second, denoted *attend same*, the monkey attend to the RDP stimulus outside the RF. These two settings are illustrated in Figure 9.6 (a). In each setting, the response of the neuron was recorded when the RDP moved in one of 12 different directions at a uniform spacing of 30°. The average firing rate recorded from one of the neurons is reproduced in Figure 9.6 (b). The average modulation ratio, defined as the ratio of the response in the attend-same condition to that in the attend-fixed condition, for the 135 neurons studied is plotted in Figure 9.6 (c).

To investigate if these results can be accounted for by the saliency hypothesis, 12 model MT neurons tuned to stimulus moving with the same uniform speed but in 12 different directions, $0°, 30°, \ldots 330°$, were constructed using the saliency model of Section 8.3. Twelve RDPs were generated using the Psychtoolbox [24], moving in each of the 12 preferred directions of the neurons.

To replicate the experiment of [168], one model neuron with preferred direction of 60° was considered, and one RDP corresponding to Pattern A and RDPs moving in the 12 directions corresponding to the 12 different directions for Pattern B were used. For each of the 12 RDPs, the three conditions were simulated, by initially training top-down feature weights using (i) Pattern A, i.e. one moving in the anti-preferred direction of the RF, (ii) on Pattern B and (iii) no training, i.e. the top-down weights were assigned equally, corresponding to a uniform prior. The output of the model neuron as a function of the direction of Pattern B, in all three conditions is shown in Figure 9.5(c). It is clear that the model replicates the multiplicative modulation observed in the recordings.

To reproduce the results of [114] in the attend fixation condition, no moving stimulus used to train the top-down weights and the response of the 12 neurons was computed without the feature selection mechanism of Section 8.4. In the case of the attend same condition, the RDP outside the RF is attended to. This was simulated by computing the responses of the 12 neurons to that RDP, and including the top-down

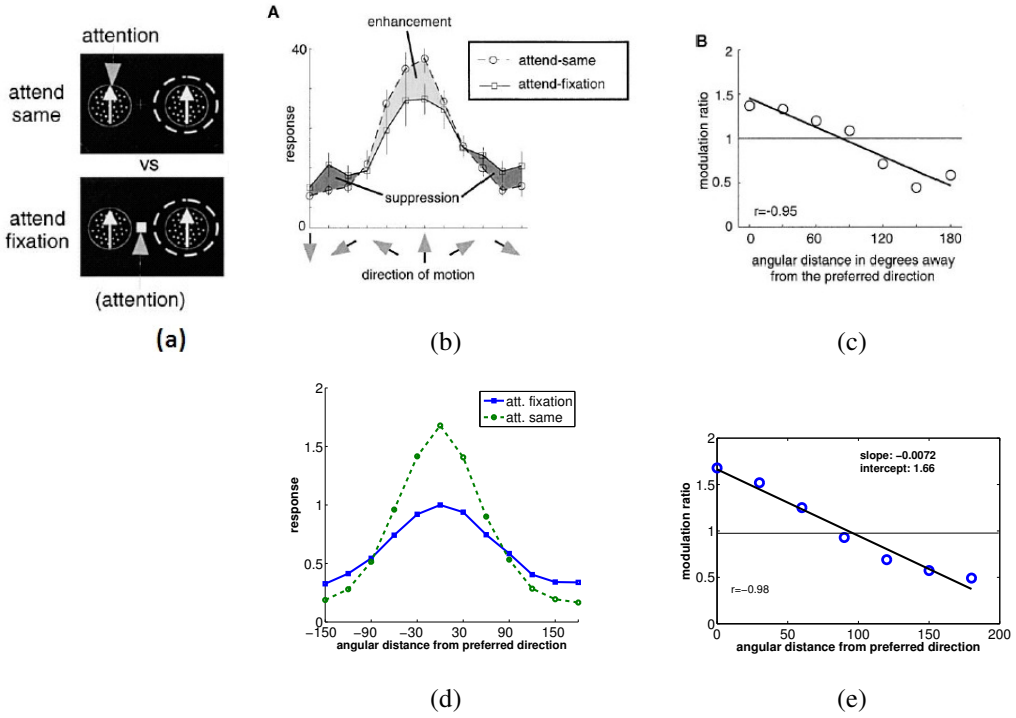


Figure 9.6: Comparison of responses of the model with the recordings from MT neurons. The top row, reproduced with permission from [114] shows (a) The panel on the top represents the *attend-same* condition, the one below represents *attend-fixation* (b) the average firing rate of an MT neuron in the two conditions as a function of the direction of RDP, and (c) average modulation ratios between the responses in the two conditions. (d) and (e) show the results obtained using the proposed model

feature weights in the saliency computations for the RDP inside the RF. The modulation ratio for each neuron was computed using (8.37). The responses of the model neurons under the two conditions are shown in Figure 9.6 (d) and the modulation ratio in Figure 9.6 (e). The model faithfully replicates the modulation trend observed in the neuronal recordings, showing enhancement for neurons whose preferred direction is close to that being attended, and suppression for anti-preferred directions. Finally, the monotonic fall of the modulation ratio is also accurately replicated by the model.

The proposed saliency based model is also qualitatively compatible with other findings from physiology. For instance, Katzner et al. [96] found that attention to a feature of the object also affects un-attended features. This is consistent with the feedback based update of the model which enhances features based on both the prior probability

of the feature being selected and the bottom-up saliency. If the un-attended features make the object salient, their weight in the saliency computation of (8.39) also increases after a few iterations. In fact, Boehler et al. [19] observed a delay of $\sim$ 80msec between the onset of attention to a feature of the object and the enhancement of irrelevant features, which can be accounted for by the delay-based feedback of feature weights in the network of Figure 8.3.

Several previously proposed models for feature-based attention also rely on a Bayesian formulation and incorporate a divisive normalization in the computation [142, 144, 102, 38]. However, these models are merely computational and often do not provide a physiological justification. On the other hand, th computations involved in the proposed network have been shown to naturally follow from the saliency based formulation, and can be mapped onto a plausible architecture. Other models for feature based attention such as the microcircuit model [11] have little neurophysiological support.

