SPIM

Thèse de Doctorat

école doctorale **sciences pour l'ingénieur et microtechniques**

UNIVERSITÉ DE BOURGOGNE

MODELES D'ATTENTION VISUELLE POUR L' ANALYSE DE SCENES DYNAMIQUES

Spatio-Temporal Saliency Detection in Dynamic Scenes Using Color and Texture Features

SATYA M. MUDDAMSETTY

SPIM

Thèse de Doctorat

école doctorale **sciences pour l'ingénieur et microtechniques**

UNIVERSITÉ DE BOURGOGNE

N°

THÈSE présentée par

SATYA M. MUDDAMSETTY

pour obtenir le

Grade de Docteur de
l'Université de Bourgogne

Spécialité : **Instrumentation et l'Informatique de l'Image**

MODELES D'ATTENTION VISUELLE POUR L' ANALYSE DE SCENES DYNAMIQUES

Spatio-Temporal Saliency Detection in Dynamic Scenes Using Color and Texture Features

Soutenue publiquement le 7 juillet 2014 devant le Jury composé de :

PHILLIPPE CARRÉ	Rapporteur	Professeur à l'Université de Poitiers
FRÉDÉRIC MORAIN-NICOLIER	Rapporteur	Professeur à l'Université de Reims
JOCELYN CHANUSSOT	Président du Jury	Professeur à l'Université de Grenoble
FABRICE MÉRIAUDEAU	Directeur de thèse	Professeur à l'Université de Bourgogne
ALAIN TRÉMEAU	Co-directeur de thèse	Professeur à l'Université de Jean Monnet
DÉSIRÉ SIDIBÉ	Co-directeur de thèse	Maître de conférences à l'Université de Bourgogne

SOMMAIRE

Contents iii

Résumé iv

Abstract v

Acknowledgements vii

List of Tables ix

List of Figures xiii

1 Introduction 1
- 1.1 Visual attention . . . 2
  - 1.1.1 Bottom-up attention approach . . . 2
  - 1.1.2 Top-down attention approach . . . 2
- 1.2 The neurology background of early vision and attention . . . 3
- 1.3 Human visual system . . . 3
  - 1.3.1 The eye structure . . . 4
  - 1.3.2 The Lateral Geniculate Nucleus (LGN) . . . 5
  - 1.3.3 Visual cortex . . . 5
  - 1.3.4 Primary visual cortex (V1) . . . 6
  - 1.3.5 Area V2 . . . 7
  - 1.3.6 Area V3 . . . 8
  - 1.3.7 Area V4 . . . 8
  - 1.3.8 Middle temporal area (MT or V5) . . . 8
  - 1.3.9 Inferotemporal cortex (IT) . . . 9
- 1.4 The background of visual attention mechanism in human brain . . . 9
- 1.5 Theories of visual attention . . . 10
  - 1.5.1 Feature Integration Theory (FIT) . . . 10
  - 1.5.2 Guided Search Theory (GST) . . . 11
  - 1.5.3 Additional theories . . . 12
    - 1.5.3.1 Attentional Engagement Theory (AET) . . . 12
    - 1.5.3.2 Boolean Map Theory (BMT) . . . 13
- 1.6 Thesis motivation . . . 14
- 1.7 Contributions and thesis structure . . . 14

2 State of the Art 17
- 2.1 Introduction . . . 17
- 2.2 Biologically inspired models of visual attention . . . 17
  - 2.2.1 Koch and Ullman model of neural architecture . . . 17
  - 2.2.2 Itti's model . . . 18

2.2.3 Marat's model . . . 20
2.2.4 Le Meur's model . . . 22
2.3 Computational models of visual attention . . . 23
2.3.1 Graph Based Visual Saliency (GBVS) model . . . 23
2.3.2 Phase spectrum of Quaternion Fourier Transform (PQFT) based saliency . . . 24
2.3.3 Saliency using natural statistics for dynamic scenes . . . 26
2.3.4 Phase discrepancy model . . . 26
2.3.5 Self-resemblance model . . . 27
2.3.6 Spatio-temporal saliency using dynamic texture . . . 27
2.3.7 Information-theoretic model . . . 29
2.3.8 Icremental Coding Length (ICL) saliency model . . . 29
2.3.9 Mancas' saliency model . . . 31
2.4 Applications of visual saliency . . . 31
2.4.1 Salient object segmentation . . . 32
2.4.2 Image re-targeting . . . 32
2.4.3 Object tracking . . . 33
2.4.4 Video compression . . . 33
2.4.5 Robotic navigation and localization . . . 34
2.5 Discussion on state-of-the-art methods . . . 34

3 Proposed spatio-temporal saliency models 37
3.1 Introduction . . . 37
3.2 Spatio-temporal saliency detection based on optical flow . . . 37
3.2.1 Static saliency map computation . . . 38
3.2.2 Dynamic saliency map using optical flow . . . 40
3.2.2.1 Step 2a : Dominant motion compensation . . . 40
3.2.2.2 Step 2b : Robust estimation of parameters . . . 41
3.2.2.3 Step 2c : Local motion estimation using dense optical flow 42
3.2.2.4 Step 2d : Temporal filtering and normalization . . . 44
3.2.3 Limitations of optical flow . . . 44
3.3 Spatio-temporal saliency model using local binary patterns . . . 45
3.3.1 Basic local binary patterns texture descriptor . . . 45
3.3.2 Spatio-temporal LBP . . . 46
3.3.3 Spation-temporal saliency using LBP-TOP descriptor . . . 47
3.3.4 Spatio-temporal saliency using color and texture features . . . 47
3.3.4.1 Post-processing . . . 48
3.4 About Fusion . . . 58
3.4.1 Different Fusion Techniques . . . 58

4 Experiments and Results 61
4.1 Experimental Setup . . . 61
4.2 Evaluation Dataset . . . 62
4.2.1 ASCMN dataset . . . 62
4.2.2 SVCL dataset . . . 62
4.3 Evaluation Measures . . . 63
4.3.1 Receiver Operating Characteristics (ROC) . . . 64
4.3.2 Normalized Scanpath Salience (NSS) . . . 64
4.3.3 Kullback-Leibler Divergence (KL) . . . 65

4.4 Evaluation of Fusion Techniques 66
4.5 Spatio-temporal saliency detection with LBP features 76
4.6 Comparison with eye-tracking data 81
4.6.1 Comparison based on evaluation metrics 81
4.6.2 Comparison based of video classes 84
4.7 Computation time details 84

5 Conclusion and Future perspectives **91**
5.1 Conclusion of the thesis 91
5.2 Future perspectives 93

Appendices **95**

References **107**

RÉSUMÉ

De nombreuses applications de la vision par ordinateur requièrent la détection, la localisation et le suivi de régions ou d'objets d'intérêt dans une image ou une séquence d'images. De nombreux modèles d'attention visuelle, inspirés de la vision humaine, qui détectent de manière automatique les régions d'intérêt dans une image ou une vidéo, ont récemment été développés et utilisés avec succès dans différentes applications. Néanmoins, la plupart des approches existantes sont limitées à l'analyse de scènes statiques et très peu de méthodes exploitent la nature temporelle des séquences d'images.

L'objectif principal de ce travail de thèse est donc l'étude de modèles d'attention visuelle pour l'analyse de scènes dynamiques complexes. Une carte de saliance est habituellement obtenue par la fusion d'une carte statitque (saliance spatiale dans une image) d'une part, et d'une carte dynamique (salience temporelle entre une série d'image) d'autre part. Dans notre travail, nous modélisons les changements dynamiques par un opérateur de texture LBP-TOP (Local Binary Patterns) et nous utilisons l'information couleur pour l'aspect spatial.

Les deux cartes de saliances sont calculées en utilisant une formulation discriminante inspirée du système visuel humain, et fuionnées de manière appropriée en une carte de saliance spatio-temporelle.

De nombreuses expériences avec des bases de données publiques, montrent que notre approche obteint des résulats meilleurs ou comparables avec les approches de la littérature.

ABSTRACT

Visual saliency is an important research topic in the field of computer vision due to its numerous possible applications. It helps to focus on regions of interest instead of processing the whole image or video data. Detecting visual saliency in still images has been widely addressed in literature with several formulations. However, visual saliency detection in videos has attracted little attention, and is a more challenging task due to additional temporal information. Indeed, a video contains strong spatio-temporal correlation between the regions of consecutive frames, and, furthermore, motion of foreground objects dramatically changes the importance of the objects in a scene. The main objective of the thesis is to develop a spatio-temporal saliency method that works well for complex dynamic scenes.

A spatio-temporal saliency map is usually obtained by the fusion of a static saliency map and a dynamic saliency map. In our work, we model the dynamic textures in a dynamic scene with Local Binary Patterns (LBP-TOP) to compute the dynamic saliency map, and we use color features to compute the static saliency map. Both saliency maps are computed using a bio-inspired mechanism of Human Visual System (HVS) with a discriminant formulation known as center surround saliency, and are fused in a proper way.

The proposed models have been extensively evaluated with diverse publicly available datasets which contain several videos of dynamic scenes. The evaluation is performed in two parts. First, the method in locating interesting foreground objects in complex scene. Secondly, we evaluate our model on the task of predicting human observers fixations. The proposed method is also compared against state-of-the art methods, and the results show that the proposed approach achieves competitive results.

In this thesis we also evaluate the performance of different fusion techniques, because fusion plays a critical role in the accuracy of the spatio-temporal saliency map. We evaluate the performances of different fusion techniques on a large and diverse complex dataset and the results show that a fusion method must be selected depending on the characteristics, in terms of color and motion contrasts, of a sequence. Overall, fusion techniques which take the best of each saliency map (static and dynamic) in the final spatio-temporal map achieve best results.

ACKNOWLEDGEMENTS

I would like to thank my director of thesis Prof. Fabrice Mériaudeau and Co-directors Prof. Alain Trémeau and Ass prof. Désiré Sidibé for giving me this PhD opportunity under their supervision. Their constant support and expert guidance helped me for the completion of my thesis. Their expertise in computer vision field have significantly improved my quality of research.

I also like to thank to all my colleagues in the le2i laboratory, Le creusot, France for being so friendly and helpful during the wonderful period of my PhD. I extend my gratitude to my best friend Ashutosh for his moral support when I am in difficult situation.

Finally, I will never forget the support given by my parents (Ranganadham and Anu Radha) and my elder and younger brothers Sathya Naresh Babu Muddamsetty and Sai Prasad Muddamsetty throughout my life in Europe. I sincerely dedicate my PhD thesis to the goddess whom I truly believe, Without her blessings I will never achieve this. I also dedicate this to my loving parents and my brothers.

LISTE DES TABLES

4.1 ASCMN dataset video types 62
4.2 Details of ASCMN dataset 62
4.3 Details of SVCL dataset 63
4.4 Fusion techniques evaluation results. Mean (Mean fusion), Max (Max fusion), AND (Multiplication fusion), MSF (Maximum skewness fusion), BTF (Binary thresholded fusion), DWF (Dynamic weight fusion), MPF (Motion priority), ITF (Information theory fusion), SIF (Scale invariant fusion). 72
4.5 Evaluation of spatio-temporal saliency detection methods. PROPOSED (with color and LBP features), LBP-TOP (LBP features only), OF (Optical Flow based), SR (Self-Resemblance) and PD (Phase Discrepancy). 77
4.6 Evaluation of spatio-temporal saliency detection methods using Arear Under Curve (AUC). PROPOSED (with color and LBP features), LBP-TOP (LBP features only), ICL [31], SUN [74], MANCAS [50] and SR [64]. 85
4.7 Evaluation of spatio-temporal saliency detection methods using Normalized Scanpath Saliency (NSS). PROPOSED (with color and LBP features), LBP-TOP (LBP features only), ICL [31], SUN [74], MANCAS [50] and SR [64]. 86
4.8 Evaluation of spatio-temporal saliency detection methods using Kullback Leibler-Divergence (KL). PROPOSED (with color and LBP features), LBP-TOP (LBP features only), ICL [31], SUN [74], MANCAS [50] and SR [64]. . 87
4.9 Results of whole dataset of videos with three evaluation metrics 87
4.10 Computation time details metrics 87
4.11 Results for the 5 classes of videos with the three evaluation metrics. 88

TABLE DES FIGURES

1.1 Bottomp-up approach experiments (Source : [24]) 2
1.2 Top-down approach experiments (Source : [24]) 3
1.3 The visual information processing in Human Visual System (source : [26]). 4
1.4 (a) The structure of eye ; (b) The laminar structure of the retina which shows the first five layers. The light must pass through these layers to reach the photoreceptors, which are situated at the back of the retina (source : [61]) . 5
1.5 Center-surround mechanism in retina (source : [61]). 6
1.6 Visual cortex region shown in human brain (source : [2]) 7
1.7 Simple flow chart of visual cortex (source : [8]) 8
1.8 Feature Integration Theory overview(Source : [68]) 11
1.9 Examples of Feature Integration Theory 11
1.10 Example studies vs. contributions vs. papers 12
1.11 A schematic overview of Boolean map Theory(Source : [32]) 13

2.1 Koch and Ullman architecture (Source : [39]) 18
2.2 Itti's model [35] 19
2.3 Marats's model [51] 21
2.4 Le Meur model [42] 22
2.5 Self-resemblance model (source : [64]) 28
2.6 Information theoretic's model (source : [16]) 30
2.7 Example of image segmentation using visual saliency. From left to right : original image, obtained saliency map, mean-shift segmentation, and segmentation based on saliency values (last two columns). Source [3]. 32
2.8 Image re-targeting using salient object detection methods. From left to right : Saliency map, detected image regions to remove, re-targeted image, and original image. Source [6]. 33
2.9 Video compression procedure using salieny detection method 34

3.1 Spatio-Temporal saliency using optical flow. 38
3.2 Examples of static saliency detection. (a) Original frame (b) Static saliency map. 49
3.3 Dominant motion estimation on two successive video frames. Top row : original video frames. Bottom left : warped image with estimated affine parameters ; Bottom right : detected relative motion. 50
3.4 The dense optical flow estimation on successive video frames. From left to right : original video frames and motion estimation of each pixel. 50
3.5 Dynamic saliency map estimation. Top row : original video frames ; Bottom left : median filtered image motion map ; Bottom right : normalized motion map. 51
3.6 Examples of dynamic saliency detection. (a) Original frame (b) Dynamic saliency 52

3.7 Examples of dynamic saliency maps failures using optical flow in complex dynamic background scenes. 53
3.8 Spatio-Temporal saliency overview. 53
3.9 Basic LBP description. 54
3.10 Circular sampling.(source : [58]) 54
3.11 LBP-TOP computation using three orthogonal planes(source : [58]). 54
3.12 Examples of saliency maps using LBP-TOP. From left to right : original video frame, saliency map in XY plane, in XT plane, in YT plane, and final spatio-temporal saliency map. 55
3.13 Examples of spatio-temporal saliency detection with LBP-TOP features. (a) Original frame ; (b) saliency map in XY plane ; (c) saliency map in XT plane ; (d) saliency map in YT plane ; (e) fused spatio-temporal saliency map. 56
3.14 Combined model with color and texture features. 57
3.15 Examples of spatio-temporal saliency detection map using color and texture features. (a) Original frame (b) Spatio-temporal saliency map. 60

4.1 Sample videos of the 5 different classes in ASCMN database. From top to bottom : ABNORMAL motion with bikes and cars moving faster than people ; SURVEILLANCE videos ; CROWD motion with increasing density from left to right ; MOVING camera videos ; and NOISE. 68
4.2 Sample images of ground truth data. The left column shows aggregated eye tracking results : each dot is the position of the eye gaze of video. The middle column shows the heatmaps and in the right column thresholded version of the heatmap is shown. 69
4.3 Examples of videos and ground truth from SVCL dataset. 70
4.4 Confusion and Common performance metric calculated from it 71
4.5 Visual comparison of different fusion techniques. From left to right : original video frame followed by Mean (Mean fusion), Max (Max fusion), AND (Multiplication fusion), MSF (Maximum skewness fusion), BTF (Binary thresholded fusion), DWF (Dynamic weight fusion), MPF (Motion priority), ITF (Information theory fusion), SIF (Scale invariant fusion). 73
4.6 Example of salient region segmentation with the *Skiing* sequence. From left to right : input frame ; detection with *Binary Threshold* and with *Motion Priority* fusion techniques. Red box indicates ground truth and green box indicates the detected salient region. 74
4.7 Quantitative fusion comparision with *Cyclists* sequence. 74
4.8 Quantitative fusion comparision with *Freeway* sequence. 74
4.9 Examples of static and dynamic saliency detection on good color contrast and motion contrast boats sequence. From left to right : *Original frame*, *Static saliency* and *Dynamic saliency*. 75
4.10 Influence of λ on the performance of the proposed method. 76
4.11 Visual comparison of spatio-temporal saliency detection of our methods and state of art methods. (a) Original frame ; (b) PROPOSED ; (c) LBP-TOP ; (d) OF [52] ; (e) SR [64] and (f) PD [77] 78
4.12 Quantitative comparision with *Boats* sequence. 79
4.13 Quantitative comparision with *Freeway* sequence. 79
4.14 Quantitative comparision with *Ocean* sequence. 80

4.15 Preprocessing of saliency maps using morphology and Gaussian filtering. From top to bottom, saliency maps obtained with PROPOSED, LBP-TOP, ICL, SUN, SR and MANCAS methods. 82
4.16 Results on the ASCMN dataset using AUC metric. 83
4.17 Results on the ASCMN dataset using NSS metric. 83
4.18 Results on the ASCMN dataset using KL metric. 84
4.19 Results for different classes using AUC metric 89
4.20 Results for different classes using NSS metric 89
4.21 Results for different classes using KL metric 89
4.22 Visual comparison of spatio-temporal saliency detection of our methods and state of art methods on ASCMN dataset. (a) Original frame ; (b) PROPOSED ; (c) LBP-TOP ; (d) ICL [31] ; (e) SUN [74] ; (f) MANCAS [50] and (g) SR [64] 90

1 Sample images of ABNORMAL class of ASCMN dataset is shown. 100
2 Sample images of SURVEILLANCE class of ASCMN dataset is shown. . . 101
3 Sample images of CROWD class of ASCMN dataset is shown. 102
4 Sample images of MOVING class of ASCMN dataset is shown. 103
5 Sample images of NOISE class of ASCMN dataset is shown. 104
6 Sample images of SVCL dynamic scenes dataset is shown. 105
7 Sample images of SVCL dynamic scenes dataset is shown. 106

1

INTRODUCTION

Visual attention is an important research topic in computer vision fields. The word attention is often used in our daily life. For instance, when we go to a new place or participate to an event, if we find something different like some people dancing or dressed differently these cues will automatically catch our attention. A motionless object which has a sudden motion or a sudden noise in a meeting room will also grab our attention. The mechanism in the brain that determines which part of the multitude of sensory data is currently of most interest is called selective attention. It is basically a process to detect a scene region which is different from the surroundings. In our thesis we will mainly focus on visual attention applied to images and videos.

Attention is one of the useful concepts for humans in their daily life and it holds an important place in computer vision applications. For example, consider a visual scene which contains many objects with various visual characteristics such as shape, color, size and texture. Some of the objects might be moving while others are static. Despite the huge amount of available information, the visual information reaching our eyes is limited as we cannot acquire the whole scene at a time. Thus we perceive only a small part of the visual field and the remaining part looks blurry to us. This smaller part of the visual field is perceived clearly with maximum acuity. It is the part of the visual scene that have more attraction than its surroundings. Understanding this mechanism is an active research area in cognitive sciences.