9.4 Discussion

This experiments in this chapter substantially strengthen the saliency hypothesis for tracking, by providing evidence that supports it in two ways. First, the biologically plausible saliency based tracking model introduced in Chapter 8 accurately predicts the results of the psychophysics experiments of Chapter 7. Second, the model also replicates electrophysiological data from MT neurons.

9.5 Acknowledgments

The text of Chapter 9, in full, is based on the material as it appears in: V. Mahadevan, and N. Vasconcelos, “Biological plausibility of the saliency hypothesis for visual tracking”, in preparation. The dissertation author was a primary researcher and an author of the cited material.

Chapter 10

Conclusions

In this work, we have shown that the center-surround discriminant saliency formulation can be used to identify salient regions in moving stimuli. The motion saliency algorithm proposed is inspired by biological vision, and is consistent with the psychophysics of motion-based perceptual grouping, and extends a discriminant formulation of center-surround saliency previously proposed for static imagery [65]. It combines spatial and temporal components of saliency in a principled manner, and is completely unsupervised. The combination of the discriminant center-surround saliency framework with the modeling power of dynamic textures leads to a robust and versatile procedure for background subtraction, which is successful even for scenes with highly dynamic backgrounds and those shot using moving cameras.

Further, we have shown that the discriminant formulation of motion saliency can be extended to perform discriminant tracking. The resultant framework is simple and computationally efficient which is consistent with what is known about the attentional mechanisms of biological vision, with an implementation that combines bottom-up center-surround discriminant saliency and spatial attention for learning, feature-based attention for feature selection, and top-down saliency for target detection. This provides a unified solution to the problems of classifier design, target detection, automatic tracker initialization, and scale adaptation.

Finally, we suggest that the connections between saliency and tracking exploited in the discriminant saliency tracker could be the basis of tracking in biological visual systems. We have provided evidence that supports this hypothesis in three ways. First,

we performed human behavior studies that show tracking requires discrimination between target and background using a center-surround mechanism, and that tracking reliability and saliency have a common dependence on feature contrast. Second, the hypothesis was shown to be neurophysiologically plausible, through construction of a tracking model that can be implemented with widely accepted models of cortical computation. Specifically, a tracking model based on MT neurons was constructed, and it was shown that saliency based tracking can be implemented with a feature selection mechanism akin to the well known phenomenon of feature-based attention in MT.

10.1 Future Work

The work in this dissertation opens up several avenues for future research.

In computer vision, the saliency based tracker proposed here can be improved to handle complete occlusions or re-entry of targets that have left the scene. It can also be extended to track multiple targets. This would involve augmenting the tracker with additional modules such as an identity management scheme that controls the initialization or termination of targets when necessary and can match potential tracks with targets.

This work is a preliminary attempt to understand the computational basis of visual tracking in biological vision. Much more research is needed to elucidate the connections between processing of saliency and tracking in the primate visual system. In this direction, more human behavior studies can be performed to understand the effect of different types of features e.g. shape, color, and motion etc. on tracking performance. Further, neurophysiological measurements from V1, MT and LIP neurons in customized experiments could be extremely helpful in validating the discriminative nature of saliency processing and tracking and to get direct evidence for the role of feature based attention in tracking

Appendix A

Implementation Details

A.1 Motion saliency and background subtraction

At each location, the center window occupied 16×16 pixels and spanned 11 frames - 5 past, current, and 5 future ($n_c = 16, \tau = 11$). The causal version of Algorithm 1 (denoted DiscSal-Causal) was implemented, by considering only the current and 10 past frames. In all cases, the surround window was set to 6 times the size of the center (i.e $96 \times 96 \times 11$). DTs with a 10-dimensional state space, patch dimension $n_p = 8$, and temporal dimension $\tau = 11$, were learned using overlapping $8 \times 8 \times 11$ patches from the center and surround windows.

For the version with a biologically plausible motion model, we used spatio-temporal Gabor filters. We considered only one spatial scale, and the spatial frequency of each Gabor filter was fixed to 0.25 cycles/pixel. Three temporal scales (temporal frequencies of 0, ± 0.25 cycles/frame) and 4 spatial orientations ($0, \pi/4, \pi/2$ and $3\pi/4$) were used, in a total of 12 filters. The standard deviation of the spatial Gaussian was set to 1, and that of the temporal Gaussian to 2

A.2 Discriminant tracking

The value of the scale parameter for GGD was set to $\beta = 0.7$.

The decay factor, λ, in estimating the parameters of the GGD in (5.20) is set to

0.35.

For the spatio-temporal features, a single spatial frequency of 0.25 cycles/pixel was used for all Gabor filters. Since most sequences considered have predominantly horizontal motion, a single spatial orientation of 0° (aligned with the horizontal axis) was used. Temporally, three frequencies of 0 cycles/frames (stationary objects) and ±0.25 cycles/frames (objects moving to the left or right) were chosen, for a total of 3 motion energy filters. The number of pyramid levels for the DCT basis functions was taken to be 2. Therefore the total number of features was $N = 3 + 64 \times 2 = 131$ features (8×8 DCT features per level plus three spatiotemporal Gabor features)

To guarantee a realistic balance between tracking performance and computation, the number of salient features K was set to 5.

The search neighborhood, $\mathcal{W}^s_{l^*}$, was set to a rectangular region centered at the current target position l^* with size twice that of the object bounding box.

For scale adaptive tracking, the range over which scale was searched is $s \in (0.8, 1.2)$, in steps of 0.05, of the current target size (both height and width vary with the aspect ration being fixed). The scale adaptation is only done once in 2 frames.

A.3 Biologically plausible model for tracking

For the MT model we use a total of 12 spatio-temporal filters, each with center frequency $(\omega_{x_j}, \omega_{y_j}, \omega_{t_j})$, $j = 0 \ldots 11$. We consider 12 neurons each tuned to motion with constant speed in one of 12 different directions spread uniformly in $(0°, 360°)$, corresponding to velocities $\bar{v}_k$, $k = 0 \ldots 11$.

The value of the scale parameter for GGD was assumed to be $\beta = 1$.

Bibliography

[1] "http://homepages.inf.ed.ac.uk/rbf/caviar/."

[2] "http://staff.science.uva.nl/ zivkovic/download.html."

[3] "http://www.svcl.ucsd.edu/projects/background_subtraction."

[4] "http://www.svcl.ucsd.edu/projects/tracking_biological/."