In the field of computer vision, the detection and recognition of objects is one of the most important and yet hardest problems. The main challenging problem for researchers - the development of a system which is able to match the human ability to recognize thousands of objects from different view points, under various illumination conditions and with occlusions - seems to lie remotely in future. Therefore, to improve the performance of detection methods researchers try to mimic simulate biological vision systems mechanisms. In robotics applications for instance, visual attention task is important for autonomous robots that act in complex and unknown environments. Focussing on the relevant data is even more important for pure vision systems. Due to this application, there have been an increase of interest in computer vision and robotics in the last decades to adapt this human *selective attention* to computational systems, with computational implementations that can work in real-time.

The goal of this chapter is to describe visual attention mechanisms and the different approaches that have been proposed to model it (section. 1.1). The human visual system and neurological background of visual attention are described in section. 1.2 and section. 1.3. Next, we introduce several attention theories in section. 1.5. Finally, we present

our thesis motivation, goals and the thesis structure are presented.

1.1/ VISUAL ATTENTION

Visual attention intuitively characterizes parts of a scene which stand out relative to their neighbouring parts and grab our attention. It is an important and fundamental research problem in neuroscience and psychology to investigate the mechanisms of human visual system in selecting the regions of interest. For the human visual system, it is impossible to perceive the entire scene simultaneously in the same sensory act. Only a small region of the scene is analysed in detail at each moment. The region that is currently attended will not always be the same region that is fixated by the eyes and the gaze will shift to the next interesting region and this process continues until the whole scene is perceived. The order in which the scene is investigated is determined by the mechanism of selective attention. Visual attention is generally processed in two approaches which are bottom-up approach and top-down approach. They are briefly described in the following sections. The detailed explanations of these two approaches are mentioned in [33]

1.1.1/ BOTTOM-UP ATTENTION APPROACH

Bottom-up attention approach is a stimulus driven derived solely from the conspicuousness of regions in a visual scene. This attentional mechanism is also called exogenous, automatic, reflexive or peripherally cued [59]. Bottom-up approach is more thoroughly investigated than top-down attention approach because the data-driven stimuli are easier to control than cognitive factors such as knowledge and expectations [24]. As an example in the Fig. 1.1 the human subject has to search the diamond shape, but the circle which is red in color immediately grabs his attention.

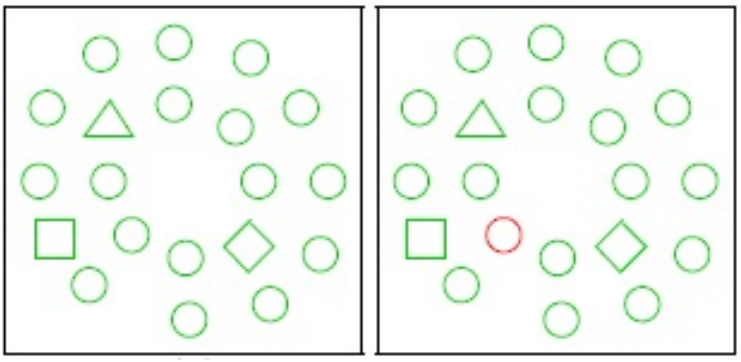


FIGURE 1.1 – Bottomp-up approach experiments (Source : [24])

1.1.2/ TOP-DOWN ATTENTION APPROACH

Top-down attention approach is goal driven, i.e. by expectations [24]. For example, car holders are more likely to see the petrol station in a street. Top-down attention refers to voluntary allocation of attention to certain features, objects or regions in space [59].

In psychophysics, top-down influences are investigated by the so-called cueing experiments. The cue directs the eye to the location of interest, whereas in the bottom-up attention it is attracted to the location. The main aspect of top-down attention is visual search, which means given a target and an image, find an instance of the target in the image. Visual search is very common in everyday life, for instance finding a friend in a crowd. Fig. 1.2 shows the cueing experiments, for the given cue (left) one has to search for it in the test image (right).

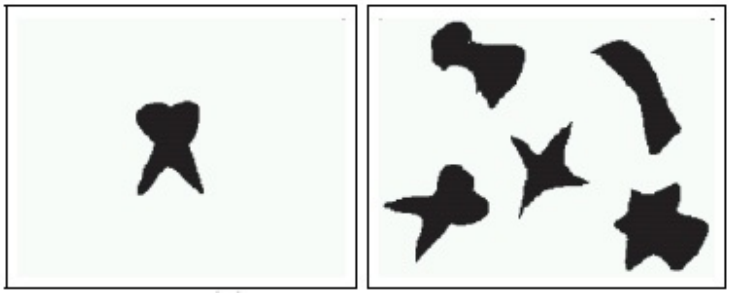


FIGURE 1.2 – Top-down approach experiments (Source : [24])

1.2/ THE NEUROLOGY BACKGROUND OF EARLY VISION AND ATTENTION

We know that visual attention takes place in the visual cortex of the human brain. The visual system is the most complex of all sensory systems and a large part of the human brain is involved in vision [66]. Due to its complexity, the study of the visual system has attracted the effort of a large number of scientists and significant progresses have been achieved in the past decades [66]. The study of the visual system allows to identify common principles of sensory information processing. The empirical findings in neurobiology and physiology, serve as motivation and guidelines for computational models of visual processing. Visual attention is a concept of human perception, so it is worth knowing more about the human visual system to get detailed information of the nature of the visual attention concepts. In section 1.3 we describe the Human Visual System (HVS). Then, in section 1.4 we briefly explain how the visual attention mechanism occurs in the human brain.

1.3/ HUMAN VISUAL SYSTEM

Before stepping into the visual attention mechanism it is necessary to recall some basics about the human visual system (HVS). HVS is a very complex system not fully understood yet. Vision is one of our most delicate and complicated senses and it is also studied intensively. According to [53], one-fourth of the brain is involved in visual information processing. More is known about vision than any other vertebrate sensory system and all this information is derived from studies of monkeys and cats. Some basic parts of the HVS are explained briefly in the next sections. The main basic parts which play major role in HVS are *eyes*, *lateral geniculate nucleus* (LGN), *visual cortex* (V1,V2,V3,V4), *middle temporal*

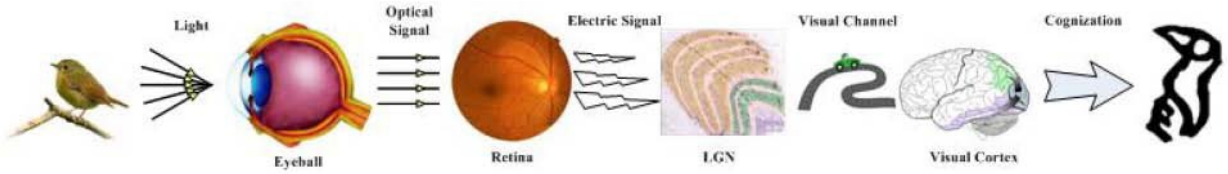


FIGURE 1.3 – The visual information processing in Human Visual System (source : [26]).

area (MT) and *inferotemporal cortex* (IT). The studies of visual physiology show that visual information processing involves four aspects : optical processing, retina processing, lateral geniculate processing and visual cortex processing [26]. The visual information processing in the HVS is illustrated in the Fig. 1.3. In the following subsections we discuss the processing in detail.

1.3.1/ THE EYE STRUCTURE

The eyes are the human organs for producing an image of the real world to the brain. Primates including humans, have sophisticated vision using two eyes, which is known as binocular vision. Visual signals pass from each eye along the millions or so fibers of the optic nerve to the optic chiasm, where some nerve fibers cross over so that both sides of brain receive signals from both eyes [53]. The left halves of both retinas project to the left visual cortex and the right halves project to the right visual cortex [53].

The retina is located in the inner surface of the eye. It contains photoreceptors that perceive only part of the visible light spectrum. There are about 125 million photoreceptors in each eye and they are specialized to turn light into electrical signals. These photoreceptors are classified into rods and cones. Rods are more sensitive to dim light and do not convey color whereas cones work in bright light and are responsible for acute detail, black and white vision and color vision. The photoreceptors are connected to horizontal and bipolar cells [8]. A bipolar cell can receive connections directly from the photoreceptors or indirectly through the horizontal cell. The output of the bipolar cells is integrated in the ganglion cells. These entire cells are responsible for the photoelectric conversion and information transmission. The structure of the eye is shown in Figure 1.4.

The processing which is done in the eye is known as optical processing and retina processing. The direct connections can be either excitatory or inhibitory. The integration of these two inputs form a center/surround response in the bipolar cell and the output of the bipolar cell is integrated in the ganglion cells which also has center-surround design with an inner circular center and a surrounding ring as illustrated in the Fig. 1.5. The ON-center/OFF-surround ganglion cells are excited when illuminated in their center and inhibited when illuminated in their surround. The OFF-center/ON-surround cells exhibit the opposite pattern. From the Fig. 1.5 (A) we can observe that ON-center cells respond maximally when the center is illuminated and the surround is not (row 3), and respond minimally in the opposite illumination condition (row 4). A partial illumination of center (row 1) or surround (row 2) results in brief excitation and inhibition, respectively. Fig. 1.5 (B) shows functioning of the OFF-center/ON-surround cells.

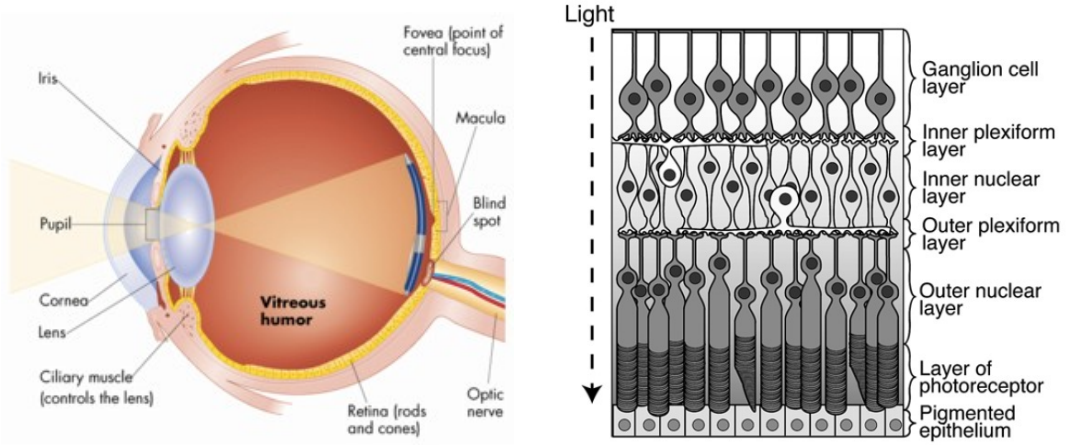


FIGURE 1.4 – (a) The structure of eye ; (b) The laminar structure of the retina which shows the first five layers. The light must pass through these layers to reach the photoreceptors, which are situated at the back of the retina (source : [61])

.

1.3.2/ THE LATERAL GENICULATE NUCLEUS (LGN)

The lateral geniculate nucleus (LGN) in the thalamus is one of the major targets of the retinal ganglion cells. LGN receives inputs from both eyes and relays these messages to the primary visual cortex (V1) via the optic radiation. The LGN is divided into six layers. The four dorsal layers receive information from P ganglion neurons and are known as *parvocellular* layers. The two ventral layers receive information from M ganglion neurons and are known as *multicellular* layers [8]. Each of the six LGN layers receives input from either the ipsilateral (same side) or contralateral (opposite side), which means the ganglion cells from the left eye project to layer 1, 4, 6 of the right LGN, and the right eye ganglion cells project to its layer 2, 3 and 5 [1]. Most of the LGN neurons have a circular centre-surround structure and they differ in sensitivity in four characteristics that are colour, contrast sensitivity, spatial resolution and latency.

The parvocellular pathway is believed to be responsible for shape and object perception while the magnocellular pathway is believed to be responsible for perception of motion and sudden changes [8]. The transmission path from eye to LGN is one way transmission. However, there are connections from the LGN to the cortex and from the cortex to the LGN. Understanding of the complete role of LGN is still in progress.

1.3.3/ VISUAL CORTEX

Physiological studies of the visual cortex show two different pathways with some connections between them. They are classified based on their functions. One is the *occipito-temporal pathway* which contains visual areas (V1, V2, and V4), and the anterior and posterior inferotemporal cortex (AIT and PIT). This pathway is related with object recognition features like color and shape. The other one is the *occipitoparietal pathway* which contains visual areas (V1, V2, and V3) and the middle temporal and medial superior temporal (MT and MST) visual areas [8]. This pathway is related with spatio-temporal characteristics of the scene such as motion direction. In this section we will give a brief

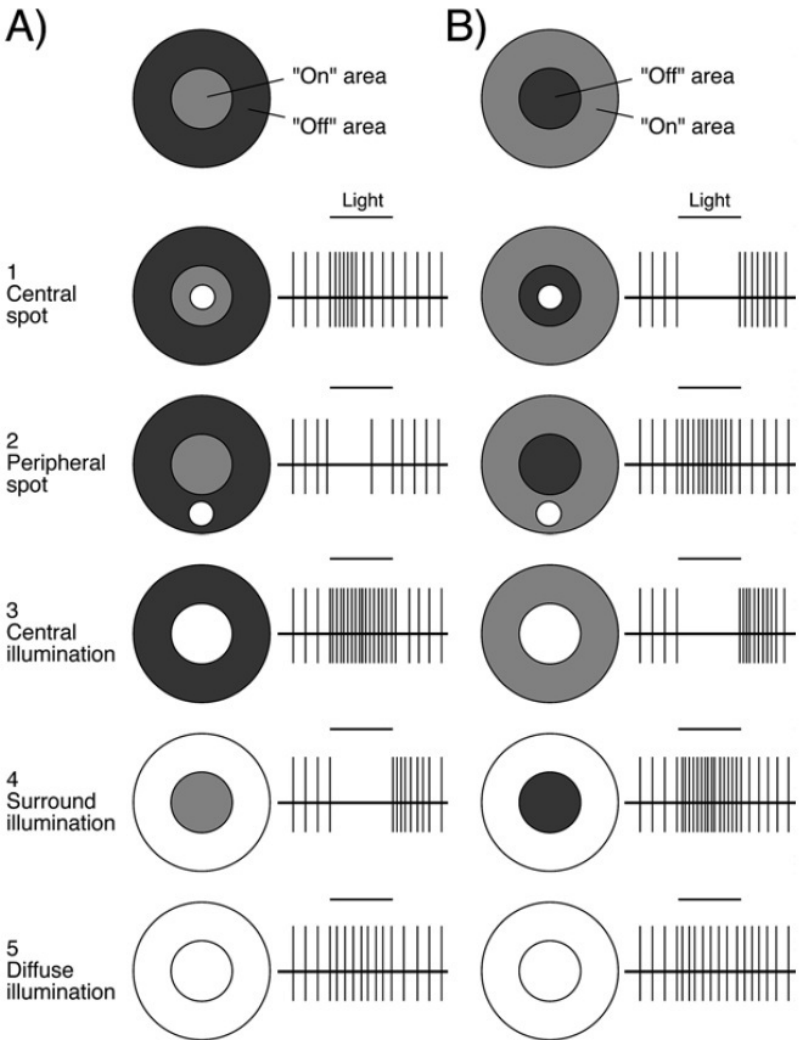


FIGURE 1.5 – Center-surround mechanism in retina (source : [61]).

overview of the most important areas which are related to object detection and perception. The different areas of the the visual cortex are shown in Fig. 1.6 and the connections between them are shown in Fig. 1.7.

1.3.4/ PRIMARY VISUAL CORTEX (V1)

During the year 1959, David Hubel and Torsten Wiesel studied the response of neurons in V1 [37]. They discovered that cats V1 neurons respond when bars and edges are presented in the receptive field. The same characteristics were later found in monkeys [8]. The neurons in V1 are classified into three types : simple cells, complex cells and end-stopped cells. Simple cells are located near the fovea which is the center most part of the eye located in center of the retina and has small receptive field (0.25-1 deg). These cells respond to bars, edges with different orientations and to spatial frequency. They can be modelled by Gabor filters. Complex cells are also sensitive to bars and orientations, but they are less sensitive than simple cells and they are directed by stimulated motion. End-stopped neurons/cells require the termination of edges or bars located in the receptive

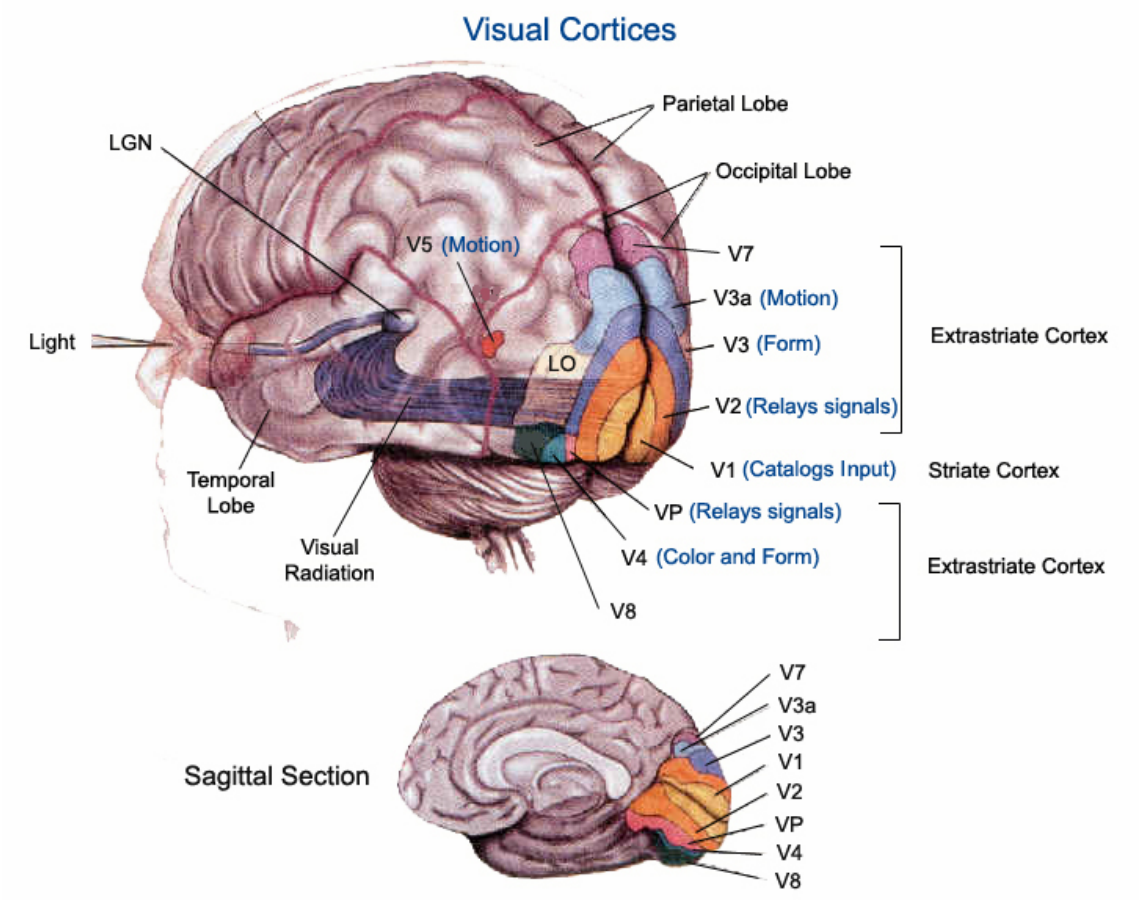


FIGURE 1.6 – Visual cortex region shown in human brain (source : [2])

field in order to respond. Finally V1 projects mainly to V2, and also to MT and V3 areas.

1.3.5/ AREA V2

Information is projected from V1 to V2 through three pathways : stereopsis and motion (magnocellular neurons), color (parvocellular neurons) and form (parvocellular neurons). The color paths in V2 are not orientation selective and more than half of them are color sensitive with center-surround antagonism. The neurons in the form pathway have selectivity to orientation, and many of them respond to termination of edges or bars. Although V2 stereopsis pathway also has orientation selectivity, they respond to orientations and bars terminations occasionally. For this reason, these neurons seem to be mainly selective for stereoscopic depth and motion [8]. The projections from V1 and V2 areas are directed mainly to the higher areas through the dorsal area MT to the parietal lobe or parietal cortex and also the ventral path of area V4 to inferotemporal cortex (IT).

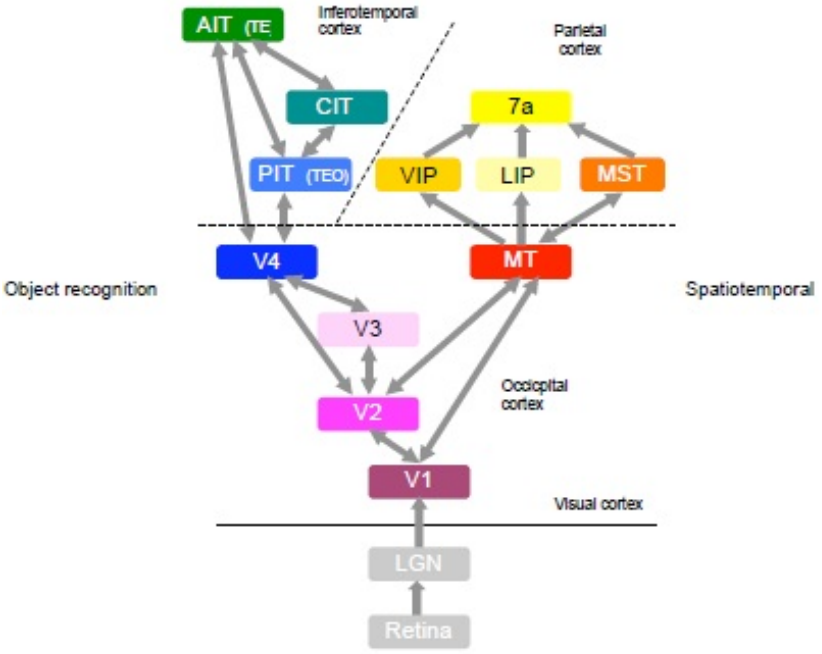


FIGURE 1.7 – Simple flow chart of visual cortex (source : [8])

1.3.6/ AREA V3

Visual cortex area V3 receives input from layers of V1 and from thick stripes of V2. V3 represents the lower visual field, and the majority of its neurons are orientation selective and exhibit high contrast selectivity.

1.3.7/ AREA V4

The visual area V4 contains many wavelength-selective cells. It also has the cells whose receptive field properties correspond to a particular perceived color, i.e these cells will mediate color constancy [61]. V4 receives input not only from the blob region of V1 and the thin-stripe regions of V2 but also from the interblob regions of V1 and the interstripe region of V2. The inputs of V4 suggest that this area in not restricted to color processing. Although the majority of cells in V4 are wavelength and color selective, some are orientation selective and some are both color and orientation selective as mentioned in [61]. So the visual area V4 is involved in both color and form processing.

1.3.8/ MIDDLE TEMPORAL AREA (MT OR V5)

The middle temporal area (MT) is one of the main area in the visual cortex and deals with motion processing. The motion selective area MT is located in the middle temporal sulcus, close to the junction of the occipital temporal and parietal lobes [61]. MT receives direct input from the layer 4B of V1 and from V2 which forms a magnocellular channel. The middle temporal area (MT) and medial superior temporal visual area (MST) in the

parietal cortex play a central role in the perception of motion [37]. Indeed, over 80% of neurons in MT respond selectively to a particular direction of motion of a stimulus.

1.3.9/ Inferotemporal cortex (IT)

The inferotemporal cortex (IT) receives mainly inputs from V2 and V4 and is believed to be the main area involved in object recognition and discrimination. As explained in [8] IT is divided into two main parts, which are posterior IT (PIT) and anterior IT (AIT). In PIT, a huge number of neurons is activated by a simple combination of features such as bars or disks varying in size, orientation or color. These neurons are called *primary cells*. AIT comprising of three-quarters of the IT area and the neurons which are required for complex features for maximal activation are known as *eloborate cells*. In [8], it is explained that neurons in IT respond to complex shapes and to the combination of such shapes with color or textures. Thus, IT is involved in filtering the objects to which we attend and foveate in visual search tasks.

1.4/ The background of visual attention mechanism in human brain

The visual attention mechanism in the human brain is still an unsolved problem in neuroscience research. Most of the researches of neuro-physiological finding on visual attention point out that there is no single brain area guiding the attention and nearly all brain areas are associated with visual information processing [23]. Most of our knowledge about the organization of visual cortex comes from behavioral, anatomical, and physiological studies of monkeys and these studies have shown that monkey cortex contains more than 30 separate visual areas [69]. The evidence for attention in physiology was first found in Superior Colliculus (SC) and later it was found in other parts such as posterior parietal cortex (PP), Visual Cortex areas V2,V4 and also in Infero Temporal area (IT) [8].

In [60], the authors explain that the three major functions concerning attentions are orienting of attention, target detection and alertness. The authors claim that the orienting of attention to a salient stimulus is carried out by the interaction of three areas in PP which are responsible for disengaging the focus of attention from its present location. SC shifts the attention to a new location, and the pulvinar is specialized in reading out the data from the indexed location. This combination of systems is called a *posterior attention system*. In the second attentional function, the detection of a target is carried out by *anterior attention system*.

In [69] the authors say that the occipitotemporal pathway or ventral stream is crucial for the identification of objects. The occipitoparietal or dorsal stream is crucial for the appreciation of the spatial relation among objects and the visual guidance of movements towards object in space. The results from functional brain imaging studies suggest the idea that top-down signals related to directed attention are generated by a distributed network of areas in frontal and parietal cortex. A network consisting of areas in the superior parietal lobule, the frontal eye field, and the supplementary eye field extending into the anterior cingulate cortex has been found to be active in a visio-spatial tasks.

It is shown in [12] that the frontal eye field is the place where a kind of saliency map is located which derives information from bottom-up and top-down influences. The authors in [11] explain that the bottom-up attentional capture, mediated by stimulus salience and/or relevance, is subserved by the temporoparietal junction. When subjects attend to and monitor a change in either a visual or an auditory stimulus presented simultaneously, blood oxygenation level dependent activation of the TPJ region of the right parietal lobe is enhanced and when the source of the attentional signal is top-down or goal-directed, the superior parietal lobule region is engaged.

Several areas have been identified has being involved into attentional process. However, the behaviour of each area of the brain as well as the interplay among them still remains an open question.

1.5/ THEORIES OF VISUAL ATTENTION

According to psychological studies there are wide varieties of theories in visual attention. The main objective of these theories is to explain the better understanding of human perception. In this section we explain some of the popular attention theories which are commonly used to design a bio-inspired attention model. The first, introduced in section 1.5.1, is the Feature Integration Theory of Treisman and the second one is the Guided Search Theory of Wolfe described in section 1.5.2.

1.5.1/ FEATURE INTEGRATION THEORY (FIT)

This theory has been widely used in modelling the visual attention mechanism. The theory was proposed by Treisman and Gelade [68]. According to FIT, different features such as color, orientation and spatial frequency, are registered early, automatically and in parallel across the visual field, a step known as pre-attentive mode. Objects are identified separately in a later stage known as attentive mode, which requires focussed attention.

It is assumed that the visual scene is initially coded with a number of separable dimensions, such as color, orientation, spatial frequency, brightness and direction of movement. In order to recombine these separate representations and to ensure the correct synthesis of features for each object in a complex display, stimulus locations are processed serially with focal attention. Any features which are present in the same fixation of attention are combined to form a single object. The focal attention provides the glue which integrates the initially separate features into unitary object shown in a master map. The FIT procedure is shown in Fig. 1.8.

One of the main statements of the FIT is that the target is detected easily, fastly and in parallel. This is known as *pop-out* effect, i.e. the target differs from the distractors in exactly one feature and the distractors are homogeneous. If it differs in more that one feature then the focal attention is required in serial search. In the example shown in Fig. 1.9 we can clearly see the pop-out effect which is also known as parallel search. In the first image (left) we can clearly see that red line has a pop-out effect, because it is totally different from its distractors (green lines). In the second image (right) the line which is vertical has a pop-out effect which this time is due to orientation.

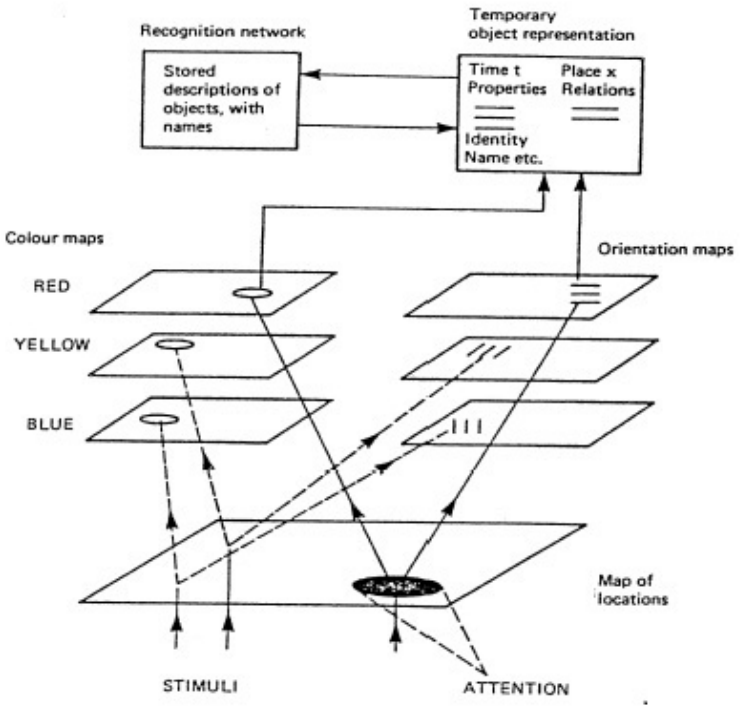


FIGURE 1.8 – Feature Integration Theory overview(Source : [68])

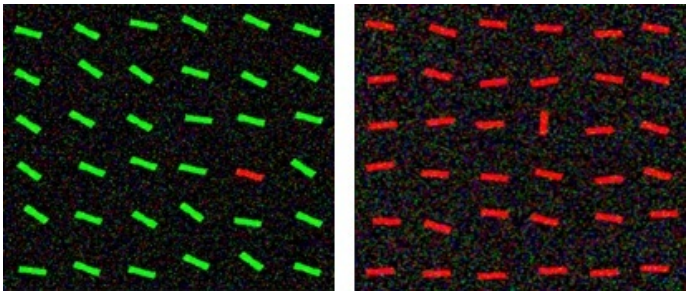


FIGURE 1.9 – Examples of Feature Integration Theory

1.5.2/ GUIDED SEARCH THEORY (GST)

The guided search model of Wolfe [71] is one of the most influential work for computational visual attention. This model was introduced to answer some of the drawbacks of FIT, such as the dichotomy between serial and parallel search. The main goal of the GST is to explain and to predict the results of visual search experiments [71]. The architecture of the guided search model is illustrated in Fig. 1.10.

This model has strong overlaps with FIT. However, FIT includes many feature maps for each color channel and for each orientation, whereas in GST there is only one map per feature type (color, orientation, etc) leading to fewer feature maps. Therefore, all color features are represented in a single map as well as all the orientations. These maps are later on fused together to create an *activation map* which is similar to the *master map* of locations in FIT. For the purpose of visual search however, the aim of parallel processing of basic features is to identify locations that are worthy of further attention.

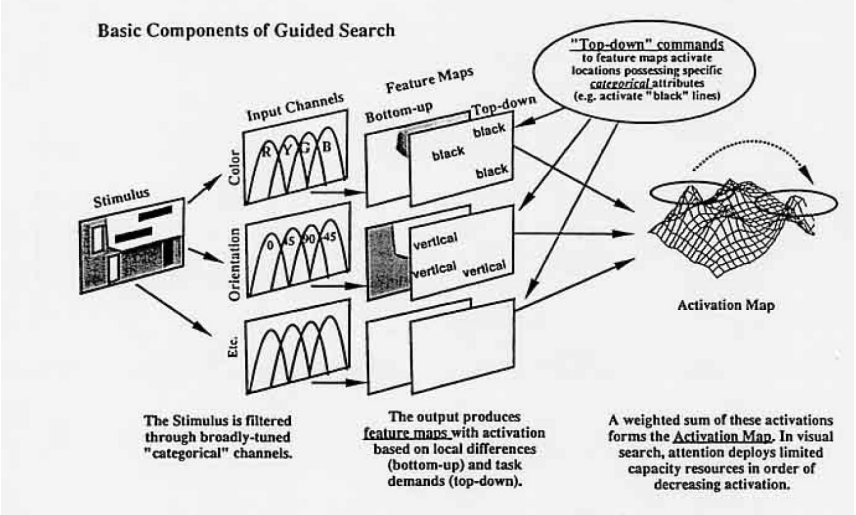


FIGURE 1.10 – Guided Search Model(Source : [71])

This is modelled as differential activation of locations in each feature map [71]. Guided search model is done in two approaches such as bottom-up activation and top-down activation.

Bottom-up activation is a measure of how unusual an item is in its present context, and the bottom-up activation for one location is based on the differences between itself and items at neighbouring loci in the relevant feature space. Thus, bottom-up activation guides towards distinctive items in the field. Top-down activation on the contrary is a user driven activation since it guides the attention to the desired item. For example, when searching for green crayons in a box of many different hues, a top-down request for *green* will activate locations that might contain the correct crayon. The target and distractors are examined to determine what categories are present in the target and decide whether they are unique or not. More weight is assigned to unique categories for each channel.

Finally, in the visual search, each feature module can be thought of as a pair of topographic maps with hills of higher activation marking locations. The attention is attracted to the hills which is shown in Fig. 1.10. The final activation map is seen by summing the activation from all the feature maps.

1.5.3/ ADDITIONAL THEORIES

Besides the Feature Integration Theory (FIT) and the Guided Search Theory (GIT), there are many variety of psychophysical theories on visual attention. Few of them are explained briefly in this section.

1.5.3.1/ ATTENTIONAL ENGAGEMENT THEORY (AET)

This theory deals with the the serial/parallel visual search dichotomy that is present in FIT. In [20], authors argue that the search is mainly parallel in groups of elements and that a continuation in search efficiency. Moreover, they argue that target-distractor and

distractor-distractor differences play a major role. That is, when the target is very dissimilar to the distractors, search is always fast (independently of the similarity or dissimilarity among distractors). When the target and distractors are similar, the heterogeneity in distractors become important for visual search. When the target is similar to the distractors and these distractors are heterogeneous, then grouping is difficult and a search close to a serial search is applied.

The attentional engagement theory is based on three components : (1) a perceptual input description such as color size which is accomplished in parallel ; (2) a selection performed by comparing the input descriptions with an internal template ; (3) a virtual short-term memory that save the selected information.

1.5.3.2/ Boolean Map Theory (BMT)

This new theory attempts to answer two questions [32] : what visual contents can an observer consciously access at one moment ? and how can an observer voluntarily select what to access ?

The Boolean Map Theory (BMT) presents a unified interpretation of a wide variety of visual attention phenomena which are treated in separate literatures. According to [32] BMT is briefly encapsulated into two propositions :

– 1. *Principle of access* : A boolean map divides the visual field into two discrete subsets : selected and not selected. This map can be associated with feature label per dimension. A feature label will provide a single overall feature description of the entire region encompassed by the map.
– 2. *Principle of Selection* : This takes place in one of two ways : (a) by selecting one feature value in one dimension or (b) combining the boolean map with the output of (a) through one of the two boolean operations (intersection and union).

The Boolean Map Theory is illustrated in Fig. 1.11.

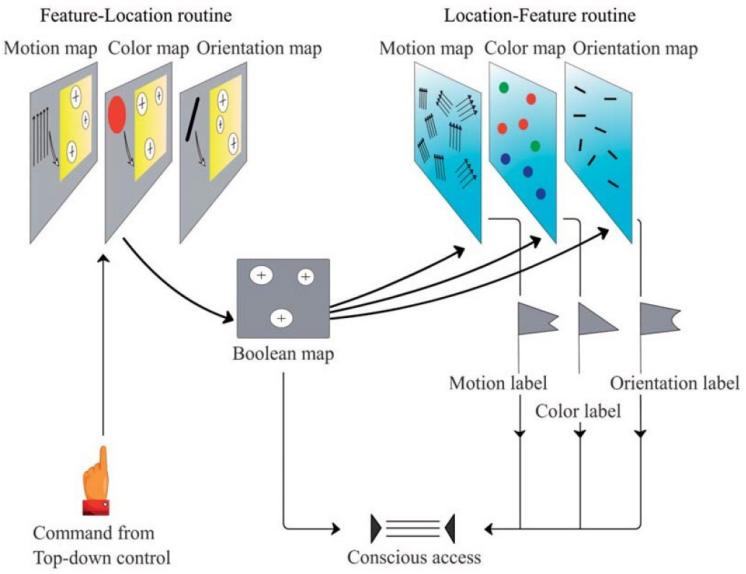


FIGURE 1.11 – A schematic overview of Boolean map Theory(Source : [32])

1.6/ THESIS MOTIVATION

Focusing on the relevant information is important for computer vision systems, especially in applications which require real-time computation such as in robotics. So, biologically motivated approaches that automatically find regions and objects of interest based on the concepts of the human visual system have proved to be effective solutions in different applications. The analysis of visual attention is considered as a very important component in the human vision system. In the visual field the region which drives human perceptual attention is called a *salient region*. The main emerging problem is how to design a model that put salient areas in conspicuous locations. Saliency detection has wide range of applications such as moving object detection, predicting human eye fixations, video summarization, video compression, action recognition and image/video quality assessment.

While saliency detection is a widely studied problem, most of the existing techniques are limited to the analysis of static images, and these approaches cannot be simply extended to the analysis of videos sequences. Dealing with videos is more a complicated issue as the perception of a video is different from that of static images due to its additional temporal information. A video contains a strong spatial-temporal correlation between the regions of consecutive frames. Furthermore, the motion of foreground objects dramatically changes the importance of the objects in a scene which leads to a different saliency map of the frame representing the scene. In addition, we know that natural scenes are composed of several dynamic entities such as moving trees, waves in water, fog, rain, snow and different illumination variations. All these characteristics make video processing for saliency evaluation a challenging task.

However, detecting salient regions and salient objects in complex dynamic scenes would be helpful in applications such as tracking, robotic navigation and localization and many more. In this thesis work, our aim is to develop a spatio-temporal saliency method that works well in the challenging conditions mentioned above.

1.7/ CONTRIBUTIONS AND THESIS STRUCTURE

The two contribution of the thesis is twofold. First, since the main approach for computing a spatio-temporal saliency map is to fuse a static saliency map with a dynamic saliency map, we analyse different possible fusion strategies to select the best ones. Secondly, we proposed a new method for spatio-temporal saliency detection based on combining color features with LBP (local binary patterns). LBP features are used here to represent dynamic textures in the videos. Experiments show the benefit of this approach on a large and diverse dataset.