[5] "http://www.svcl.ucsd.edu/projects/tracking/results.html."

[6] "See attached supplementary material."

[7] A. Adam, E. Rivlin, and I. Shimshoni, "Robust Fragments-based Tracking using the Integral Histogram," in *IEEE Conference on Computer Vision and Pattern Recognition.*, 2006, pp. 798–805.

[8] E. H. Adelson and J. R. Bergen, "Spatiotemporal energy models for the perception of motion." *Journal of the Optical Society of America A*, vol. 2, no. 2, pp. 284–299, 1985. [Online]. Available: citeseer.ist.psu.edu/ adelson85spatiotemporal.html

[9] S. Agarwal and D. Roth, "Learning a sparse representation for objection detection," in *Proceedings European Conference on Computer Vision*, vol. 4, 2002, pp. 113–130.

[10] R. Allen, P. Mcgeorge, D. Pearson, and A. B. Milne, "Attention and expertise in multiple target tracking," *Applied Cognitive Psychology*, vol. 18, no. 3, pp. 337–347, 2004.

[11] S. Ardid, X. Wang, and A. Compte, "An integrated microcircuit model of attentional processing in the neocortex," *Journal of Neuroscience*, vol. 27, no. 32, p. 8486, 2007.

[12] S. Avidan, "Ensemble tracking," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 2, pp. 261–271, 2007.

[13] B. Babenko, M.-H. Yang, and S. Belongie, "Visual tracking with online multiple instance learning," *IEEE Conference on Computer Vision and Pattern Recognition*, vol. 0, pp. 983–990, 2009.

[14] M. M. Bence P. lveczky, Stephen A. Baccus, "Segregation of object and background motion in the retina," *Nature*, vol. 423, pp. 401–408, 2003.

[15] S. Birchfield and S. Rangarajan, "Spatiograms versus histograms for region-based tracking," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2005, pp. 1158–1163.

[16] J. Bisley and M. Goldberg, "Attention, intention, and priority in the parietal lobe," *Annual review of neuroscience*, vol. 33, pp. 1–21, 2010.

[17] M. Black and A. Jepson, "Eigentracking: Robust matching and tracking of articulated objects using a view-based representation," *International Journal of Computer Vision*, vol. 26, no. 1, pp. 63–84, 1998.

[18] E. Blaser, Z. Pylyshyn, and A. O. Holcombe, "Tracking an object through feature-space," *Nature*, vol. 408, pp. 196–199, 2000.

[19] C. Boehler, M. Schoenfeld, H. Heinze, and J. Hopf, "Object-based Selection of Irrelevant Features Is Not Confined to the Attended Object," *Journal of Cognitive Neuroscience*, no. Early Access, pp. 1–9, 2010.

[20] E. Borenstein and S. Ullman, "Learn to segment," in *Proceedings European Conference on Computer Vision*, 2004, pp. 315–328.

[21] R. T. Born, J. Groh, R. Zhao, and S. J. Lukasewycz, "Segregation of object and background motion in visual area MT: Effects of microstimulation on eye movements," *Neuron*, vol. 26, pp. 725–734, 2000.

[22] R. T. Born and R. B. H. Tootell, "Segregation of global and local motion processing in primate middle temporal visual area," *Nature*, vol. 357, pp. 497–499, 1992.

[23] R. Born and D. Bradley, "Structure and Function of Visual Area MT," *Annu. Rev. Neurosci*, vol. 28, pp. 157–89, 2005.

[24] D. H. Brainard, "The psychophysics toolbox," *Spatial Vision*, vol. 10, pp. 433–436, 1997.

[25] L. Bretzner and T. Lindeberg, "Feature Tracking with Automatic Selection of Spatial Scales," *Computer Vision and Image Understanding*, vol. 71, no. 3, pp. 385–392, 1998.

[26] N. Bruce and J. Tsotsos, "Saliency, attention, and visual search: An information theoretic approach," *Journal of Vision*, vol. 9, no. 3, 2009.

[27] ——, "Saliency based on information maximization," in *Advances in Neural Information Processing Systems 18*, Y. Weiss, B. Schölkopf, and J. Platt, Eds. Cambridge, MA: MIT Press, 2006, pp. 155–162.

[28] R. Buccigrossi and E. Simoncelli, "Image compression via joint statistical characterization in the wavelet domain," *IEEE Transactions on Image Processing*, vol. 8, pp. 1688–1701, 1999.

[29] A. Bugeau and P. Perez, "Detection and segmentation of moving objects in highly dynamic scenes," in *Computer Vision and Pattern Recognition*, 2007.

[30] M. Carandini, J. Demb, V. Mante, D. Tolhurst, Y. Dan, B. Olshausen, J. Gallant, and N. Rust, "Do we know what the early visual system does?" *Journal of Neuroscience*, vol. 25, pp. 10 577–10 597, 2005.

[31] P. Cavanagh, "Attention-based motion perception," *Science*, vol. 257, no. 5076, pp. 1563–1565, 1992. [Online]. Available: http://www.jstor.org/stable/2879947

[32] P. Cavanagh and G. A. Alvarez, "Tracking multiple targets with multifocal attention," *Trends in Cognitive Sciences*, vol. 9, no. 7, pp. 349 – 354, 2005. [Online]. Available: http://www.sciencedirect.com/science/article/B6VH9-4GCX1TC-8/2/be01ca0d7179cb1bb28e6c9b9d780255

[33] J. Cavanaugh, W. Bair, and J. Movshon, "Nature and interaction of signals from the receptive field center and surround in macaque V1 neurons," *Journal of Neurophysiology*, vol. 88, pp. 2530–2546, 2002.

[34] A. B. Chan and N. Vasconcelos, "Efficient computation of the kl divergence between dynamic textures," Dept. of ECE, UCSD, Tech. Rep. SVCL-TR-2004-02, 2004. [Online]. Available: http://www.svcl.ucsd.edu/publications/techreports/SVCL-TR-2004-02.pdf

[35] ——, "Probabilistic kernels for the classification of auto-regressive visual processes," in *IEEE Conference on Computer Vision and Pattern Recognition*, vol. 1, 2005, pp. 846–851.

[36] L. Cheng, S. Wang, D. Schuurmans, T. Caelli, and S. Vishwanathan, "An online discriminative approach to background subtraction," in *Advanced Video and Signal Based Surveillance*, 2006, p. 2.

[37] H. Chernoff, "On the distribution of the likelihood ratio," *The Annals of Mathematical Statistics*, vol. 25, no. 3, pp. 573–578, 1954. [Online]. Available: http://links.jstor.org/sici?sici=0003-4851\%28195409\%2925\%3A3\%3C573\%3AOTDOTL\%3E2.0.CO\%3B2-P

[38] S. Chikkerur, T. Serre, C. Tan, and T. Poggio, "What and where: A Bayesian inference theory of attention," *Vision Research*, 2010.

[39] R. Collins, Y. Liu, and M. Leordeanu, "On-line selection of discriminative tracking features," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 10, pp. 1631 – 1643, October 2005.

[40] R. Collins, "Mean-shift blob tracking through scale space," in *IEEE Conference on Computer Vision and Pattern Recognition*, vol. 2, 2003.

[41] D. Comaniciu, V. Ramesh, and P. Meer, "Kernel-based object tracking," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 5, pp. 564–577, 2003. [Online]. Available: http://dx.doi.org/10.1109/TPAMI.2003.1195991

[42] T. Cover and J. Thomas, *Elements of Information Theory*. New York: John Wiley & Sons Inc., 1991.

[43] D. Cremers and F. Soatto, "Dynamic texture segmentation," in *International Conference on Computer Vision*, 2003, pp. 1236–1242. [Online]. Available: citeseer.ist.psu.edu/cremers03dynamic.html

[44] R. Cucchiara, C. Grana, M. Piccardi, and A. Prati, "Detecting moving objects, ghosts, and shadows in video streams," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 10, pp. 1337–1342, October 2003.

[45] J. C. Culham, S. A. Brandt, P. Cavanagh, N. G. Kanwisher, A. M. Dale, and R. B. Tootell, "Cortical fmri activation produced by attentive tracking of moving targets." *J Neurophysiol*, vol. 80, no. 5, pp. 2657–2670, November 1998. [Online]. Available: http://jn.physiology.org/cgi/content/abstract/80/5/2657

[46] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, June 2005.

[47] M. N. Do and M. Vetterli, "Wavelet-based texture retrieval using generalized gaussian density and kullback-leibler distance," *IEEE Transactions on Image Processing*, vol. 11, no. 2, pp. 146–158, 2002.

[48] M. Doran and J. Hoffman, "Spatial attention in multiple object tracking: Evidence from ERPs," *Journal of Vision*, vol. 8, no. 6, p. 505, 2008.

[49] G. Doretto, A. Chiuso, Y. N. Wu, and S. Soatto, "Dynamic textures," *International Journal of Computer Vision*, vol. 51, no. 2, pp. 91–109, 2003.

[50] J. Duncan, "Selective attention and the organization of visual information." *Journal of Experimental Psychology: General*, vol. 113, no. 4, pp. 501–517, 1984.

[51] A. Elgammal, D. Harwood, and L. Davis, "Non-parametric model for background subtraction," in *European Conference on Computer Vision*, 2000, pp. 751–757.

[52] A. Elgammal, R. Duraiswami, D. Harwood, and L. S. Davis, "Background and foreground modeling using nonparametric kernel density for visual surveillance," in *Proceedings of the IEEE*, vol. 90, no. 7, Jul. 2002, pp. 1151–1163. [Online]. Available: citeseer.ist.psu.edu/elgammal02background.html

[53] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results," http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html.

[54] J. Feldman and P. D. Tremoulet, "Individuation of visual objects over time," *Cognition*, vol. 99, no. 2, pp. 131 – 165, 2006. [Online]. Available: http://www.sciencedirect.com/science/article/B6T24-4G0YTJJ-1/2/b5f49179cd12e2d541e5889d7c19a629

[55] R. Fergus, P. Perona, and A. Zisserman, "Object class recognition by unsupervised scale-invariant learning," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2003.

[56] F. Fleuret and I. Guyon, "Fast binary feature selection with conditional mutual information," *Journal of Machine Learning Research*, vol. 5, pp. 1531–1555, 2004.

[57] W. Förstner, "A framework for low level feature extraction," in *Proceedings of European Conference on Computer Vision*, 1994, pp. 383–394.

[58] Y. Freund and R. E. Schapire, "1997, a decision-theoretic generalization of on-line learning and an application to boosting," in *European Conference on Computational Learning Theory*, 1995, pp. 23–37. [Online]. Available: http://citeseer.ist.psu.edu/freund95decisiontheoretic.html

[59] D. Gao, S. Han, and N. Vasconcelos, "Discriminant saliency, the detection of suspicious coincidences, and applications to visual recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 6, p. 989, 2009.

[60] D. Gao, V. Mahadevan, and N. Vasconcelos, "On the plausibility of the discriminant center-surround hypothesis for visual saliency," *Journal of Vision*, vol. 8, no. 7, pp. 1–18, 6 2008. [Online]. Available: http://journalofvision.org/8/7/13/

[61] D. Gao and N. Vasconcelos, "Bottom-up saliency is a discriminant process," in *International Conference on Computer Vision*, 2007.

[62] ——, "Discriminant interest points are stable," in *IEEE Conference on Computer Vision and Pattern Recognition*, Jun 2007.

[63] ——, "Discriminant saliency for visual recognition from cluttered scenes," in *Advances in Neural Information Processing Systems*, 2005.

[64] ——, "V1 is an optimal saliency detector," in *Computational Cognitive Neuroscience Conference (CCNC)*, 2007.

[65] ——, "Decision-theoretic saliency: computational principle, biological plausibility, and implications for neurophysiology and psychophysics," *Neural Computation*, vol. 21, pp. 239–271, Jan 2009.

[66] H. Grabner and H. Bischof, "On-line boosting and vision," *IEEE Conference on Computer Vision and Pattern Recognition*, vol. 1, pp. 260–267, 2006.

[67] H. Grabner, C. Leistner, and H. Bischof, "Semi-supervised on-line boosting for robust tracking," in *European Conference on Computer Vision*, 2008, pp. 234–247.