The presented research work was published in two international conference paper. The conference paper have been in International Conference on Image processing (ICIP-2013) and also in International Conference on Pattern Recognition (ICPR-2014).

The remaining of the thesis is structured as follows :

– **Chapter 2 : State of the art**
  In this chapter we present a survey of existing spatio-temporal saliency detection methods found in recent literature. The main objective of this chapter is to study about the

similarities and differences among different methods and highlights their strengths and weaknesses. The chapter also discusses about some of the potential applications of the these models in computer vision.

- **Chapter 3 : Proposed spatio-temporal saliency models**
  This chapter describes the proposed spatio-temporal saliency detection method. This chapter starts with a description of a method that uses color features for static saliency detection and optical flow for dynamic saliency. We then improve the model computing motion information using local binary patterns. The chapter also reviews and compares different fusion approaches for spatio-temporal saliency detection.
- **Chapter 4 : Experiments and results**
  In this chapter we present extensive experimental results of the proposed method. First, we describe the datasets used for evaluations and the evaluation metrics adopted. Second, we presented the experimental results and discussion on the results. The proposed method is compared against other state-of-art techniques.
- **Chapter 5 : Conclusion and future work**
  This thesis is concluded by summarizing the main contributions of thesis and by suggesting some possible future research directions associated with this work.
- **Appendix**
  In appendix we provided the list of the publications related to this thesis and the some sample frames of the all the videos from the datasets used for evaluation.

2

State of the Art

This chapter gives an overview of existing research relevant to the contribution made in this thesis. The literature survey consists of two main parts. The first part presents a broad classification of spatio-temporal saliency detection techniques and their limitations. The second part presents an overview of applications of visual saliency in computer vision.

2.1/ Introduction

Over the recent past years the research on visual saliency has been tremendously increasing and many saliency models for different applications have been proposed. Most of the methods are limited to static images. A recent survey of existing static saliency detection methods can be found in [13, 14]. Since we focus on the analysis of video sequences in our work, we will mainly review in this chapter spatio-temporal saliency detection methods. We broadly classify these models into two categories : biologically inspired approaches and purely computational approaches.

2.2/ Biologically inspired models of visual attention

Biological approaches try to follow the principles of human visual system (HVS) in designing a saliency model. In this section we will introduce some of the most important biological inspired models for spatio-temporal saliency detection in videos.

2.2.1/ Koch and Ullman model of neural architecture

This was the first approach designed on the basis of biologically inspired principles of visual attention. This method, introduced by Koch and Ullman [39], was not implemented as an algorithm, but provided the algorithmic reasoning serving as a foundation for later implementations. The main idea of this approach is that several features are computed in parallel and their conspicuousness are collected in a saliency map. The Winner Take All (WTA) network finds the most salient region in this map, which is finally routed to a central representation where complex processing tasks take place. These complex tasks are restricted to the region of interest. The architecture of the model is shown in Fig. 2.1.

The model follows the principles of the feature integration theory (FIT) [68], see section 1.5.1. A number of elementary features such as color, orientation, direction of movement and disparity are represented in parallel in different topographical maps, which correspond to the early representation. A selective mapping from the early topographic representation into a more central non-topographic representation is done using certain rules. The central representation at any instant contains the properties of only a single location in the visual scene.

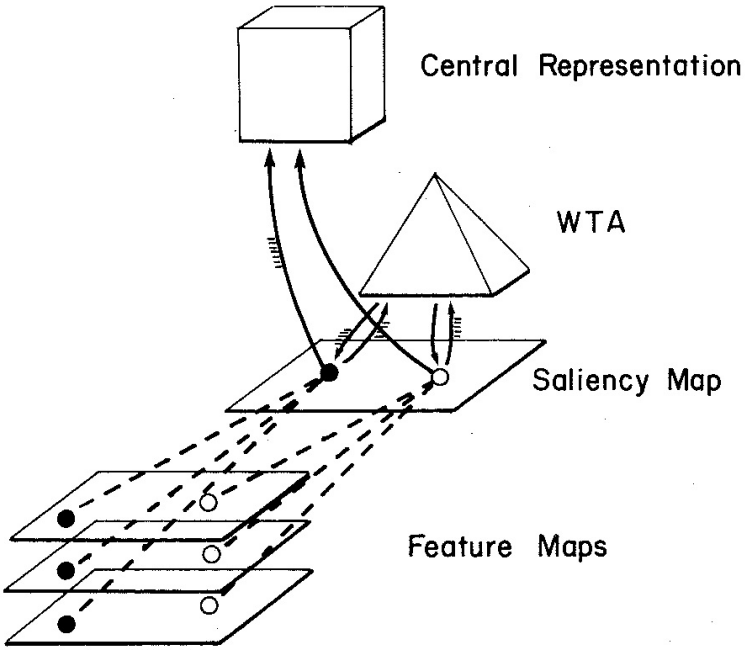


FIGURE 2.1 – Koch and Ullman architecture (Source : [39])

The main important contribution of Koch and Ullman's work is the WTA network. This network localizes the most active unit in the saliency map and shows how the selection of a maximum is implementable by neural networks. This approach is strongly biologically motivated and shows how such a mechanism might be realized in the human brain. After selecting the most salient region by the WTA, this region is routed into a central representation which at any instant contains only the properties of a single location in the visual scene. This routing is also known as selective routing model. However, a clear explanation of how the routing is performed in central representation is not mentioned by the authors. The mechanism for inhibiting the selected region causing an automatic shift towards the nest conspicuous location is known as inhibition of return (IOR). The proposed architecture is bottom-up. It is not discussed how top-down factors influence the selection of salient regions.

2.2.2/ ITTI'S MODEL

This model of visual attention is inspired by the behaviour and the neuronal architecture of the early primate visual system [44]. It is based on FIT [68] which explains the human visual search strategies. In this model, a visual input is first decomposed into a set of to-

pographic feature maps. Different spatial locations then compete for saliency within each map, so that only locations which locally stand out from their surround can persist. All these feature maps are processed purely in a bottom-up manner into a master "saliency map". In primates, such map is believed to be located in the posterior parietal cortex (PP). This model consequently represents a complete account of bottom-up saliency and doest not require any top-down guidance to shift attention. Moreover, it provides a massively parallel method for the the selection of a small number of interesting image locations.

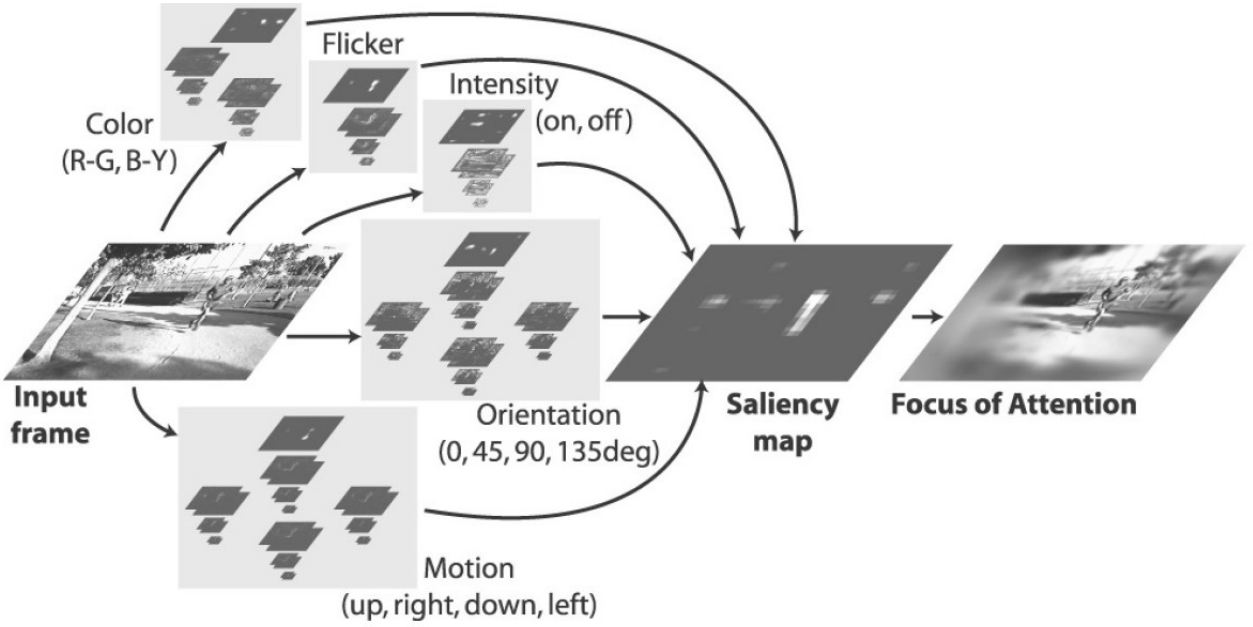


FIGURE 2.2 – Itti's model [35]

Fig. 2.2 shows the structure of the model. A static color image is given as input and is processed in nine spatial scales which are created using Gaussian pyramids. Each feature is computed by a set of linear "center-surround" operations. These center-surround operations are particularly well-suited to detect the locations which stand out from their surroundings which is a general computational principle in the retina, lateral geniculate nucleus and primary visual cortex. Center-surround is implemented in the model as the difference between fine and coarse scales.

This model is computed in three steps : extraction of early visual feature maps, computation of the saliency map, and application of WTA strategy to find the attended location. In the first step, four color channels are created : $R = r-(g+b)/2$ for Red, $G = g-(r+b)/2$ for Green, $B = b-(r+g)/2$ for Blue, and $Y = (r+g)/2-|r-g|/2-b$ for Yellow and an intensity image $I = (r+g+b)/3$ is obtained from the input color image. For each color channels and the intensity image, Gaussian pyramids $R(\sigma), G(\sigma), B(\sigma)$, $Y(\sigma)$ and $I(\sigma)$ are obtained using $\sigma \in [0..8]$. A center-surround difference between a center fine scale c and a surround coarser scale s yields the feature maps. The first set of feature maps is concerned with intensity contrast $I(c,s)$ computed as follows :

$$I(c,s) = |I(c) - I(s)| \tag{2.1}$$

A second set of maps is similarly constructed for the color channels. $RG(c, s)$ are created in the model to simultaneously account for Red/Green and Green/Red double apponency and $BY(c, s)$ for Blue/Yellow and Yellow/Blue double apponency :

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

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

where $\ominus$ indicates the scale difference between two maps.

Local orientation information is obtained from I using oriented Gabor pyramids $O(\sigma, \theta)$, where $\theta \in \{0°, 45°, 90°, 135°\}$. The orientation contrast $O(c, s, \theta)$ is computed for different orientations as :

$$O(c, s, \theta) = |O(c, \theta) \ominus O(s, \theta)|. \tag{2.4}$$

So, totally 42 feature maps are computed, 6 for intensity, 12 for color, and 24 for orientation. To extend this model to spatio-temporal saliency the authors add flicker and motion features. The extended model is described in [35].

The feature maps are combined into three "conspicuity" maps : $\bar{I}$ for intensity, $\bar{C}$ for color, and $\bar{O}$ for orientation. These three conspicuity maps are normalized and summed into the final saliency map as in Eq. 2.5, where $N(.)$ is normalization operator which highlights the conspicuous locations.

$$S = \frac{1}{3}(N(\bar{I}) + N(\bar{C}) + N(\bar{O})). \tag{2.5}$$

2.2.3/ MARAT'S MODEL

This spatio-temporal saliency model was proposed by Marat *et al.* [51] to predict eye movement during video free viewing. This model is inspired by the biology of the first steps of HVS, from the retina cells to the complex cells of the primary visual cortex. The visual signal first enters the retina where its is decomposed into two outputs signals. These two outputs signals are then split into elementary feature maps by cortical-like filters. These feature maps are used to form two saliency maps which are then fused into a master saliency map. Fig. 2.3 shows a description of the model.

The retina model has two outputs which are formed by different ganglion cells : a *parvocellular output* which provides the detailed information which can be simulated by extracting the high spatial frequencies of an image, and a *magnocellular output* which responds rapidly and provides global information that can be simulated by using lower spatial frequencies.

The parvocellular output enters to primary visual cortex which is modelled as cortical-like filters and the visual information is decomposed into different spatial frequencies using Gabor filters. These filters are a good compromise of resolution between the frequency and spatial domains. The number of orientations and frequencies are respectively fixed at $N_\theta = 6$ and $N_f = 4$, for a total of 24 filters. The output of each filter corresponds to an intermediate map which is the equivalent of some of the elementary features maps in FIT. All these intermediate maps are then normalized and added together to obtain a static saliency map M_s for each frame.

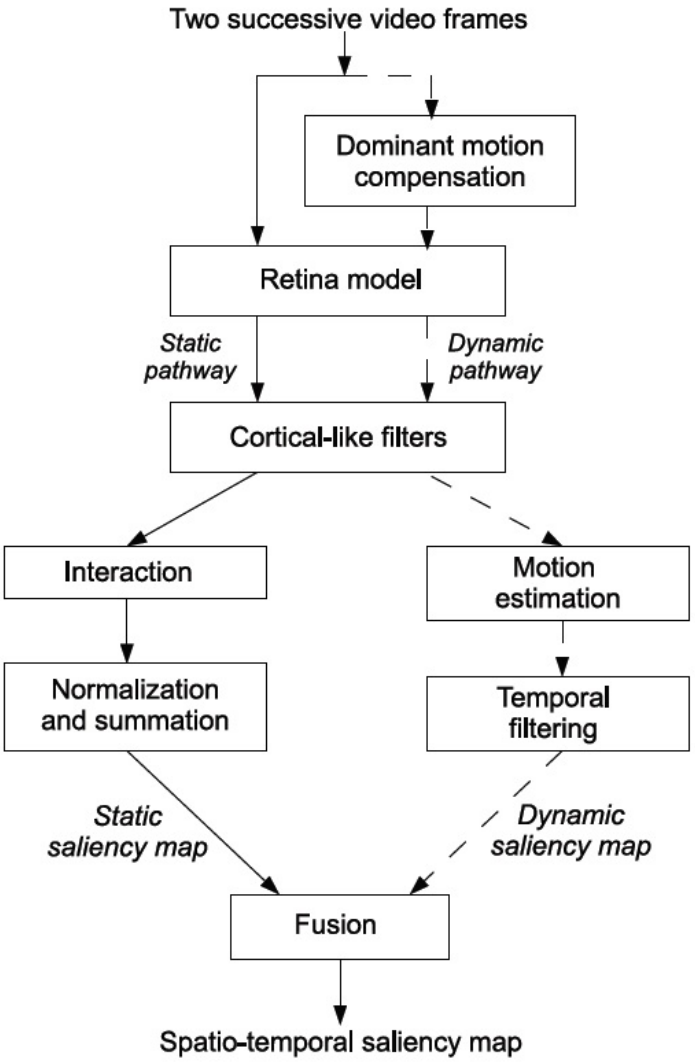


FIGURE 2.3 – Marats's model [51]

In [51], authors assumed that visual attention is attracted by motion contrast and they defined the motion of regions against the background. The first step is thus the compensation of the background motion between two successive frames which is done by carrying out a robust multi-resolution estimation of an affine parametric motion model. Optical flow is used to estimate the motion of every pixel, and a temporal median filter is applied to remove noise and finally gives a dynamic saliency map M_d for each frame.

The final spatio-temporal saliency map is obtained by the fusion of the static saliency map M_s and the dynamic saliency map M_d, as given by the following equation :

$$M = \alpha M_S + \beta M_d + \gamma M_s \times M_d, \tag{2.6}$$

where $\alpha = max(M_s)$, $\beta = skewness(M_d)$ and $\gamma = \alpha\beta$.

2.2.4/ LE MEUR'S MODEL

In [42], the authors present a coherent computational approach to the modelling of bottom-up visual attention. This model is mainly based on the current understanding of the HVS behaviour and uses as main features contrast sensitivity functions, perceptual decomposition, visual masking and center-surround interactions. The model was later improved to include two chromatic saliency maps which enhance its capacity to detect visually important locations and a temporal saliency is additionally computed to get a spatio-temporal saliency maps [41]. The improved model is illustrated in the Fig. 2.4.

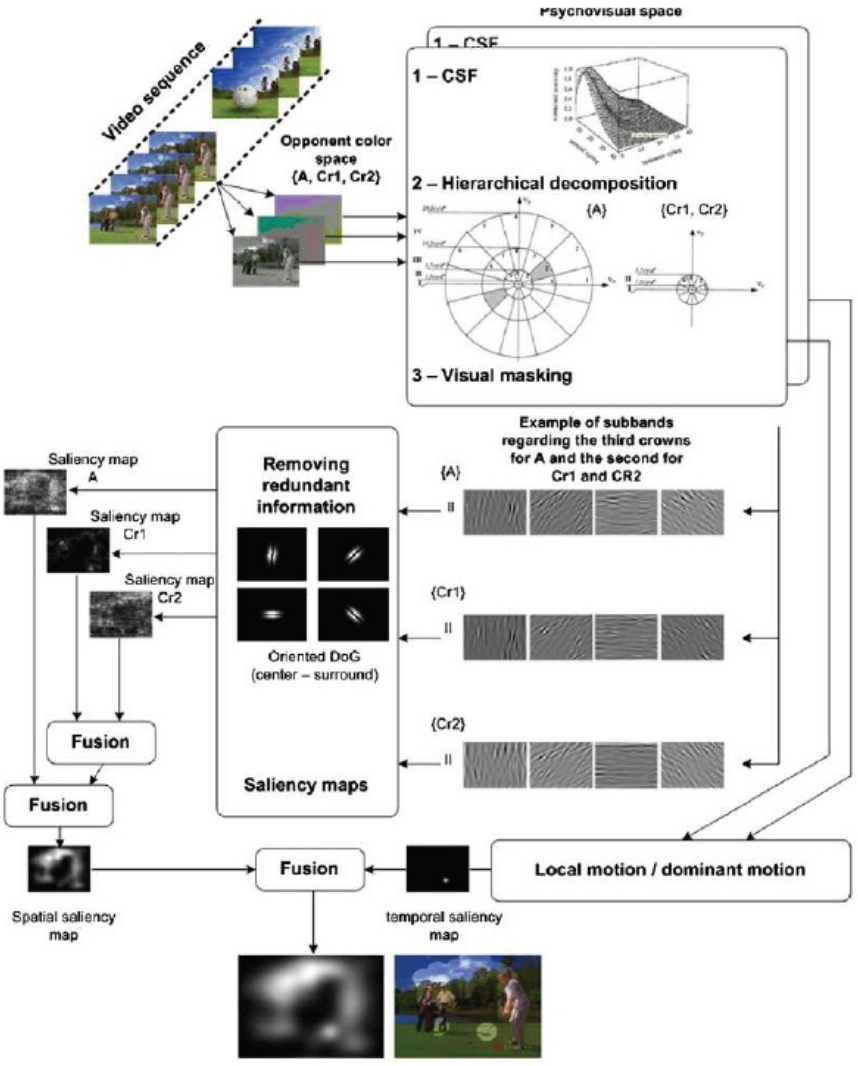


FIGURE 2.4 – Le Meur model [42]

To compute static saliency, the data is expressed in the same unit and three contrast sensitivity functions (CSF) are used. The CSF shows how sensitivity varies with spatial frequency and orientation. This modulation is called visual masking. A hierarchical decomposition is then conducted splitting the 2D spatial frequency domain both in spatial radial frequency and in orientation. This decomposition is applied to each of the three perceptual components. The visual system cannot process all visual information at once,

two kinds of mechanisms are required to cope with this biological limitation. The first one selects small part of the visual field on which a close inspection is performed and the second one is more passive than the first one and its role is to suppress the redundancy of the visual information. A center-surround filter is used to suppress the irrelevant data. After applying the center-surround filters, three saliency maps are derived for the three color channels and they are fused to get static saliency map.