[68] B. Han and L. Davis, "On-Line Density-Based Appearance Modeling for Object Tracking," in *IEEE International Conference on Computer Vision*, 2005, p. 1499.

[69] S. Han and N. Vasconcelos, "Biologically Plausible Saliency Mechanisms Improve Feedforward Object Recognition," *Vision Research*, 2010.

[70] C. Harris and M. Stephens, "A combined corner and edge detector," in *Alvey Vision Conference*, 1988.

[71] E. Hayman and J. Eklundh, "Statistical background subtraction for a mobile observer," in *In Proceedings International Conference on Computer Vision, 2003.*, 2003. [Online]. Available: citeseer.ist.psu.edu/hayman03statistical.html

[72] D. Heeger, "Models for motion perception," *Ph.D.thesis, Univ. of Pennsylvania, 1987.*, 1987.

[73] ——, "Optical flow from spatiotemporal filters," *International Journal of Computer Vision*, vol. 1, no. 4, pp. 279–302, 1988.

[74] G. Heidemann, "Focus-of-attention from local color symmetries," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 7, pp. 817–830, 2004.

[75] M. Heikkila and M. Pietikainen, "A texture-based method for modeling the background and detecting moving objects," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 4, pp. 657–62, 2006.

[76] T. Herrington and J. Assad, "Temporal sequence of attentional modulation in the lateral intraparietal area and middle temporal area during rapid covert shifts of attention," *The Journal of Neuroscience*, vol. 30, no. 9, p. 3287, 2010.

[77] J. Ho, K. Lee, M. Yang, and D. Kriegman, "Visual tracking using learned linear subspaces," in *IEEE Conference on Computer Vision and Pattern Recognition*, vol. 1, 2004.

[78] T. S. Horowitz, S. B. Klieger, D. E. Fencsik, K. K. Yang, G. A. Alvarez, and J. M. Wolfe, "Tracking unique objects." *Percept Psychophys*, vol. 69, no. 2, pp. 172–184, Feb 2007.

[79] G. A. Horridge, "The evolution of visual processing and the construction of seeing systems," *Proceedings of the Royal Society of London. Series B, Biological Sciences*, vol. 230, no. 1260, pp. 279–292, 1987.

[80] J. Huang and D. Mumford, "Statistics of Natural Images and Models," in *IEEE Conference on Computer Vision and Pattern Recognition*, 1999, pp. 541–547.

[81] D. H. Hubel and T. N. Wiesel, "Receptive fields and functional architecture in two nonstriate visual areas (18 and 19) of the cat," *Journal of Neurophysiology*, vol. 28, pp. 229–289, 1965.

[82] J. Intriligator and P. Cavanagh, "The spatial resolution of visual attention," *Cognitive Psychology*, vol. 43, pp. 171–216, 1997.

[83] M. Isard and A. Blake, "Condensation conditional density propagation for visual tracking," *International Journal of Computer Vision*, vol. 29, pp. 5–28, 1998.

[84] L. Itti, "Automatic foveation for video compression using a neurobiological model of visual attention," *IEEE Transactions on Image Processing*, vol. 13, no. 10, pp. 1304–1318, 2004.

[85] ——, "The ilab neuromorphic vision c++ toolkit: Free tools for the next generation of vision algorithms," *The Neuromorphic Engineer*, vol. 1, no. 1, p. 10, Mar 2004.

[86] L. Itti and P. Baldi, "A principled approach to detecting surprising events in video," in *Computer Vision and Pattern Recognition*, 2005, pp. 631–637.

[87] ——, "Bayesian surprise attracts human attention," in *Advances in Neural Information Processing Systems*, vol. 19. Cambridge, MA: MIT Press, 2006, su;mod;bu;td;eye, pp. 1–8.

[88] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 11, pp. 1254–1259, 1998.

[89] L. Itti and C. Koch, "A saliency-based search mechanism for overt and covert shifts of visual attention," *Vision Research*, vol. 40, pp. 1489–1506, 2000.

[90] R. B. Ivry and A. Cohen, "Asymmetry in visual search for targets defined by differences in movement speed," *J Exp Psychol Hum Percept Perform*, vol. 18, pp. 1045–1057, 1992.

[91] N. Jacobson, Y. Lee, V. Mahadevan, N. Vasconcelos, and T. Nguyen, "A novel approach to fruc using discriminant saliency and frame segmentation," *Image Processing, IEEE Transactions on*, vol. 19, no. 11, pp. 2924–2934, 2010.

[92] A. D. Jepson, D. J. Fleet, and T. F. El-Maraghi, "Robust online appearance models for visual tracking," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 10, pp. 1296–1311, 2003. [Online]. Available: http://dx.doi.org/10.1109/TPAMI.2003.1233903

[93] T. Kadir and M. Brady, "Scale, saliency and image description," *International Journal of Computer Vision*, vol. 45, pp. 83–105, Nov. 2001.

[94] D. Kahneman, A. Treisman, and B. J. Gibbs, "The reviewing of object files: Object-specific integration of information," *Cognitive Psychology*, vol. 24, no. 2, pp. 175–219, April 1992. [Online]. Available: http://dx.doi.org/10.1016/0010-0285(92)90007-O

[95] K. Karmann and A. V. Brandt, "Moving object recognition using an adaptive background memory," in *Time-Varing Image Processing and Moving Object Recognition Vol 2*, 1990, pp. 289–296.

[96] S. Katzner, L. Busse, and S. Treue, "Attention to the color of a moving stimulus modulates motion-signal processing in macaque area MT: evidence for a unified attentional system," 2009.

[97] Y. Kazanovich and R. Borisyuk, "An oscillatory neural model of multiple object tracking," *Neural computation*, vol. 18, no. 6, pp. 1413–1440, 2006.

[98] B. Keane and Z. Pylyshyn, "Is motion extrapolation employed in multiple object tracking? Tracking as a low-level, non-predictive function," *Cognitive psychology*, vol. 52, no. 4, pp. 346–368, 2006.

[99] C. Koch and S. Ullman, "Shift in selective visual attention: towards the underlying neural circuitry," *Hum. Neurobiol.*, vol. 4, pp. 219–227, 1985.

[100] D. Koller, J. Weber, and J. Malik, "Robust multiple car tracking with occlusion reasoning," in *European Conference on Computer Vision (1)*, 1994, pp. 189–196. [Online]. Available: citeseer.ist.psu.edu/article/koller93robust.html

[101] S. Kullback, *Information Theory and Statistics*. Dover Publications, New York, 1968.

[102] J. Lee and J. Maunsell, "A normalization model of attentional modulation of single unit responses," *PLoS One*, vol. 4, no. 2, 2009.

[103] J. Lewis and D. Van Essen, "Corticocortical connections of visual, sensorimotor, and multimodal processing areas in the parietal lobe of the macaque monkey," *The Journal of comparative neurology*, vol. 428, no. 1, pp. 112–137, 2000.

[104] Y. Li, "On incremental and robust subspace learning," *Pattern Recognition*, vol. 37, no. 7, pp. 1509–19, 2004.

[105] Z. Li, "A saliency map in primary visual cortex," *Trends in Cognitive Sciences*, vol. 6, no. 1, pp. 9–16, 2002.

[106] Y. li Tian and A. Hampapur, "Robust salient motion detection with complex background for real-time video surveillance," in *IEEE Workshop on Applications of Computer Vision*, 2005, pp. 30–35.

[107] R. Lin, D. Ross, J. Lim, and M. Yang, "Adaptive discriminative generative model and its applications," *Advances in neural information processing systems*, pp. 801–808, 2004.

[108] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proceedings International Conference on Computer Vision*, 1999, pp. 1150–1157.

[109] E. Maggio and A. Cavallaro, "Hybrid particle filter and mean shift tracker with adaptive transition model," in *ICASSP*, 2005.

[110] V. Mahadevan and N. Vasconcelos, "Saliency-based discriminant tracking," *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, vol. 0, pp. 1007–1013, 2009.

[111] ——, "Background subtraction in highly dynamic scenes," *IEEE Conference on Computer Vision and Pattern Recognition*, vol. 1, 2008.

[112] T. Makovski and Y. Jiang, "Feature binding in attentive tracking of distinct objects," *Visual cognition*, vol. 17, no. 1, pp. 180–194, 2009.

[113] J. Malik and P. Perona, "Preattentive texture discrimination with early vision mechanisms," *J. Opt. Soc. Am. A*, vol. 7, no. 5, pp. 923–932, May 1990.

[114] J. Martinez-Trujillo and S. Treue, "Feature-based attention increases the selectivity of population responses in primate visual cortex," *Current Biology*, vol. 14, no. 9, pp. 744–751, 2004.

[115] H. Masnadi-Shirazi, V. Mahadevan, and N. Vasconcelos, "On the design of robust classifiers for computer vision," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2010, pp. 779–786.

[116] J. Maunsell and S. Treue, "Feature-based attention in visual cortex," *TRENDS in Neurosciences*, vol. 29, no. 6, pp. 317–322, 2006.

[117] A. Mittal and N. Paragios, "Motion-based background subtraction using adaptive kernel density estimation," in *Computer Vision and Pattern Recognition*, 2004.

[118] A. Monnet, A. Mittal, N. Paragios, and V. Ramesh, "Background modeling and subtraction of dynamic scenes," in *Computer Vision and Pattern Recognition*, 2003.

[119] A. Murray, D. Basu, "Motion tracking with an active camera," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 16, no. 5, pp. 449–459, 1994.

[120] V. Navalpakkam and L. Itti, "Search goal tunes visual features optimally," *Neuron*, vol. 53, no. 4, pp. 605–617, 2007.

[121] H. Nguyen and A. Smeulders, "Robust tracking using foreground-background texture discrimination," *International Journal of Computer Vision*, vol. 69, no. 3, pp. 277–293, 2006.

[122] P. Noriega and O. Bernier, "Real time illumination invariant background subtraction using local kernel histograms," in *British Machine Vision Conference*, vol. 3, 2006, pp. 979–988.

[123] H. C. Nothdurft, "The role of local contrast in pop-out of orientation, motion and color," *Investigative Ophthalmology and Visual Science*, vol. 32, no. 4, p. 714, 1991.

[124] ——, "Texture segmentation and pop-out from orientation contrast," *Vision Research*, vol. 31, no. 6, pp. 1073–1078, 1991.

[125] ——, "Feature analysis and the role of similarity in preattentive vision," *Perception and Psychophysics*, vol. 52, no. 4, pp. 355–375, 1992.

[126] ——, "The conspicuousness of orientation and motion contrast," *Spatial Vision*, vol. 7, pp. 341–363, 1993.

[127] ——, "The role of features in preattentive vision: Comparison of orientation, motion and color cues," *Vision Research*, vol. 33, no. 14, pp. 1937–1958, 1993.

[128] ——, "Salience from feature contrast: additivity across dimensions," *Vision Research*, vol. 40, pp. 1183–1201, 2000.

[129] ——, "Salience from feature contrast: variations with texture density," *Vision Research*, vol. 40, pp. 3181–3200, 2000.

[130] L. Oksama and J. Hyn, "Is multiple object tracking carried out automatically by an early vision mechanism independent of higher-order cognition? an individual difference approach." *Visual Cognition*, vol. 11, no. 5, pp. 631 – 671, 2004. [Online]. Available: http://search.ebscohost.com/login.aspx?direct=true&db=a9h&AN=13310802&site=ehost-live

[131] N. M. Oliver, B. Rosario, and A. P. Pentland, "A Bayesian computer vision system for modeling human interactions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 831–43, 2000.

[132] P. V. Overschee and B. D. Moor, "N4sid: Subspace algorithms for the identification of combined deterministic-stochastic systems," *Automatica*, vol. 30, pp. 75–93, 1994.

[133] S. E. Palmer, *Vision Science: Photons to Phenomenology*. The MIT Press, 1999.

[134] D. R. Patzwahl and S. Treue, "Combining spatial and feature-based attention within the receptive field of mt neurons," *Vision Research*, vol. 49, no. 10, pp. 1188 – 1193, 2009, visual Attention: Psychophysics, electrophysiology and neuroimaging. [Online]. Available: http://www.sciencedirect.com/science/article/B6T0W-4W1SRMB-2/2/d478e352f65411fd27e2c4529bdba2d3

[135] M. Piccardi, "Background subtraction techniques: a review," in *IEEE International Conference on Systems, Man and Cybernetics*, Oct 2004, pp. 3099– 3104.