To compute temporal saliency map the author assumed that motion contrast is one of the most visual attractors. The main aim of the temporal saliency map computation rests on the relative motion occurring in the retina. The difference between the local and the dominant motion gives relative motion. The local motion $\vec{V}_{local}$ at each point is computed by block matching in different resolutions. The local motion does not necessarily reflect the motion contrast, for example when the camera is fixed. The dominant motion is represented by a 2D parametric model, $\vec{V}_{\theta}(s)$; where θ is the parameter vector and $\vec{V}_{\theta}(s)$ is the motion vector at pixel s. Finally, the relative motion representing the motion contrast is given by :

$$\vec{V}_{relative}(s) = \vec{V}_{\theta}(s) - \vec{V}_{local}(s). \tag{2.7}$$

The combination of different saliency maps into a unique saliency map known as spatio-temporal saliency map is the final step. This is achieved by :

$$S(s) = F(S_1, S_2)(s) = \text{IntraMap}(s) + \text{InterMap}(s), \tag{2.8}$$

where

$$\text{IntraMap}(s) = \frac{S_1(s)}{\text{NearestMax}_1} + \frac{S_2(s)}{\text{NearestMax}_2}, \tag{2.9}$$

$$\text{IntraMap}(s) = \frac{S_1(s)}{\text{NearestMax}_1} * \frac{S_2(s)}{\text{NearestMax}_2}, \tag{2.10}$$

with S_1 and S_2 the normalized spatial and temporal maps and NearestMax_1 , NearestMax_2 the value of the nearest local maximum regarding the current position for the component S_1 and S_2 respectively.

2.3/ Computational models of visual attention

In this section we will introduce some of the most important computational models of spatio-temporal saliency which have close impact on our work. We start by introducing the Graph Based Visual Saliency (GBVS) model [30], then the method based on Quaternion Fourier Transform (PQFT) [28], next we describe the SUN model of saliency [74] followed by phase discrepancy [77], self-resemblance [64], spatio-temporal saliency using dynamic textures [48], information-theoretic model [16], incremental coding length based saliency [31], and a saliency model proposed by Mancas [50].

2.3.1/ Graph Based Visual Saliency (GBVS) model

This is a bottom-up saliency model proposed by the authors in [30] and it is based on graph computations known as Graph Based Visual Saliency (GBVS). The method mainly

consists of three stages : the first one is the extraction of feature vectors at locations over the image plane ; the second stage forms an activation map on the feature channels, and the final stage normalizes them in order to highlight the conspicuity and admits the combination with other maps.

In this method, feature maps are extracted at multiple spatial scales using a scale-space pyramid. Three feature maps are derived from intensity, color and orientation similarly to [44]. For a given feature map M the goal is to compute an activation map A at every location (i, j) of the image. To compute the activation map, a fully connected graph G_A over all grid locations of each feature map is built and weights between the nodes are assigned proportional to the similarity of feature values and spatial distance. The dissimilarity between two positions (i, j) and (p, q) in the feature map, with respect to feature values $M(i, j)$ and $M(p, q)$, is computed as follows :

$$d((i, j)||(p, q)) = log\frac{M(i, j)}{M(p, q)} \tag{2.11}$$

The directed edge from node (i, j) to node (p, q) is then assigned a weight proportionally to their dissimilarity and their distance on lattice M given by :

$$w((i, j)||(p, q)) = d((i, j)||(p, q)) * F(i - p, j - q), \tag{2.12}$$

where $F(a, b) = exp(-\frac{a^2+b^2}{2\sigma^2})$ and σ is a free parameter.

The resulting graphs G_A are treated as Markov chains by normalizing the weights of the outbound edges of each node to 1 and by defining an equivalence relation between nodes and states, as well as between edge weights and transition probabilities. The stationary distributions of the graphs are adopted as activation and saliency maps. In the equilibrium distributions, the nodes which are highly dissimilar to their surrounding nodes are assigned large values. In the final stage, the activation maps are normalized to emphasize conspicuous locations, and are combined to form a single overall saliency map.

2.3.2/ Phase spectrum of Quaternion Fourier Transform (PQFT) based saliency

In [28], the authors present a quaternion representation of an image which is composed of intensity, color and motion features. This method follows the principle of Phase spectrum of Fourier Transform (PFT) to compute the saliency map. For a given image $I(x, y)$ the steps are summarized as follows :

$$f(x, y) = F(I(x, y)), \tag{2.13}$$

$$p(x, y) = P(f(x, y)), \tag{2.14}$$

$$sM(x, y) = g(x, y) * ||F^{-1}[e^{i.p(x,y)}]||^2, \tag{2.15}$$

where F and F^{-1} denote the Fourier Transform and Inverse Fourier Transform, respectively. $P(f)$ represents the phase spectrum of the image and $g(x, y)$ is a 2D gaussian filter.

The authors claim that the phase spectrum of an image represents the local image information and the location with less periodicity or less homogeneity in the image shows a "pop-out" effect in the reconstruction of the image's phase spectrum. This motived the

author to propose a model of visual saliency based on PFT which was later extended to spatio-temporal saliency map using quaternion representation of an image and quaternion Fourier Transform by taking color and motion information into consideration.

For each frame at time t, two color channels are computed taking into account color opponency as follows :

$$RG(t) = R(t) - G(t); \quad BY(t) = B(t) - Y(t)$$

The intensity channel and motion channel are given by :

$$I(t) = \frac{r(t) + g(t) + b(t)}{3}; \quad M(t) = |I(t) - I(t - \tau)|$$

The four feature channels are then combined in form of a quaternion image $q(t)$ as :

$$q(t) = M(t) + RG(t)\mu_1 + BY(t)\mu_2 + I(t)\mu_3, \tag{2.16}$$

where $\mu_i, i = 1, 2, 3$ satisfies $\mu_i^2 = -1, \mu_1 \perp \mu_2, \mu_2 \perp \mu_3, \mu_1 \perp \mu_3, \mu_3 = \mu_1\mu_2$.

In Eq. 2.16 , the quaternion representation of an image is calculated and each pixel is denoted as $q(n, m, t)$, where (n, m) is the location of the pixel and t is the time point. The QFT representation can then be computed as follows :

$$Q[u, v] = F_1[u, v] + F_2[u, v]\mu_2, \tag{2.17}$$

where

$$F_i[u, v] = \frac{1}{\sqrt{MN}} \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} e^{-\mu_1 2\pi((mv/M)+(nu/N))} f_i(n, m), \tag{2.18}$$

$$f_i(n, m) = \frac{1}{\sqrt{MN}} \sum_{v=0}^{M-1} \sum_{u=0}^{N-1} e^{\mu_1 2\pi((mv/M)+(nu/N))} F_i[u, v], \tag{2.19}$$

with $f_i(n, m)$ the inverse Fourier transform and (n, m), (u, v) the locations of each pixel in spatial and frequency domain, respectively, and N and M the images's height and width.

The frequency domain representation $Q(t)$ of $q(t)$ is given in polar form as follows.

$$Q(t) = \|Q(t)\|e^{\mu\phi(t)}, \tag{2.20}$$

where $\phi(t)$ is the phase spectrum of $Q(t)$ and μ is the unit pure quaternion. Finally the spatio-temporal saliency sM is computed by

$$S(t) = g * \|q'(t)\|^2, \tag{2.21}$$

where g is a Gaussian filter and $q'(t)$ is the reconstruction of $Q(t)$

The spatio-temporal saliency model computed by PQFT considers features such as motion, color, intensity, and orientations. These features are represented in a quaternion image which allows parallel processing, making the method fast.

2.3.3/ SALIENCY USING NATURAL STATISTICS FOR DYNAMIC SCENES

The main goal of this model is to detect potentially important targets and allocate computational resources to them for further processing. To achieve such a goal, the authors in [75] suggest that the pre-attentive process must estimate the probability of a target given the visual features at every location in the visual field. This methods is based on a Bayesian framework from which bottom-up saliency emerges naturally, using image statistics derived from a large collection of natural images.

The probability of a target at a point z in the visual field is given by :

$$S_z = p(C = 1|F = f_z, L = I_z), \tag{2.22}$$

where C is a binary variable that is 1 when the target is present at the current point and 0 otherwise. $L = I_z$ represents the location of the point z and $F = f_z$ is the visual feature at this point.

The features used are the intensity I, Red-Green (RG) and Blue-Yellow (BY) channels which are calculated as follows :

$$I = r + g + b \tag{2.23}$$

$$RG = r - g \tag{2.24}$$

$$BY = b - \frac{r+g}{2} - \frac{min(r,g)}{2}, \tag{2.25}$$

where r, g and b are the Red, Green and Blue channels of the image.

These features are then processed by a bank of spatio-temporal separable linear filters designed to be both efficient and in line with the HVS. Difference of Gaussians (DoG) filters are used as spatial filters.

2.3.4/ PHASE DISCREPANCY MODEL

The phase discrepancy model was developed to detect moving objects against dynamic background [77]. This algorithm is based on the fact that the displacement of the foreground and the background can be represented by the phase change of the Fourier spectra, and the motion of background objects can be extracted by phase discrepancy in an efficient way.

To detect salient moving objects from the dynamic background, the method follows the idea of predictive coding. First, the next frame is predicted only considering background movements. Let the function $f(\mathbf{x}, t)$ denotes the observation at time t, where $\mathbf{x} = [x_1, x_2]^T$ is the 2-dimensional vector of the spatial location. Given the background B_t, the appearance of the background B_{t+1} in the next frame is predicted based on the intensity constancy assumption :

$$f(\mathbf{x}, t) = f(\mathbf{x} + \mathbf{v}, t + 1), \tag{2.26}$$

where $\mathbf{x} \in B_t \cap B_{t+1}$ and $\mathbf{v}$ is the motion vector. The saliency of the moving object is computed by taking the prediction error as a likelihood function of the appearance of moving objects at a certain location. In other words the reconstruction error map $s(\mathbf{x}, t)$ can be considered as a saliency map which is given in Eq. 2.27 :

$$s(\mathbf{x}, t) = [f(\mathbf{x} + \mathbf{v}, t + 1) - f(\mathbf{x}, t)]^2. \tag{2.27}$$

To generate the saliency map, displacement vectors are computed in the Fourier domain. The method was evaluated on a dataset created by the authors with manually generated groundtruth for moving objects.

2.3.5/ SELF-RESEMBLANCE MODEL

A novel unified framework for both static and temporal saliency detection is proposed by Seo and Milanfar [64]. This method is based on a bottom-up framework and uses local regression kernels as local features which differs from conventional filter responses. Local regression kernels capture the underlying local structure of the image very well even in the presence of significant distortions. In [64], authors use a non parametric kernel density estimation for such features, which results in a saliency map constructed from a local self-resemblance measure indicating likelihood of saliency.

The saliency at a pixel is defined as a posterior probability distribution $S_i = Pr(y_i = 1|\mathbf{F})$, where the feature matrix F, contains features not only from the center, but also from the surrounding region., and y_i is a binary random variable that indicates whether a pixel is salient or not ; $y_i = 1$ if the pixel x_i is salient and $y_i = 0$ otherwise. Assuming that every pixel has the same prior probability of being salient, and that the probability of the features is uniform over the features space, the saliency measure reduces to the conditional probability $p(\mathbf{F}|y_i = 1)$ using Bayes' theorem. This conditional probability is estimated based on non-parametric kernel density estimation.

Finally the saliency value at the pixel i is given by :

$$S_i = \frac{1}{\sum_{j=1}^{N} exp(\frac{-1+\rho(\mathbf{F}_i,\mathbf{F}_j)}{\sigma^2}}, \tag{2.28}$$

where $\rho(\mathbf{F}_i, \mathbf{F}_j)$ is the "Matrix Cosine Similarity (MCS)" between the center feature matrix $\mathbf{F}_i$, and a neighboring feature matrix $\mathbf{F}_j$.

A brief description of the method is illustrated in Fig. 2.5. For more details, please refer to [64].

2.3.6/ SPATIO-TEMPORAL SALIENCY USING DYNAMIC TEXTURE

In this method [48], the authors proposed a spatio-temporal saliency detection algorithm based on center-surround mechanism which is inspired by biological mechanism of motion-based perceptual grouping. This method is built to address the challenges in complex natural scenes which contain dynamic entities such as swaying trees, waves, snow, rain and smoke filled environments. The saliency of a location is the power of a predefined set of features to discriminate between the visual stimuli in a center and surround window. The features used in this method are spatio-temporal video patches and are modelled as dynamic textures.

The saliency detection problem is treated as a two classes classification problem between stimuli of interest (salient) and background. The saliency of a location is then given by the ability of the visual features extracted from that location in discriminating the two classes. The dynamic textures (DT) model of [19] is used due to its ability to account for variability, while jointly modelling the spatio-temporal characteristics of visual stimulus. The dynamic

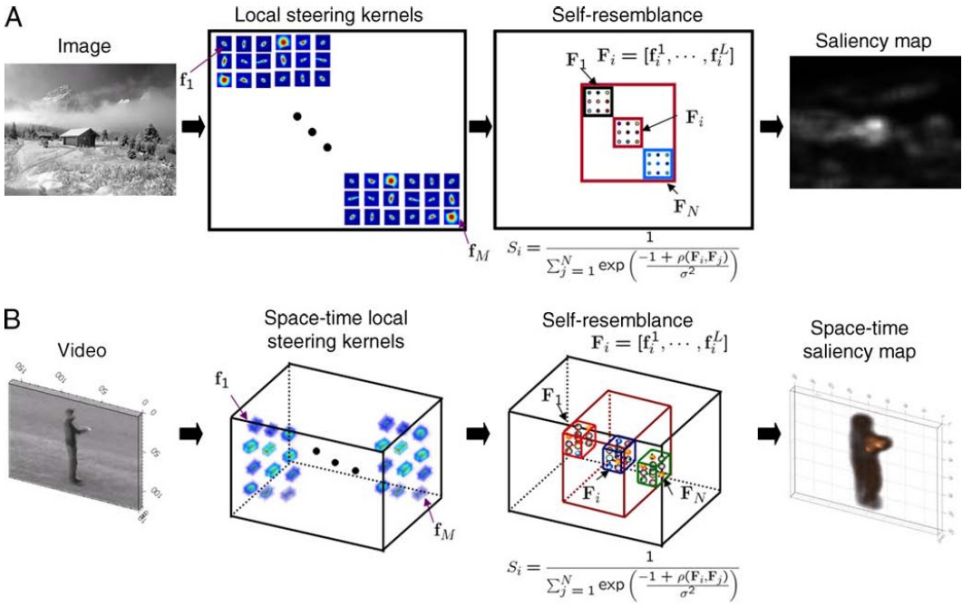


FIGURE 2.5 – Self-resemblance model (source : [64])

texture is an autoregressive model which represents the appearance of a stimulus $\mathbf{y}_t \in \Re^m$, where the pixels of the two-dimensional visual stimulus at time t are represented as a column vector of length m, as a linear function of a hidden state process $\mathbf{x}_t \in \Re^n (n << m)$ subject to Gaussian observation noise. The state and appearance processes form a linear dynamical system :

$$\mathbf{x}_t = \mathbf{A}\mathbf{x}_{t-1} + \mathbf{v}_t, \tag{2.29}$$

$$\mathbf{y}_t = \mathbf{C}\mathbf{x}_t + \mathbf{w}_t, \tag{2.30}$$

where $\mathbf{A} \in \Re^{m \times n}$ is the state transition matrix, $\mathbf{C} \in \Re^{m \times n}$ the observation matrix, and $\mathbf{v}_t$ and $\mathbf{w}_t$ are Gaussian state and observation noise processes, respectively. The model is parametrized by $\Theta = (\mathbf{A}, \mathbf{C}, \mathbf{Q}, \mathbf{R}, \mu_1, \mathbf{S}_1)$, with $\mathbf{Q}$ and $\mathbf{R}$ the covariances of the state and observation noise respectively, and μ_1 and $\mathbf{S}_1$ defining the initial state.

To compute the saliency $S(l)$ of a location l, a collection of spatio-temporal patches are extracted from a centre window (centred at l) and a surround window. The dynamic texture (DT) parameters are learned for τ frames from center, surround, and total windows, to obtain the densities $p_{\mathbf{y}|C(l)}(\mathbf{y}(\tau)|1)$,$p_{\mathbf{y}|C(l)}(\mathbf{y}(\tau)|0)$, and $p_Y(\mathbf{y}(\tau))$, respectively. $S(l)$ is finally computed with Eq. 2.31

$$S(l) = \sum_{c=0}^{1} p_{C(l)}(c) KL(p_{Y|C(l)}(\mathbf{y}|c) \| p_{\mathbf{y}}(\mathbf{y})) \tag{2.31}$$

where $KL(p \parallel q)$ is the Kullback-Leibler (KL) divergence between the probability distributions.

This model was then evaluated on the dynamic scenes dataset [9]. This dataset contains different challenges like moving water, camera motion, peds, smoke and illumination changes, snow etc. The spatio-temporal saliency maps are then computed for each frame of the video sequences and these saliency maps are compares with the ground truth of

manually segmented salient objects using ROC analysis. This model was used for background subtraction application purpose.

2.3.7/ Information-theoretic model

The authors in [16] proposed a new visual saliency model for detecting moving objects in a video sequence using spatio-temporal volumes. In this method the spatial and temporal saliency are calculated separately and they are fused with a dynamic fusion method. Fig. 2.6 shows the block diagram of the method.

The spatio-temporal saliency measure is based on information theory, self-information in this context, and is called information saliency map (ISM). The ISM is a two dimensional array where each entry reflects the saliency of the corresponding pixel in the video frame. To compute the ISM for the frame I_t, a three dimensional spatio-temporal (ST) volume V is constructed from the current frame I_t and the previous $(N-1)$ frames, i.e. $I_1, I_2, ..., I_{N-1}$. The ST volume is then divided into smaller ST sub-volumes of size $M \times M \times N$. For each sub-volume a spatial vector set $X = \{x_0, x_1, ..., x_{N-1}\}$ is constructed for the patch (x_0) in frame I_t and its $N-1$ spatial neighbourhoods. The temporal saliency of a patch $I_T(r,s)$ is computed as :

$$\begin{aligned} I_T(x_0) &= I_T(r,s) && (2.32)\\ &= -log_2[P(I_T(r,s)|V(r,s)] && (2.33)\\ &= -log_2(P(x_0|X), && (2.34) \end{aligned}$$

where $V(r,s)$ is the sub-volume constructed at (r,s) and x_0 is the vector form of the patch $I_t(r,s)$.

To compute the conditional probability $P(x_0|X)$ in the above equation from the image data, a non-parametric approach using kernel density estimation (KDE) is employed. The spatial saliency is computed in a similar way and the final ST saliency is obtained by fusion of the spatial and temporal saliency maps as follows :

$$I(x_0) = -log_2(x_0|X \cup X^{'}), \quad (2.35)$$

where X and X' are respectively the temporal and spatial neighborhoods.

2.3.8/ Icremental Coding Length (ICL) saliency model

The authors in [31] proposed a dynamic saliency visual attention model based on the rarity of features. They introduced the Incremental Coding Length (ICL) to measure the perspective entropy gain of each feature using sparse coding techniques to represent features. More specifically, image patches are represented as a linear combination of sparse coding basis functions which are referred as features. The set of basis functions are learned so that they yield a sparse representation of natural image patches.

Then, authors proposed to use ICL as a computational principle based on features. This principle aims to optimize the immediate energy distribution in the system in order to achieve an energy economic representation of its environment. The ICL is defined by the feature activity ratio distribution. The distribution can be defined over space as well as over time.