[136] D. Pokrajac and L. J. Latecki, "Spatiotemporal blocks-based moving objects identification and tracking," in *IEEE Visual Surveillance and Performance Evaluation of Tracking and Surveillance,*, 2003. [Online]. Available: citeseer.ist.psu.edu/article/koller93robust.html

[137] M. Posner, C. Snyder, and B. Davidson, "Attention and the detection of signals," *Journal of Experimental Psychology: General*, vol. 109, no. 2, pp. 160–174, 1980.

[138] ——, "Attention and the detection of signals," *Journal of Experimental Psychology: General*, vol. 109, no. 2, pp. 160–174, 1980.

[139] C. Privitera and L. Stark, "Algorithms for defining visual regions-of-interest: comparison with eye fixations," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, pp. 970–982, 2000.

[140] Z. W. Pylyshyn and R. W. Storm, "Tracking multiple independent targets: evidence for a parallel tracking mechanism." *Spatial vision*, vol. 3, no. 3, pp. 179–197, 1988. [Online]. Available: http://view.ncbi.nlm.nih.gov/pubmed/3153671

[141] D. Ramanan, D. Forsyth, and A. Zisserman, "Tracking people by learning their appearance," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 65–81, 2007.

[142] R. Rao, "Bayesian inference and attentional modulation in the visual cortex," *Neuroreport*, vol. 16, no. 16, p. 1843, 2005.

[143] Y. Ren, C. Chua, and Y. Ho, "Motion detection with nonstationary background," *Machine Vision and Applications*, vol. 13, no. 5-6, pp. 332–343, 2003.

[144] J. Reynolds and D. Heeger, "The normalization model of attention," *Neuron*, vol. 61, no. 2, pp. 168–185, 2009.

[145] D. Ross, J. Lim, R. Lin, and M. Yang, "Incremental learning for robust visual tracking," *International Journal of Computer Vision*, vol. 77, no. 1-3, pp. 125–141, May 2008.

[146] N. Rust, V. Mante, E. Simoncelli, and J. Movshon, "How MT cells analyze the motion of visual patterns," *Nature Neuroscience*, vol. 9, no. 11, pp. 1421–1431, 2006.

[147] Y. Saalmann, I. Pigarev, and T. Vidyasagar, "Neural mechanisms of visual attention: how top-down feedback highlights relevant locations," *Science*, vol. 316, no. 5831, p. 1612, 2007.

[148] H. Sakata, H. Shibutani, and K. Kawano, "Functional properties of visual tracking neurons in posterior parietal association cortex of the monkey," *J Neurophysiol*, vol. 49, no. 6, pp. 1364–1380, 1983. [Online]. Available: http://jn.physiology.org

[149] B. Schiele and J. Crowley, "Where to look next and what to look for," in *Intelligent Robots and Systems (IROS)*, 1996, pp. 1249–1255.

[150] A. B. Sekuler and R. Sekuler, "Collisions between moving visual targets: what controls alternative ways of seeing an ambiguous display?" *Perception*, vol. 28, no. 4, pp. 415–432, 1999. [Online]. Available: http://view.ncbi.nlm.nih.gov/pubmed/10664783

[151] T. Serre, L. Wolf, S. Bileschi, M. Riesenhuber, and T. Poggio, "Robust object recognition with cortex-like mechanisms," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 3, pp. 411–426, 2007.

[152] A. Sha'ashua and S. Ullman, "Structural saliency: the detection of globally salient structures using a locally connected network," in *Proceedings International Conference on Computer Vision*, 1988, pp. 321–327.

[153] K. Sharifi and A. Leon-Garcia, "Estimation of shape parameter for generalized gaussian distributions in subband decompositions of video," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 5, no. 1, pp. 52–56, 1995.

[154] Y. Sheikh and M. Shah, "Bayesian modeling of dynamic scenes for object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 11, pp. 1778–92, 2005.

[155] R. Shumway and D. Stoffer, "An approach to time series smoothing andforecasting using the EM algorithm," *Journal of Time Series Analysis*, vol. 3, no. 4, pp. 433–467, 1982.

[156] E. Simoncelli and D. Heeger, "A model of neuronal responses in visual area MT," *Vision Research*, vol. 38, no. 5, pp. 743–761, 1998.

[157] A. Srivastava, A. Lee, E. Simoncelli, and S. Zhu, "On advances in statistical modeling of natural images," *Journal of mathematical imaging and vision*, vol. 18, no. 1, pp. 17–33, 2003.

[158] C. Stauffer and W. Grimson, "Adaptive background mixture models for real-time tracking," in *IEEE Computer Vision and Pattern Recognition*, vol. 2, 1999, pp. 2246–2252.

[159] ——, "Adaptive background mixture models for real-time tracking," in *IEEE Conference on Computer Vision and Pattern Recognition*, vol. 2, 1999, pp. 246–252.

[160] K. Tanaka, K. Hikosaka, H. Saito, M. Yukie, Y. Fukada, and E. Iwai, "Analysis of local and wide-field movements in the superior temporal visual areas of the macaque monkey," *Journal of Neuroscience*, vol. 6, no. 1, p. 134, 1986.

[161] B. Tatler, "The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions," *Journal of Vision*, vol. 7, no. 14, 2007.

[162] A. Tavakkoli, M. Nicolescu, and G. Bebis, "A novelty detection approach for foreground region detection in videos with quasi-stationary backgrounds," in *International Symposium on Visual Computing*, 2006.

[163] A. Torralba, "Modeling global scene factors in attention," *JOSA A*, vol. 20, no. 7, pp. 1407–1418, 2003.

[164] K. Toyama and G. D. Hager, "Incremental focus of attention for robust visual tracking," in *International Journal of Computer Vision*, 1996, pp. 189–195.

[165] K. Toyama, J. Krumm, B. Brumitt, and B. Meyers, "Wallflower: Principles and practice of background maintenance," in *International Conference on Computer Vision (1)*, 1999, pp. 255–261. [Online]. Available: citeseer.ist.psu.edu/toyama99wallflower.html

[166] K. Toyama and Y. Wu, "Bootstrap initialization of nonparametric texture models for tracking," in *European Conference on Computer Vision*, 2000.