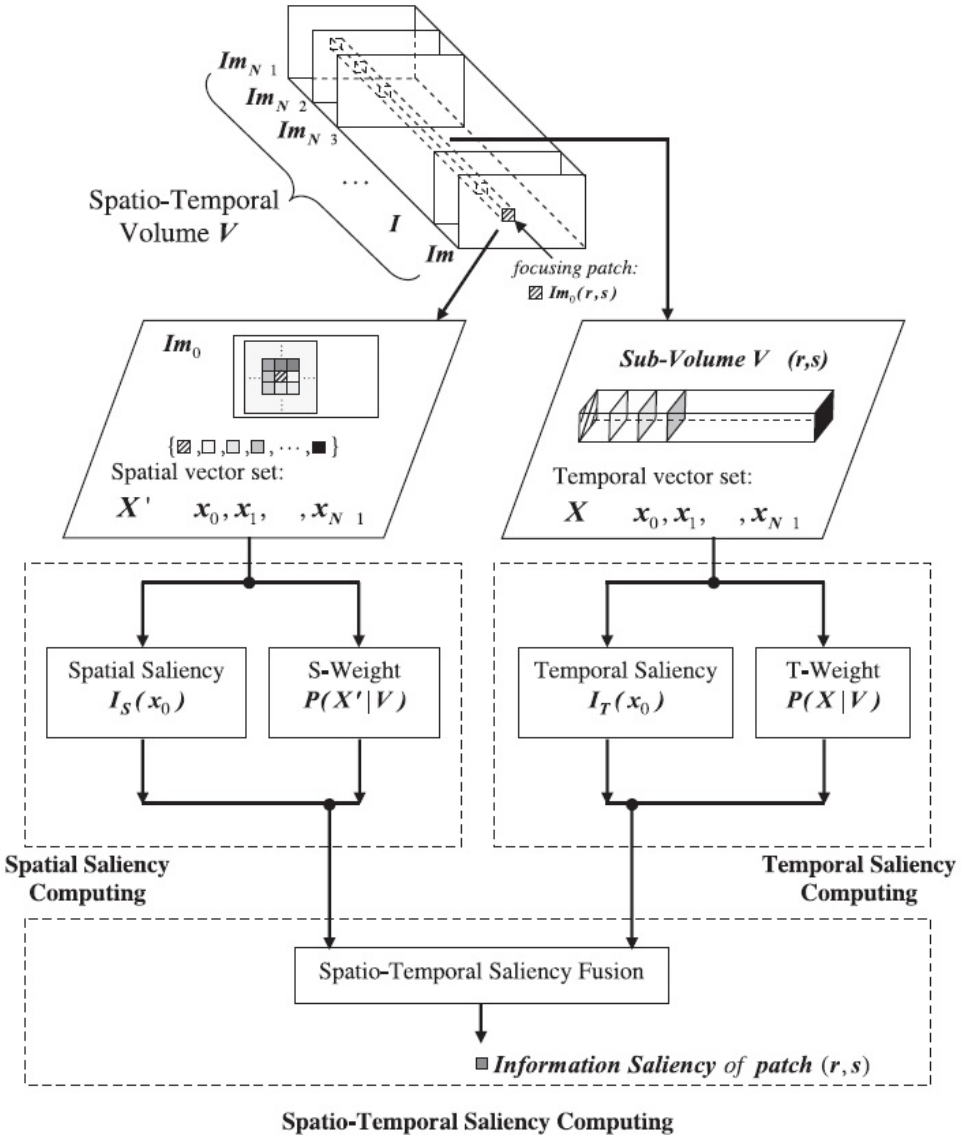


FIGURE 2.6 – Information theoretic's model (source : [16])

An image $\mathbf{X}$ is divided into patches as $\mathbf{X} = [\mathbf{x}^1, \mathbf{x}^2, ...\mathbf{x}^k, ..]$, where $\mathbf{x}^k$ is a vectorized image patch. The image is then coded using the learnt dictionary to get a set of features $\mathbf{W} = [\mathbf{w}^1, \mathbf{w}^2, ...\mathbf{w}^i, ..]$. The activity ratio p_i for i^{th} feature is therefore defined as :

$$p_i = \frac{\sum_k |\mathbf{w}_i \mathbf{x}^k|}{\sum_i \sum_k |\mathbf{w}_i \mathbf{x}^k|}. \tag{2.36}$$

AUthors finally define the ICL of each feature based on the entropy $H(\mathbf{p})$ of the probability function of feature activities $\mathbf{p} = [p_1, p_2, ...]$:

$$ICL(p_i) = \frac{\partial H(\mathbf{p})}{\partial p_i} = -H(\mathbf{p}) - p_i - \log p_i - p_i \log p_i. \tag{2.37}$$

Finally for the given image $X = [\mathbf{x}^1, \mathbf{x}^2, ...\mathbf{x}^n]$ the saliency map $M = [\mathbf{m}_1, \mathbf{m}_2, ..\mathbf{m}_n]$ is defined

by :

$$\mathbf{m}_k = \sum_{i \in S} d_i \mathbf{W}_i \mathbf{x}^k, \tag{2.38}$$

where d_i is the amount of energy received by each feature, $\mathbf{W}$ is the learned feature vector from sparse coding, and S is the learned salient feature set.

2.3.9/ Mancas' saliency model

Mancas *et al.* proposed a bottom-up saliency method based on global rarity quantificationin [50]. The model is an extension of the work in [49], and it is based on a multi-scale approach using features extracted from optical flow. The computation is done in three steps : 1) Motion features extraction 2) Spatio-temporal filtering of the features 3) From feature detection to feature selection.

In the first step, motion vectors are extracted from the video frames, each frame being divided into square cells. The motion vectors contain the information of speed and motion direction. The features are discretized into 4 directions (North, South, West and East) and five speeds (very slow, slow, mean , fast and very fast).

In the second step, a spatio-temporal low-pass filter is applied to each of the discretized feature channels (4 directions and 5 speeds). The space and time dimensions are separated. The frames are first spatially low-pass filtered. Then a weighted sum is made on the time dimension to provide lower weight to the older frames.

In the final step, the resulting filtered images from the 9 feature channels (4 directions and 5 speeds) are separated into 5 bins each and the self-information $I(b_i)$ of the pixels for a given bin b_i is computed as described in equation. 2.39 :

$$I(b_i) = -\log\left(\frac{H(b_i)}{\text{Card}(B)}\right), \tag{2.39}$$

where $H(b_i)$ is the value of the histogram H for the bin b_i and Card(B) the cardinality of the frame (size of the frame in pixels). After computing the saliency map for each feature channel, a maximum operator is applied to put together the 4 directions into a single saliency map and the 5 speeds into a second saliency map. These two saliency maps gives indication about the rarity of the statistics of a given video volume at two different scales for a given feature.

2.4/ Applications of visual saliency

Visual attention has recently became an active area of research in the fields of computer vision and robotics. It provides an intuitive method to determine the most interesting region of an image in a bio- inspired way, and has appeared as a promising approach to improve computational vision systems. Visual saliency has been used in different applications such as segmentation, image re-targeting, video quality assessment and tracking. In this section we discuss several application scenarios in the field of computer vision and robotics.

FIGURE 2.7 – Example of image segmentation using visual saliency. From left to right : original image, obtained saliency map, mean-shift segmentation, and segmentation based on saliency values (last two columns). Source [3].

2.4.1/ SALIENT OBJECT SEGMENTATION

Visual saliency has been used widely for object segmentation. Segmentation is the problem of grouping parts of an image together according to some measure of similarity. It is a well studied problem that has two major issues : setting the starting points for segmentation (seeds) and choosing the similarity criterion to segment regions. Achanta *et al.* [5] proposed a method for salient object detection in image that outputs full resolution saliency maps in which well defined boundaries are preserved by retaining substantially more frequency content from the original image. This full resolution saliency maps is then used for segmenting the objects in the image. The segmentation of objects is done by using adaptive thresholding [4]. In this technique the authors oversegmente the input image using K-means clustering and retain only those segments whose average saliency is greater than a constant threshold. This threshold is set at 10% of the maximum normalized saliency value. The binary maps representing the salient object are thus obtained by assigning ones to pixels of chosen segments and zeros to the rest of the pixels. The segmentation procedure using saliency detection is illustrated in Figure 2.7. In [17], a simple and efficient algorithm which yields full resolution saliency maps is proposed. The algorithm is based on regional contrast without any prior knowledge and the saliency maps are used for segmentation. Several other saliency object detection methods has been developed for segmentation of objects in images and videos [7, 56, 62].

2.4.2/ IMAGE RE-TARGETING

Image re-targeting is another important application of saliency detection in recent years. The diversity of today's display device sizes and aspect ratios demands smarter ways of re-targeting. Content aware image re-targeting methods aim to arbitrarily change image aspect ratio while preserving visually prominent features. Adaptive warping [70] and seam carving [10] are methods that accentuate visually important content with minimal loss of original content. The existing re-targeting methods mostly rely on grayscale intensity gradient map which shows higher energy only at edges of objects. In [6] the authors developed a computationally efficient, noise robust re-targeting scheme based on seam carving by using saliency maps which assigns higher importance to visually prominent whole regions and not just the edges. An example of image re-targeting in shown in Figure 2.8. The concept of image re-targeting can also be applied to videos due to development and

FIGURE 2.8 – Image re-targeting using salient object detection methods. From left to right : Saliency map, detected image regions to remove, re-targeted image, and original image. Source [6].

the flexibility of watching video on portable devices. This evoke growing interest in automatic video adaption that seeks to change the resolution from larger to smaller video sources while faithfully conveying the original information. However, video re-targeting is a hard problem, since it is a very subjective task to map human cognition into the automated process. In [45] the author proposed a saliency based video re-targeting method. In this method, authors incorporate the features from phase spectrum of quaternion Fourier Transform (PQFT) saliency in spatial domain and global motion residual based on matched features points by the Kanade-Lucas-Tomasi (KLT) tracker in temporal domain.

2.4.3/ OBJECT TRACKING

Tracking of objects in dynamic environments is important in applications such as video surveillance and robotics. Huge research has been done on tracking problem and produced a diverse set of approaches and rich collection of tracking algorithms [73]. Appearance based methods are most popular among these. These methods rely uniquely on models of object appearance and do not take background into account. In the biological world, object tracking is tightly related to attentional tasks such as the guidance of eye movements. That is salient locations become the focus of attention (FoA). Therefore, FoA can be used for tracking. In [67], Toyama *et al.* proposed an incremental FoA procedure to combine multiple trackers which leads to increase robustness. In [65], authors proposed an object tracking method with particle filter using visual saliency. They show that using visual saliency improves the robustness of object tracking, in particular it helps overcome the challenges such as illumination variations, occlusions and similarity between the target object and the background. Authors in [47] introduced a saliency based discriminant tracking method using dynamic textures and showed very good results for some complex scenes.

2.4.4/ VIDEO COMPRESSION

When a human observer looks at natural images or video clips only a small region around the center of his eye fixation is captured at high resolution and the rest of the region looks

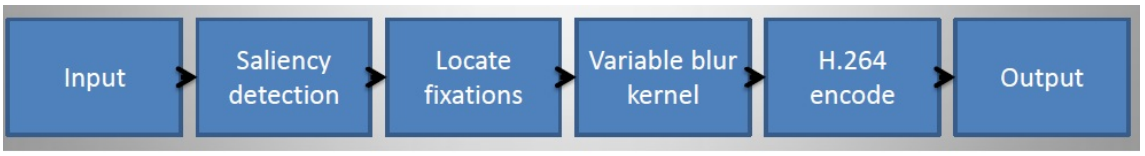


FIGURE 2.9 – Video compression procedure using salieny detection method

blurred due to the non uniform distribution of photoreceptor [28]. So it is necessary to consider this fact when compressing images or videos, to get uniform quality. Saliency based methods have been introduced in the past years. In [34], an attention based video compression method is introduced. In this method regions of interest are selected based on a non-linear integration of low-level visual cues. A dynamic foveation filter then blurs every frame, increasingly with distance from salient locations and finally the resulting sequence is compressed using H.264. The video compression procedure using saliency detection is illustrated in Figure 2.9.

2.4.5/ ROBOTIC NAVIGATION AND LOCALIZATION

Another application scenario of an attention system in robotics is the detection of landmarks for localization and navigation, especially in outdoor environments and open areas. The ability to get to specific locations in an environment in a timely manner is fundamental in creating a fully autonomous mobile robot system. To this end, a robot not only needs to be able to move in its environment, but it also have to identify its location. The existing approaches try to solve this problem using proximity sensors such as Laser Range Finder (LRF) and Global Positioning System (GPS). But these methods may fail in complex environments, where no GPS signals is received for example. To overcome these problems saliency detection can facilitate the search of landmarks during operation by selecting interesting regions. For example, in [15] a vision based navigation and localization system using biologically inspired scene understanding models is developed. The authors utilized salient features to localize the robot accurately and for the navigation purpose the salient regions are used to perform visual feedback control to direct its heading and go to a user-provided goal location.

In [25], authors present a simultaneous localization and mapping (SLAM) system which is based on a biologically motivated feature selection strategy. A visual attention system detects salient features which are highly discriminative, ideal candidates for visual landmarks which are easy to re-detect. Features are tracked over several frames to determine stable landmarks and to estimate their 3D position in the environment.

2.5/ DISCUSSION ON STATE-OF-THE-ART METHODS

In this chapter we reviewed several state-of-the-art methods for spatio-temporal saliency detection in dynamic scenes and some of the potential application of saliency detection. We have classified the state-of-the-art methods into two categories : biologically inspired and pure computational methods. These methods have been developed by authors

for different applications such as object detection, background subtraction, image/video compression and robotic navigation.

We observe that these methods suffer several limitations and are not able to satisfy all the challenges of natural video sequences. Let us note that natural video scenes have many dynamic entities such as moving water, snow, smoke etc. Along with these dynamic entities, if the video has additional camera motion it becomes very complex. The perception of videos is different than for static images due to its additional temporal information. So, dealing with temporal information is important in spatio-temporal saliency. All the reviewed methods use features such as color, intensity, orientation and motion, flicker informations for temporal saliency. A spatio-temporal saliency is then obtained by fusing spatial and temporal saliency maps. In [28] the authors used frame difference as a motion feature, but using this frame difference the motion feature is not estimated properly. This type of computation of temporal saliency maps may not succeed in complex videos. In [51] [41] the motion feature is computed using optical flow. Though the motion features computed using optical flow may solve some problems, it fails when the scenes have complex background. In [16] the authors compute the temporal saliency using temporal patches. This model can detect the temporal saliency in static video scenes. If the video has camera motion this method completely fails. In [64] the temporal information is computed by 3D-Local Steering Kernels (3D-LSK). This method some how solves the problem of detecting the spatio-temporal saliency in complex environments. Due to its invariant properties it can also detect the saliency in videos with distortions. In [48] the authors modelled the video patches using dynamic textures whereas center-surround mechanism is applied on the video patches. This method is developed to deal with the complex dynamic scenes so they dealt the temporal information in a proper manner. After analysing these methods we observe that most of the methods fail to address the complex dynamic scenes. This is because due to lack of proper temporal saliency modelling. None of these methods haven't deal with temporal information properly. Moreover, most of these methods are restricted to detect salient regions. Only [48] [16] methods detect salient objects. It is also very difficult to access which method is good in predicting salient regions because, the state of art methods use different evaluation criteria and different dataset to evaluate the performance of their models. A fair comparison should be done in order to compare any methods with common evaluation metrics. In order to address these issues, we propose a new model which is explained in chapter 3.

3

PROPOSED SPATIO-TEMPORAL SALIENCY MODELS

Detecting spatio-temporal salient objects and regions is a challenging task in complex dynamic scenes as explained in chapter 2. In this chapter, we describe our proposed spatio-temporal saliency models to deal with the challenges of dynamic scenes such as dynamic background. We will first present a method based on optical flow estimation, and then a method based on local binary patterns.

3.1/ INTRODUCTION

In chapter 2, we reviewed several state-of-the-art methods for spatio-temporal saliency detection in dynamic scenes. As we have seen, most of these methods are based on the fusion of a static map which computes spatial saliency, and a dynamic map which computes temporal saliency. We follow a similar approach and propose two spatio-temporal saliency models, one based on optical flow and another dynamic textures computed with local binary patterns(LBP) respectively.

We will first describe the model based on optical flow, then describe the model based on LBP. We also compare different fusion techniques for obtaining the final spatio-temporal saliency map.

3.2/ SPATIO-TEMPORAL SALIENCY DETECTION BASED ON OPTICAL FLOW

In this section we describe our first proposed spatio-temporal saliency model based on optical flow. In this model we compute the static saliency map using color features and the dynamic saliency map using an optical-flow motion estimation method. These two maps are finally fused to get the spatio-temporal (ST) saliency map. Each of the three steps is explained in detail in the following subsections whereas a block diagram of the proposed method is shown in Figure. 3.1.

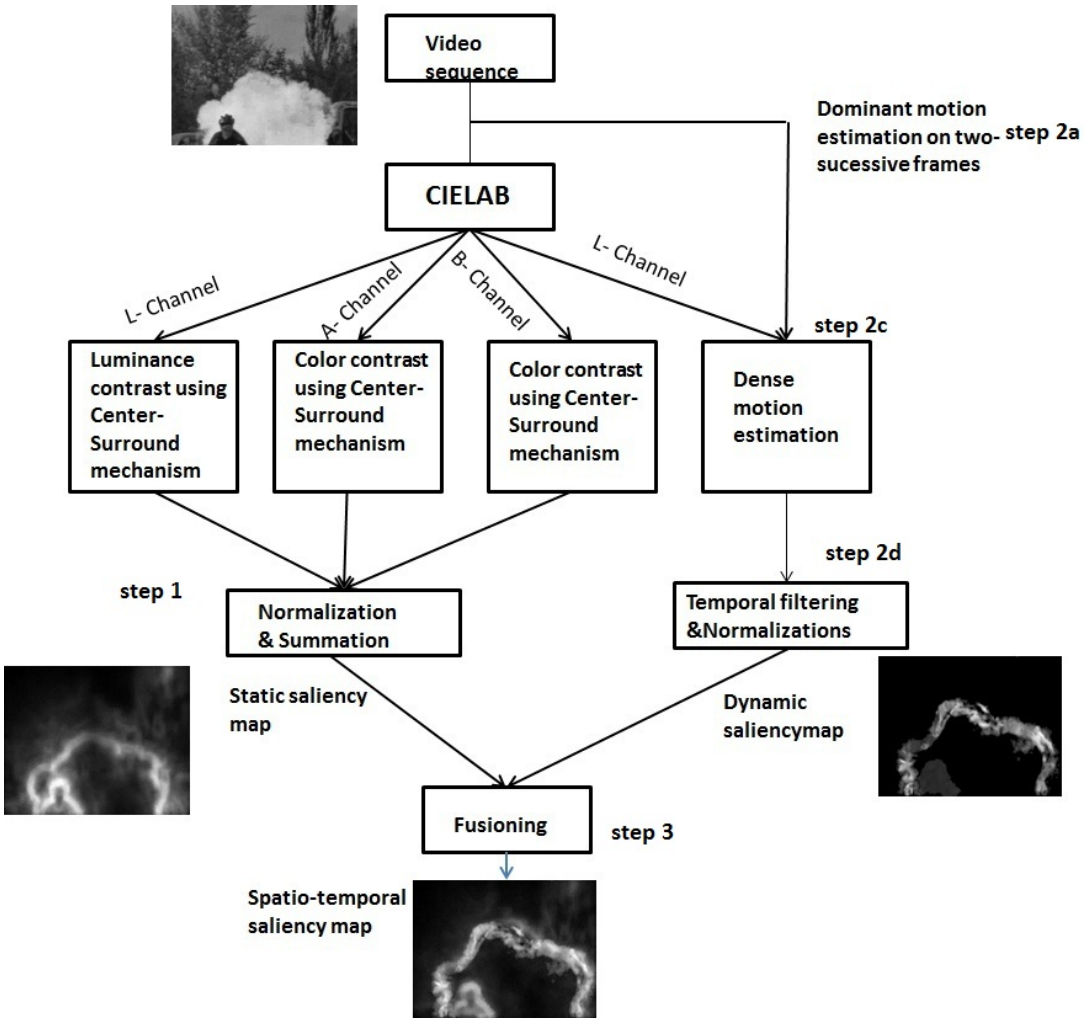


FIGURE 3.1 – Spatio-Temporal saliency using optical flow.