[167] A. Treisman and G. Gelade, "A feature-integratrion theory of attention," *Cognitive Psychology*, vol. 12, no. 1, pp. 97–136, 1980.

[168] S. Treue and J. Trujillo, "Feature-based attention influences motion processing gain in macaque visual cortex," *Nature*, vol. 399, no. 6736, pp. 575–579, 1999.

[169] P. Tseng, R. Carmi, I. Cameron, D. Munoz, and L. Itti, "Quantifying center bias of observers in free viewing of dynamic natural scenes," *Journal of Vision*, vol. 9, no. 7, 2009.

[170] M. Vasconcelos and N. Vasconcelos, "Natural image statistics and low-complexity feature selection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 2, pp. 228–244, 2008.

[171] N. Vasconcelos, "Feature selection by maximum marginal diversity," in *Advances in Neural Information Processing Systems*, 2002.

[172] ——, "Feature selection by maximum marginal diversity: optimality and implications for visual recognition," in *Proceedings IEEE Conference on Computer Vision and Pattern Recognition*, vol. 1, 2003, pp. 762–769.

[173] F. A. J. Verstraten, P. Cavanagh, and A. T. Labianca, "Limits of attentive tracking reveal temporal properties of attention," *Vision Research*, vol. 40, no. 26, pp. 3651 – 3664, 2000. [Online]. Available: http://www.sciencedirect.com/science/article/B6T0W-41WBMWC-9/2/9b649c00559d56168e300e968aa18874

[174] R. Vidal and D. Singaraju, "A closed form solution to direct motion segmentation," in *Computer Vision and Pattern Recognition*, 2005, pp. II: 510–515.

[175] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *IEEE Conference on Computer Vision and Pattern Recognition*, vol. 1, 2001, pp. 511–518.

[176] R. Visser, N. Sebe, and E. Bakker, "Object recognition for video retrieval," in *International Conference on Image and Video Retrieval*, 2002, pp. 250–259. [Online]. Available: citeseer.ist.psu.edu/visser02object.html

[177] E. Vul, M. Frank, G. Alvarez, and J. Tenenbaum, "Explaining human multiple object tracking as resource-constrained approximate inference in a dynamic probabilistic model," *Advances in Neural Information Processing Systems*, vol. 22, pp. 1955–1963, 2009.

[178] M. J. Wainwright, O. Schwartz, and E. P. Simoncelli, "Natural image statistics and divisive normalization: Modeling nonlinearities and adaptation in cortical neurons," in *Probabilistic Models of the Brain: Perception and Neural Function*, R. Rao, B. Olshausen, and M. Lewicki, Eds. Cambridge, MA: MIT Press, 2002, pp. 203–222.

[179] K. Walker, T. Cootes, and C. Taylor, "Locating salient object features," in *Proceedings British Machine Vision Conf.*, 1998, pp. 557–566.

[180] J. Y. A. Wang and E. H. Adelson, "Representing Moving Images with Layers," *The IEEE Transactions on Image Processing Special Issue: Image Sequence Compression*, vol. 3, no. 5, pp. 625–638, September 1994. [Online]. Available: citeseer.ist.psu.edu/wang94representing.html

[181] A. Watson and A. Ahumada, Jr, "Model of human visual-motion sensing," *Journal of the Optical Society of America A*, vol. 2, no. 2, pp. 322–341, 1985.

[182] L. Wixson, "Detecting salient motion by accumulating directionally-consistent flow," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 774–780, 2000. [Online]. Available: citeseer.ist.psu.edu/wixson99detecting.html

[183] J. M. Wolfe, "Guided search 2.0: A revised model of visual search," *Psychonomic Bulletin & Review*, vol. 1, no. 2, pp. 202–238, 1994.

[184] ——, "Visual search," in *Attention*, H. Pashler, Ed. UK: Psychology Press, 1998, pp. 13–74.

[185] C. R. Wren, A. Azarbayejani, T. Darrell, and A. Pentland, "Pfinder: Real-time tracking of the human body," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 780–785, 1997. [Online]. Available: citeseer.ist.psu.edu/wren97pfinder.html

[186] J. Xiao and M. Shah, "Motion layer extraction in the presence of occlusion using graph cut," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 10, pp. 1644–1659, Oct 2005. [Online]. Available: citeseer.ist.psu.edu/xiao04motion.html

[187] S. Yantis, "The neural basis of selective attention," *Current Directions in Psychological Science*, vol. 17, no. 2, p. 86, 2008.

[188] A. Yarbus, *Eye movements and vision*. New York: Plenum, 1967.

[189] A. Yilmaz, "Object tracking by asymmetric kernel mean shift with automatic scale and orientation selection," in *IEEE Conference on Computer Vision and Pattern Recognition, 2007.*, 2007, pp. 1–6.

[190] A. Yilmaz, O. Javed, and M. Shah, "Object tracking: A survey," *ACM Computing Surveys*, vol. 38, no. 4, p. 13, 2006.

[191] L. Zhang, M. Tong, and G. Cottrell, "Sunday: Saliency using natural statistics for dynamic analysis of scenes," in *Proceedings of the 31st Annual Cognitive Science Conference, Amsterdam, Netherlands*, 2009.

[192] L. Zhang, M. Tong, T. Marks, H. Shan, and G. Cottrell, "Sun: A bayesian framework for saliency using natural statistics," *Journal of Vision*, vol. 8, no. 7, 2008.

[193] J. Zhong and S. Sclaroff, "Segmenting foreground objects from a dynamic textured background via a robust Kalman filter," in *Proceedings of IEEE International Conference on Computer Vision*, vol. 1, 2003, p. 44.

[194] Y. Zhong, A. Jain, and M. Dubuisson-Jolly, "Object tracking using deformable templates," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 5, pp. 544–549, 2000.

[195] Q. Zhu, S. Avidan, and K.-T. Cheng, "Learning a sparse, corner-based representation for time-varying background modelling," in *International Conference on Computer Vision*, 2005. [Online]. Available: citeseer.ist.psu.edu/article/koller93robust.html

[196] Z. Zivkovic, "Improved adaptive Gaussian mixture model for background subtraction," in *International Conference on Pattern Recognition*, 2004.