3.2.1/ STATIC SALIENCY MAP COMPUTATION

There are several methods proposed in the literature for computing a static saliency map in an image. In our work, we used a saliency detection method based on context information [27]. Our choice is motivated by the fact that this method proves to be the best in a recent evaluation of saliency detection methods [14]. This context saliency model is based on four principles of human visual attention that are :

– 1a. Low level considerations such as contrast and color.
– 1b. Global considerations, which suppress frequently-occurring features.
– 1.c Visual organization rules.
– 1.d High-level factors.

According to the principle (1a), areas with distinctive colors should be assigned higher saliency value. The principle (1b) helps to suppress frequently occurring features and according to principle (1c), the salient pixels are grouped together.

The saliency is computed in three steps. In the first step, local and global single scale saliency is computed for each pixel i in an image. The pixel i is salient if its appearance

is unique following principle (1a). So, a pixel i is considered salient if the appearance of the patch p_i centered at pixel i is distinctive with respect to all other image patches. The dissimilarity measure between the patches p_i and p_j is defined by :

$$d(p_i, p_j) = \frac{d_{color}(p_i, p_j)}{1 + c.d_{position}(p_i, p_j)}, \tag{3.1}$$

where d_{color} represents the Euclidean distance between the vectorized patches p_i and p_j of sizes 7×7 in CIElab color space which are normalized to the range $[0, 1]$, and $d_{position}$ is the Euclidean distance between the position of patches p_i and p_j. c is a constant scalar value set to $c = 3$ in our experiments (changing the value of c does not significantly affect the final result).

To evaluate a patch's uniqueness, there is no need to incorporate its dissimilarity to all the image patches. So for every patch p_i, we search for the K most similar patches q_k, $k = 1, \ldots, K$, in the image. The pixel i is considered salient when its dissimilarity $d(p_i, q_k)$ is high $\forall k \in [1, K]$. The single-scale saliency value of a pixel i at scale r is defined by Eq. 3.2 :

$$S_i^r = 1 - e^{\frac{1}{K}\sum_{k=1}^{K} d(p_i^r, q_k^r)}, \tag{3.2}$$

where p_i^r is the patch centered at pixel i at scale r.

In the second stage, multi-scale saliency is computed by considering different scales of the processed image. These multiple scales are utilized by representing each pixel i by the set of multi-scale image patches centered at it. The pixel i is considered as salient if it is consistently different from other pixels in multiple scales. So, to get a global saliency of the pixel i , it needs to consider multi-scale K most similar patches which are different from it. So for the patch p_i at scale r, we consider as candidate neighbours all the patches in the image whose scales are $R_q = \{r, \frac{1}{2}r, \frac{1}{4}r\}$. From all these patches, the K most similar patches according to Eq.3.1. So, Eq.3.2 cab be rewritten as :

$$S_i^r = [1 - exp\{-\frac{1}{K}\sum_{k=1}^{K} d(p_i^r, q_k^{r^k})\}]. \tag{3.3}$$

where $r_k \in R_q$. The multi-scale saliency is defined for each scales as in Eq. 3.3, and the saliency of the pixel i is taken as the average of its saliency at different scales :

$$\bar{S}_i = \frac{1}{M}\sum_{r\in R} S_i^r, \tag{3.4}$$

where the saliency maps S_i^r at each scale are normalized to the range $[0, 1]$, and M is the number of scales.

The final step includes the immediate context of the salient object. The immediate context suggests that areas that are close to the foci of attention should be explored significantly more than far-away regions. The visual context is simulated by extracting the most attended localized areas at each scale produced by Eq. 3.2. A pixel i is considered as a focus of attention at scale r which is normalized to the range $[0, 1]$, if the dissimilarity measure of Eq : 3.2 exceeds a given threshold $(S_i^r > 0.8)$. Then, each pixel outside the attended

areas is weighted according to its Euclidean distance to the closest attended pixel using the following equation :

$$\hat{S}_i = \frac{1}{M} \sum S_i^r (1 - d_{foci}^r(i)), \tag{3.5}$$

where M is the total number of scales and $d_{foci}(i)$ is the Euclidean positional distance between pixel i and the closest focus of attention pixel at scale r, normalized to the range $[0, 1]$.

In our experiments, we compute the saliency for all image pixels and consider patches of size 7×7 with 50% overlap for the nearest neighbours search. The scale used are $100\%, 80\%, 50\%, 30\%$ of the original image scale. Some examples of obtained static saliency maps are shown in Fig. 3.2.

3.2.2/ Dynamic saliency map using optical flow

In this model of spatio-temporal saliency, we compute the dynamic saliency map based on motion features. As we know, human beings are able to see stable and moving components in a moving scene effortlessly. We assume that visual attention is attracted by motion contrast and we define it as the motion of some regions of the scene against the background.

The computation of the dynamic saliency map is performed in few steps. Firstly, two successive video frames are processed to estimate the dominant motion, i.e. the background motion due to the camera using an affine parametric model. The affine parameters are then used to compensate for the background camera motion. Then, a local dense motion is estimated on the compensated frames. Finally, a median filter is applied in temporal direction on the estimated motion vectors to remove temporal noise.

3.2.2.1/ Step 2a : Dominant motion compensation

Before computing the dynamic saliency map, it is necessary to compensate for the motion due to the camera. This helps to estimate the relative motion of objects in the scene with respect to the background. This background motion is computed using a 2D parametric affine motion estimation algorithm [46], which provides the dominant motion between two successive frames using a robust multi-resolution estimation.

The parametric model chosen here is an affine one with 6 parameters as follows

$$\begin{cases} v_x = a_1 + a_2.x + a_3.y \\ v_y = a_4 + a_5.x + a_6.y \end{cases},$$

where $[a_1, \ldots, a_6]^T$ is the vector of parameters and v_x and v_y are the vectorial components of the dominant motion computed at position (x, y). The estimation of these parameters is based on the well known *brightness constancy assumption* [46]. Let $I(x, y, t)$ be the image brightness at a point (x, y) at the time t, then the brightness constancy assumption is given by :

$$I(x, y, t) = I(x + u\delta t, y + v\delta t, t + \delta t), \tag{3.6}$$

where (u, v) is the horizontal and vertical image velocity at a point and δt is the time step.

The direct way to use the brightness constancy equation is to formulate the data conservation error measure using the sum-of squared difference (SSD) measure computed over a region W of the image. The error associated with a given displacement $[u, v]^T$ is formulated as :

$$E_D(u, v) = \sum_{(x,y)\in W} [I(x, y, t) - I(x + u\delta t, y + v\delta t, t + \delta t)]^2. \tag{3.7}$$

The equation 3.7 can be rewritten by taking the Taylor series approximation of $I(x+u\delta t, y+v\delta t, t+\delta)$ which is given by :

$$I(x + u\delta t, y + v\delta t, t + \delta) = I(x, y, t) + \frac{\partial I}{\partial x}\delta u + \frac{\partial I}{\partial x}\delta v + \frac{\partial I}{\partial x}\delta t. \tag{3.8}$$

By substituting the equation. 3.8 in the equation. 3.7, we get :

$$E_D = \sum_{(x,y)\in W} (I_x(x, y, t)u + I_y(x, y, t)v + I_t(x, y, t))^2, \tag{3.9}$$

where I_x, I_y and I_t indicate partial derivatives of the brightness function I with respect to x, y and t respectively. This last equation can be written as :

$$E_D(\mathbf{u}) = \sum_W ((\nabla I)^T \mathbf{u} + I_t)^2, \tag{3.10}$$

where $\nabla I = [I_x, I_y]^T$ denotes the local brightness gradient vector, and $\mathbf{u} = [u, v]^T$ denotes the flow vector. The flow field is modeled as a parametric function of the image coordinates $\mathbf{u}(x, y; \mathbf{a})$, and the parameters $\mathbf{a}$ are obtained by minimizing the sum of the errors over a neighborhood W :

$$E_D(\mathbf{u}) = \sum_W ((\nabla I)^T \mathbf{u}(\mathbf{a}) + I_t)^2. \tag{3.11}$$

3.2.2.2/ Step 2b : Robust estimation of parameters

In order to make the estimation of the affine parameters less sensitive to noise, a robust estimator is used. In particular, authors in [46] use an M-estimator. M-estimators are based on the generalization of the maximum likelihood estimators (MLE) which minimize the following objective function :

$$\min_\theta \sum_{i=1}^{n} \rho(x_i; \theta), \tag{3.12}$$

where ρ is some function.

The above problem can be solved using the likelihood equations :

$$\min_\theta \sum_{i=1}^{n} \psi(x_i; \theta) = 0, \tag{3.13}$$

with $\psi(x_i; \theta) = \frac{\partial \rho(x_i;\theta)}{\partial \theta}$.

The choice of the function ρ defines the robust M-estimator. If ρ is the squared errors function as in the previous section, then the error $\rho(x;\theta)$ increases at an accelerating rate. It is therefore desirable to use error functions such that the error increase becomes constant when a threshold is reached. This behaviour reduces the effect of outliers on the estimation. For example, the least squared estimator corresponds to $\rho(x) = x^2$, and the L_1 estimator to $\rho(x) = |x|$.

In [46], the Geman-McClure error function is used. This error function is defined by :

$$\rho(x,\sigma) = \frac{x^2}{x^2+\sigma}, \tag{3.14}$$

where σ is a control parameter that can be seen as an outlier threshold. The influence function for this error function is given by $\psi(x,\sigma) = \frac{2x\sigma}{(\sigma+x^2)^2}$.

Using the Geman-McClure error function, the regression problem is reformulated as :

$$E_D(\mathbf{a}) = \sum_W \rho((\nabla I)^T \mathbf{u}(\mathbf{a}) + I_t, \sigma), \tag{3.15}$$

and solved iteratively using the following update equation :

$$a_i^{(n+1)} = a_i^{(n)} - \omega \frac{1}{T_{a_i}} \frac{\partial E_D}{\partial a_i} \ \forall i. \tag{3.16}$$

The algorithm starts by constructing a Gaussian pyramid of the images. At the coarse level all the affine parameters are initialized to zero. The translational terms a_1 and a_4 are solved first by performing n iterations of the update equation. In our experiments, we use $\omega = 1.995$ and fix the maximum number of iterations to 30. The change in the affine parameters is then computed using the iterative update scheme. The process is repeated until the full resolution is reached. The obtained parameters are then used to warp the images using bilinear interpolation.

An example of dominant motion compensation is shown in Fig. 3.3 where the camera is moving from left to right. A four level pyramid was used to estimate the dominant motion parameters. With the estimated parameters the first frame is back-warped using bilinear interpolation. The relative motion between the two frames can then be computed using the warped frames.

3.2.2.3/ Step 2c : Local motion estimation using dense optical flow

After the dominant motion compensation between two successive frames, we estimate the local motion on each frame using a dense optical flow based method developed by [21]. This method uses polynomial expansion to approximate the neighbours of a pixel. Quadratic polynomials are used to express the local signal model as :

$$f_1(\mathbf{x}) = \mathbf{x}^T \mathbf{A}_1 \mathbf{x} + \mathbf{b}_1^T \mathbf{x} + c_1, \tag{3.17}$$

where $\mathbf{A}$ is a symmetric matrix, $\mathbf{b}$ a vector and c a scalar. These coefficients are estimated using a weighted least square fitted to the signals values in the neighborhood.

Let assume that the signal undergoes a global displacement $\mathbf{d}$, then we can construct a new signal f_2 as follows :

$$f_2(\mathbf{x}) = \mathbf{x}^T\mathbf{A}_1\mathbf{x} + (\mathbf{b}_1 - 2\mathbf{A}_1\mathbf{d})^T\mathbf{x} + \mathbf{d}^T\mathbf{A}_1\mathbf{d} - \mathbf{b}_1^T\mathbf{d} + c_1 = \mathbf{x}_T\mathbf{A}_2\mathbf{x} + \mathbf{b}_2^T\mathbf{x} + c_2. \tag{3.18}$$

Equating the coefficients in the above two quadratic polynomials, we get :

$$\mathbf{A}_2 = \mathbf{A}_1 \tag{3.19}$$

$$\mathbf{b}_2 = \mathbf{b}_1 - 2\mathbf{A}_1\mathbf{d} \tag{3.20}$$

$$c_2 = \mathbf{d}^T\mathbf{A1d} - \mathbf{b}_1^T\mathbf{d} + c_1 \tag{3.21}$$

The displacement $\mathbf{d}$ is finally obtained by :

$$\mathbf{d} = \frac{1}{2}\mathbf{A}_1^{-1}(\mathbf{b}_2 - \mathbf{b}_1) \tag{3.22}$$

To make to computation of the displacement vector in Eq. 3.22 less sensitive to noise, we make the assumption that the displacement field varies slowly, so that we can integrate information over a neighborhood of each pixel. More formally, we minimize the following function :

$$\sum_{\Delta\mathbf{x}\in I} w(\Delta\mathbf{x})|\mathbf{A}(\mathbf{x}+\Delta\mathbf{x})\mathbf{d}(\mathbf{x}) - \Delta\mathbf{b}(\mathbf{x}+\Delta\mathbf{x})|^2, \tag{3.23}$$

where $w(\Delta\mathbf{x})$ is a weight function for the points in the neighborhood.

The minimum is obtained for

$$\mathbf{d}(\mathbf{x}) = (\sum w(\mathbf{x})\mathbf{A}^T\mathbf{A})^{-1}\sum w(\mathbf{x})\mathbf{A}^T\Delta\mathbf{b}). \tag{3.24}$$

To improve the robustness, the displacement is parameterized according to some motion model. For an affine motion model, we have [21] :

$$d_x(x,y) = a_1 + a_2x + a_3y + a_7x^2 + a_8xy, \tag{3.25}$$

$$d_y(x,y) = a_4 + a_5x + a_6y + a_7xy + a_8y^2, \tag{3.26}$$

which can be rewritten as

$$\mathbf{d} = \mathbf{Sp} \tag{3.27}$$

with

$$\mathbf{S} = \begin{bmatrix} 1 & x & y & 0 & 0 & 0 & x^2 & xy \\ 0 & 0 & 0 & 1 & x & y & xy & y^2 \end{bmatrix}, \tag{3.28}$$

$$\mathbf{p} = [a_1\ a_2\ a_3\ a_4\ a_5\ a_6\ a_7\ a_8]^T. \tag{3.29}$$

Inserting into Eq. 3.23 to obtain a weighted least square problem.

$$\sum_i w_i||\mathbf{A}_i\mathbf{S}_i\mathbf{p} - \Delta\mathbf{b}_i||^2, \tag{3.30}$$

which we can solved to get the parameters $\mathbf{p}$.

The displacement estimation can be done from two different approaches, iterative displacement estimation and multiscale displacement estimation. In the iterative approach the same polynomial expansion coefficients are used in all iterations and they need only to be computed once. The displacement obtained at one iteration is used as prior for the next one. The second approach uses a multiscale representation. It starts at a coarse scale to get a rough but reasonable displacement estimate and propagates this through finer scales to obtain increasingly more accurate estimates. The dense optical flow estimation is illustrated in the Fig. 3.4.

3.2.2.4/ STEP 2D : TEMPORAL FILTERING AND NORMALIZATION

We apply a temporal median filter to remove noise. The idea is that if a pixel has motion in one frame but not in the previous ones, then it is most probably noise resulting from motion estimation. In our work, we apply the temporal median filter to the magnitude of the motion vectors estimated in five successive frames (the current one and the previous four frames).

After temporal filtering, the last step to find the most salient regions is the normalization step. This normalization step is similar to the method proposed by [44], and salient regions are found based on their dissimilarity from neighbours. The steps for finding the salient regions from the median filtered image are as follows :

1. Normalize the temporal median filtered image M in the range $[01]$,
2. Find the maximum, $\max(M)$, and average, $\bar{M}$, values of the normalized maps,
3. Multiply each normalized map by $(\max(M) - \bar{M})^2$,
4. Remove from the map values those are less than 20% of the maximum value.

An example of filtered and normalized dynamic saliency map is shown in Fig. 3.5.

3.2.3/ LIMITATIONS OF OPTICAL FLOW

In the chapter 1 we have studied the main challenges to address in complex scenes whereas in chapter 2 we reviewed the state-of-the-art methods and their limitations. Based on these studies, firstly we computed the static saliency maps using color features and dynamic saliency map using optical flow. From the experimental results, we have observed that dynamic saliency maps using optical flow methods are relevant in scenes which have good motion contrast. In Figure. 3.6 we can observe that these scenes have good motion contrast. We can observe that in peds and boats sequences in Figure. 3.6. In these sequences the pixels has uniform motion due to this the dynamic saliency are estimated properly. Whereas in the Figure. 3.7 we observe the failures of optical flow in complex scenes. In the bottle sequence the motion of water is not uniform, due to this the dynamic saliency map is noisy and it is very difficult to localize the salient object in this sequence. Likewise the skiing and chopper sequences are noisy because of snow and fog. Due to this, each pixel in the sequence has non-uniform motion which makes the computation of saliency map very difficult using optical flow. So, to overcome these limitations we propose a new saliency model based LBP-TOP in section. 3.3

3.3/ Spatio-temporal saliency model using local binary patterns

In the previous section, (Section 3.2), we used optical fow for dynamic saliency map computation. As discussed in Section 3.2.3, the optical flow based method properly detects salient objects in dynamic scenes with high motion contrast. It fails when the scene has more complex background such as sea waves, snow, fog, or moving tree, which is frequent in real world scenarios.

The detection of salient objects in dynamic scenes with complex background is important for a wide range of applications like object detection and segmentation, target tracking and action recognition. So, to overcome the problems of optical flow and to deal with complex dynamic background scenes, we propose a new method for spatio-temporal saliency detection based on dynamic textures to model the varying appearance of dynamic scenes with time. Dynamic textures or temporal textures are textures that show stationary properties in time. In step 3 of the proposed approach Here, dynamic textures are modelled by local binary patterns computed in three orthogonal planes (LBP-TOP) to combine motion and appearance simultaneously. This dynamic texture representation using LBP-TOP has been shown to perform well in dynamic texture recognition [76]. The spatio-temporal saliency that we propose is computed by using a center surround contrast measure applied to each of three planes separately. An overview of the approach is shown in Fig. 3.8, and the detailed explanations are given in the following subsections.

3.3.1/ Basic local binary patterns texture descriptor

Local binary operator, introduced by Ojala *et al.* [54] is a powerful texture descriptor. It is based on the assumption that the texture locally has two complementary aspects, a pattern (local spatial pattern) and a strength (local grayscale contrast). The original version of the local binary pattern (LBP) operator works on 3×3 pixels blocks and labels the image pixels by thresholding them with the center value and considering the result as a binary number. The resulting histogram of these $2^8 = 256$ different labels is then used as a texture descriptor. The basic LBP is illustrated in Figure. 3.9.

Consider a grayscale image $I(x, y)$ and let g_c denote the graylevel of an arbitrary pixel (x, y), i.e $g_c = I(x, y)$. Let g_p denote the graylevel value of a sampling point in an evenly space circular neighborhood of P sampling points and radius R around point (x, y) :

$$g_p = I(x_p, y_p), p = 0, \ldots, P - 1;$$

and

$$x_p = x + R\cos(2\pi p/P),$$

$$y_p = y - R\sin(2\pi p/P).$$

The local texture of the image $I(x, y)$ is characterized by the joint distribution of graylevel values :

$$T = t(g_c, g_0 - g_c, g_1 - g_c, ...g_{P-1} - g_c). \tag{3.31}$$

The generic LBP operator is derived from this joint distribution by summing the thresholded differences weighted by powers of two :

$$LBP(x_c, y_c) = \sum_{p=1}^{P} s\left(g_p - g_c\right) 2^p, \qquad s(x) = \begin{cases} 1 & \text{if } x \geq 0 \\ 0 & \text{otherwise} \end{cases}. \tag{3.32}$$

Different circular sampling schemes are possible as shown in Fig. 3.10.

3.3.2/ SPATIO-TEMPORAL LBP

Dynamic or temporal textures are textures with motion that exhibit some stationary properties in time. The major difference between a dynamic texture (DT) and an ordinary texture is that the notion of self-similarity, central to conventional image texture, is extended to the spatio-temporal domain, thus DT combines appearance and motion simultaneously. DTs encompass the different difficulties mentioned in 3.2.3 such as moving trees, snow, rain, fog, crowd etc. Therefore, we adapt DT to model the varying appearance of dynamic scenes with time.

Several approaches have been developed to represent dynamic textures and a review of these methods can be found in [18]. In our work, we model DT using local binary patterns computed in orthogonal planes (LBP-TOP). The LBP-TOP operator extends LBP to temporal domain by computing the co-occurrences of local binary patterns on three orthogonal planes (LBP-TOP) such as XY, XT and YT planes (step 3a). The XT and YT planes provide information about the space-time transitions while the XY plane provides spatial information. LBPT-TOP uses three orthogonal planes which intersect at the center pixel. The LBP code is extracted from each separate plane and then concatenated into a single histogram. The procedure is illustrated in Figure. 3.11.

With the basic LBP operator a circular sampling is used with a single radius parameters. However, setting the radius in the time axis to be equal to the radius in the space axis is not reasonable for dynamic textures as mentioned in [76]. Assume for example that we have a DT with an image resolution of over 300 by 300, and a frame rate of less than 12. In a neighbouring area with a radius of 8 pixel in the X and Y axis the texture might still keep its appearance, but within the same temporal intervals in the T axis, the texture changes drastically. So there must be different radius parameters in space and time domains. In the XT and YT planes, different radii can be assigned to simple neighbouring points in space and time, so the tradition circular sampling is extended to elliptical sampling.

For a spatio-temporal volume $X \times Y \times T$, a histogram of DT is defined as follows :

$$H_{i,j} = \sum_{x,y,t} I\{f_j(x, y, t) = i\}, \tag{3.33}$$

where $i = 0, .., n_j - 1; j = 0, 1, 2$ and n_j is the number of different labels produced by the LBP operator in the jth plane ($j = 0$ for XY, $j = 1$ for XT and $j = 2$ for YT). $f_i(x, y, t)$ is the LBP code of central pixel (x, y, t) in the jth plane, and $I\{A\} = \begin{cases} 1 & \text{if A is true} \\ 0 & \text{if A is false} \end{cases}$.

The obtained histograms are finally normalized to get a coherent description of the dynamic texture. The XY plane contains information about the appearance, and in the planes XT and YT planes co-occurrence statistics of motion in horizontal and vertical directions are included.

3.3.3/ SPATION-TEMPORAL SALIENCY USING LBP-TOP DESCRIPTOR

After modelling the dynamic textures using LBP-TOP, we compute the ST- saliency using center-surround mechanism which is a biologically inspired formulation. Indeed, it has been proven that the motion contrast is processed in the Middle-Temporal (MT) area with a center-surround mechanism. Center-surround is a discriminant formulation in which the feature distribution of the center of visual stimuli is compared with the feature distribution of surrounding stimuli.

For each pixel location $l = (x_c, y_c)$, we extract a center region r_C and a surrounding region r_S both centred at l. We set the size of the surrounding region to be six time larger than the center region. We then compute the feature distributions $\mathbf{h}_c$ and $\mathbf{h}_s$ of both regions as histograms and define the saliency of pixel l as the dissimilarity between these two distributions. More specifically, the saliency $S(l)$ of pixel at location l is given by :

$$S(l) = \chi(\mathbf{h}_c, \mathbf{h}_s) = \sum_{i=1}^{B} \frac{(\mathbf{h}_c(i) - \mathbf{h}_s(i))^2}{(\mathbf{h}_s(i) + \mathbf{h}_c)/2}, \tag{3.34}$$

where $\mathbf{h}_c$ and $\mathbf{h}_s$ are the histogram distributions of r_C and r_S respectively, B is the number of bins of the histogram, and χ is the Chi-square distance measure.

Note that we separately apply center-surround mechanism to each of the three planes XY, XT and YT. Hence, we compute three different saliency maps based on the three distributions derived from LBP-TOP(step 3b).

The final step of the method consists in fusing the previous three maps into a single spatio-temporal saliency map. This is done in two steps. First, the two maps containing temporal information, i.e. the saliency maps from XT and YT planes, are fused to get a dynamic saliency map(step 3c). Then, this dynamic saliency map is fused with the static saliency map from the XY plane(step 3d). It is clear that the fusion technique adopted affects directly the quality of the obtained final spatio-temporal saliency map. In Section 3.4, we discuss this combining different fusion methods. In this section, we will use the *Dynamic Weighted Fusion* (DWF) scheme and this choice will be made clear in the next chapter, as this technique shows best performance when compared against other fusion techniques. In DWF the weights are computed by taking the ratio between the means of both the maps to combine, so they are updated from frame to frame. Let S_{XT} and S_{YT} be saliency maps obtained in the XT and YT planes respectively. They are fused into a dynamic saliency map M_D as follows :

$$M_D = \alpha_D S_{YT} + (1 - \alpha_D) S_{XT}, \tag{3.35}$$

where $\alpha_D = \frac{mean(S_{YT})}{mean(S_{XT}) + mean(S_{YT})}$.

The obtained dynamic map M_D and the static map $M_S = S_{XY}$ are fused in a similar manner. Some example of saliency maps obtained using LBP-TOP in three orthogonal planes such (i.e XT-plane, YT- plane and XY-planes) are shown in Fig. 3.12. Figure. 3.13 shows the example of fused spatio-temporal saliency maps.

3.3.4/ SPATIO-TEMPORAL SALIENCY USING COLOR AND TEXTURE FEATURES

We observe that the spatial saliency map derived from the XY plane fails to highlight salient objects of some scenes because LBP-TOP does not use color features. Figure. 3.13

shows examples of saliency maps obtained from LBP-TOP features. As it can be seen in the second column of Figure. 3.13, the saliency map obtained in the XY plane fails to highlight objects of interest, for example the moving persons in the second row of the figure.

This observation motivates us to replace the LBP features computed in XY plane by color features and compute the spatial saliency map following the context-aware method of Goferman et al. [27] explained in Section. 3.2.1, since this static saliency detection performs well.

The temporal saliency is here computed as described in Section 3.3.3. That is, we consider only the two planes XT and YT which give motion information in temporal direction with respect to X and Y planes separately. The two maps, one for each plane, are then fused into a single dynamic saliency map using the DWT fusion scheme explained in Section 3.3.3(second proposal for step 3d).

3.3.4.1/ POST-PROCESSING

After obtaining the spatial and temporal saliency maps, denoted respectively M_S and M_D, they are fused into the final spatio-temporal saliency map as :

$$M_F = \alpha M_D + (1 - \alpha) M_S, \tag{3.36}$$

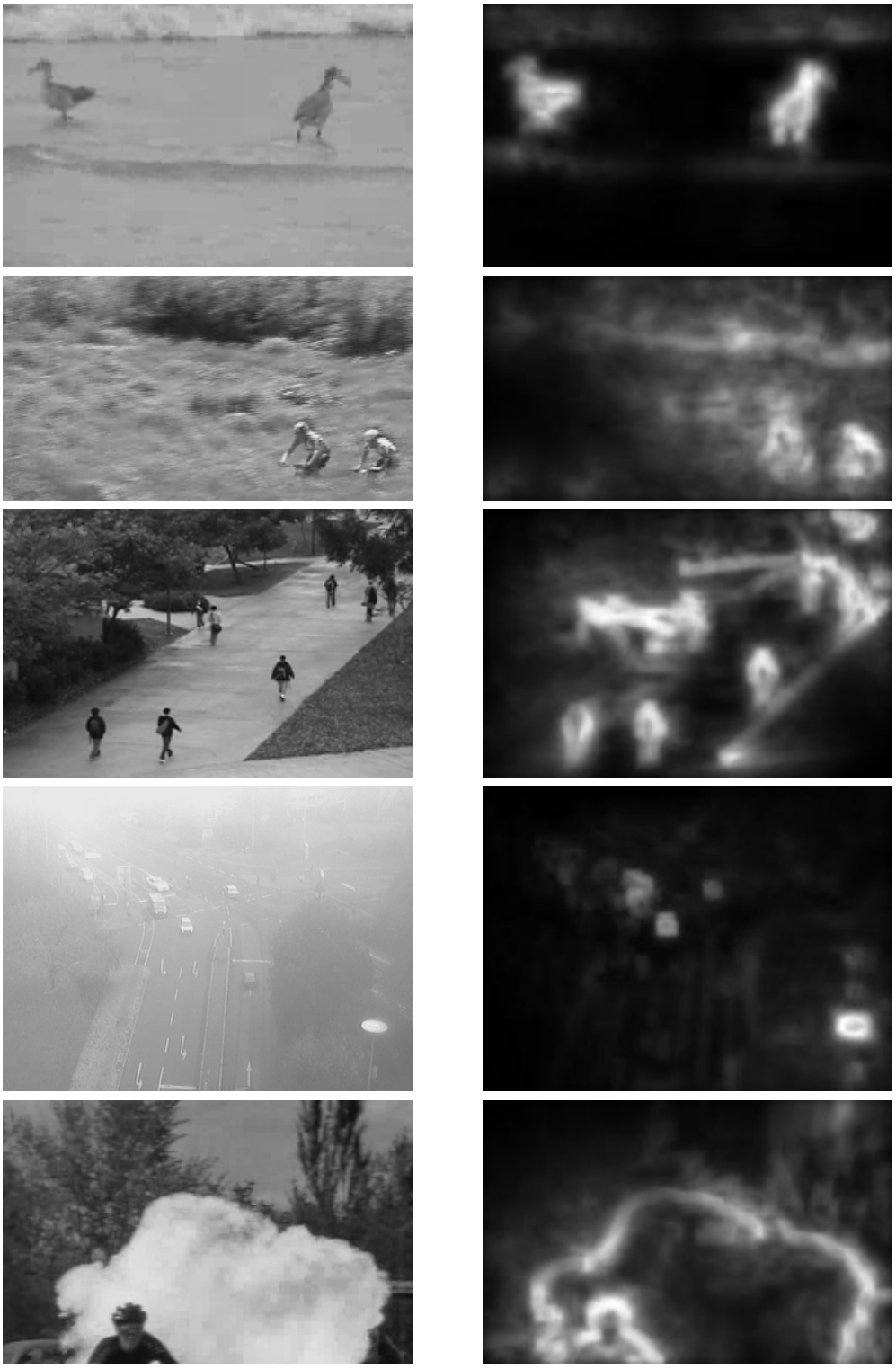
with $\alpha = \frac{mean(M_D)}{mean(M_D)+mean(M_S)}$, and M_F the final saliency map.

The last step of our method consists in applying a post-processing scheme with the goal of suppressing isolated pixels or group of pixels with low saliency values (step 3d). We start by finding pixels whose saliency value is above a defined threshold (0.5 in our experiments, the final saliency map M_F being normalized to have values in $[0, 1]$). Then, we compute the spatial distance $D(x, y)$ from each pixel to the nearest non-zero pixel in the thresholded map. The spatio-temporal saliency map M_F is finally refined using the following equataion :

$$M_F(x, y) = e^{\frac{-D(x,y)}{\lambda}} \times M_F(x, y), \tag{3.37}$$

where λ is a constant set to $\lambda = 0.5$. The influence of this last parameter is studied in the next section showing experimental results.

A block diagram of the proposed spatio-temporal saliency detection method is shown in Figure 3.14 and the spatio-temporal saliency of the proposed methods are illustrated in Figure 3.15.



(a) (b)

FIGURE 3.2 – Examples of static saliency detection. (a) Original frame (b) Static saliency map.

FIGURE 3.3 – Dominant motion estimation on two successive video frames. Top row : original video frames. Bottom left : warped image with estimated affine parameters ; Bottom right : detected relative motion.

FIGURE 3.4 – The dense optical flow estimation on successive video frames. From left to right : original video frames and motion estimation of each pixel.

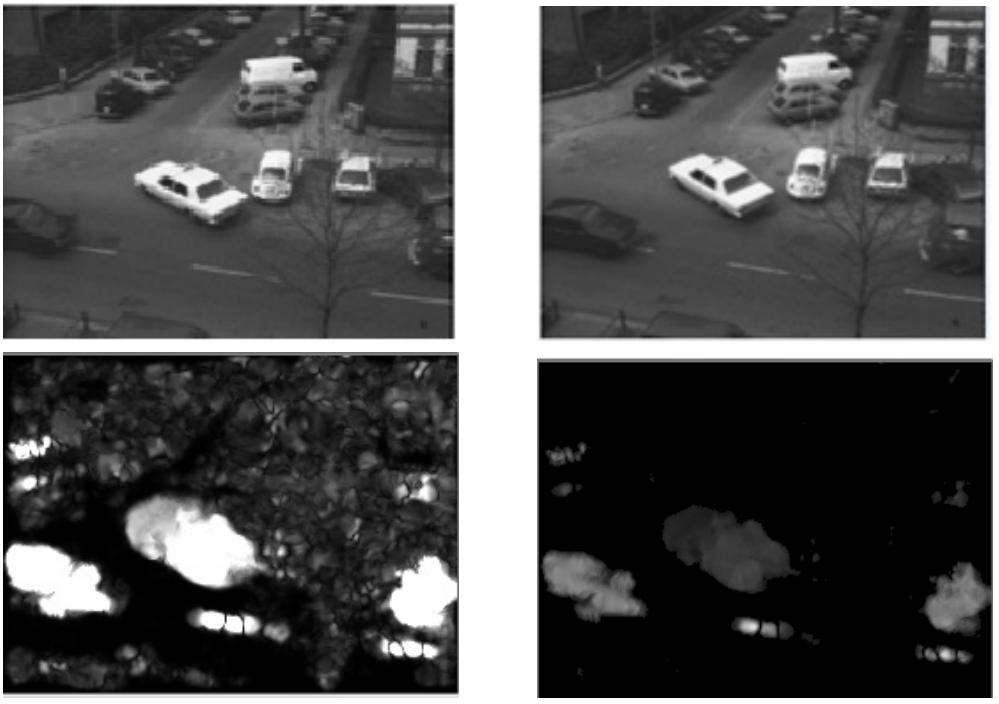
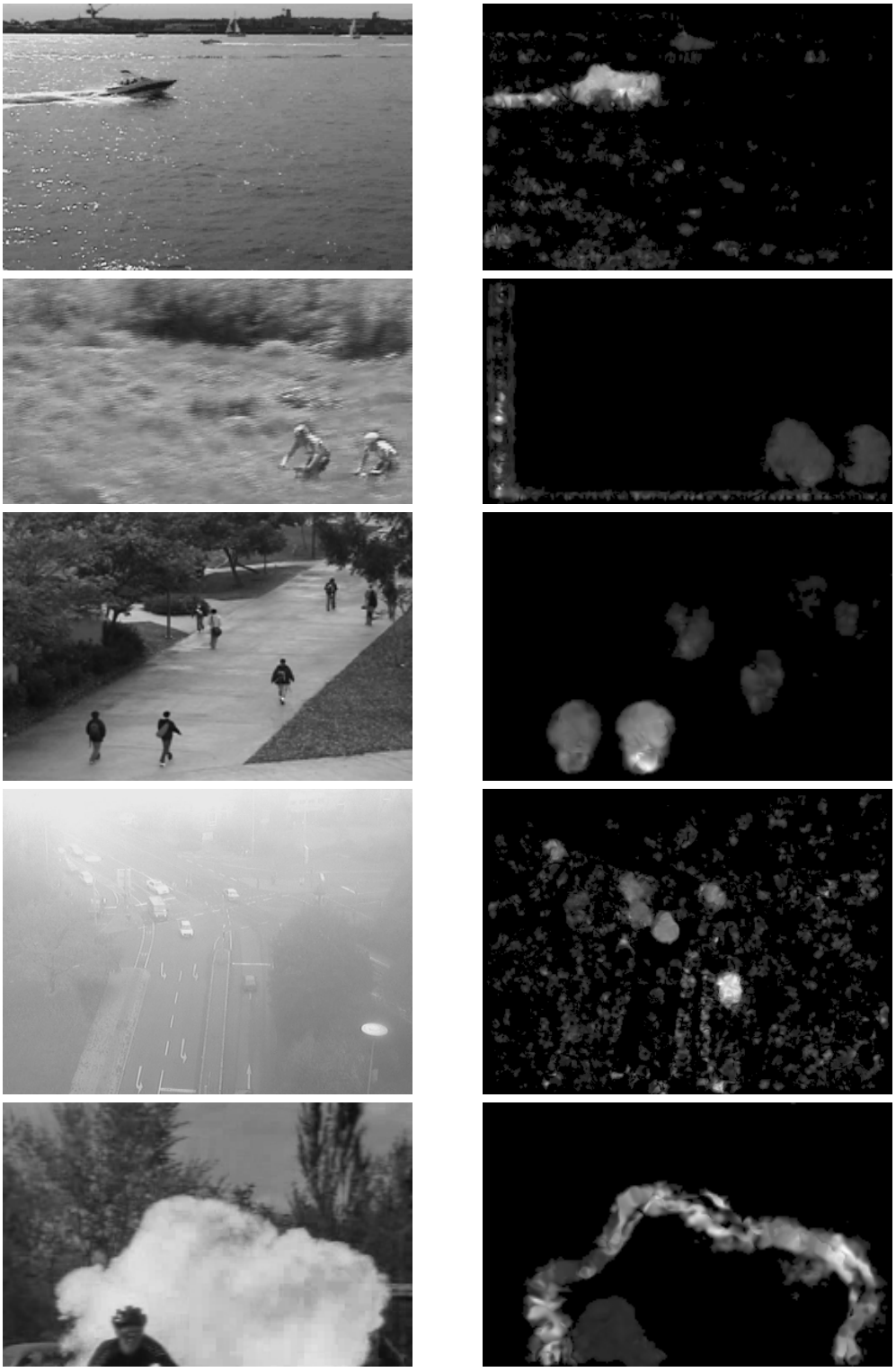


FIGURE 3.5 – Dynamic saliency map estimation. Top row : original video frames ; Bottom left : median filtered image motion map ; Bottom right : normalized motion map.



(a) (b)

FIGURE 3.6 – Examples of dynamic saliency detection. (a) Original frame (b) Dynamic saliency

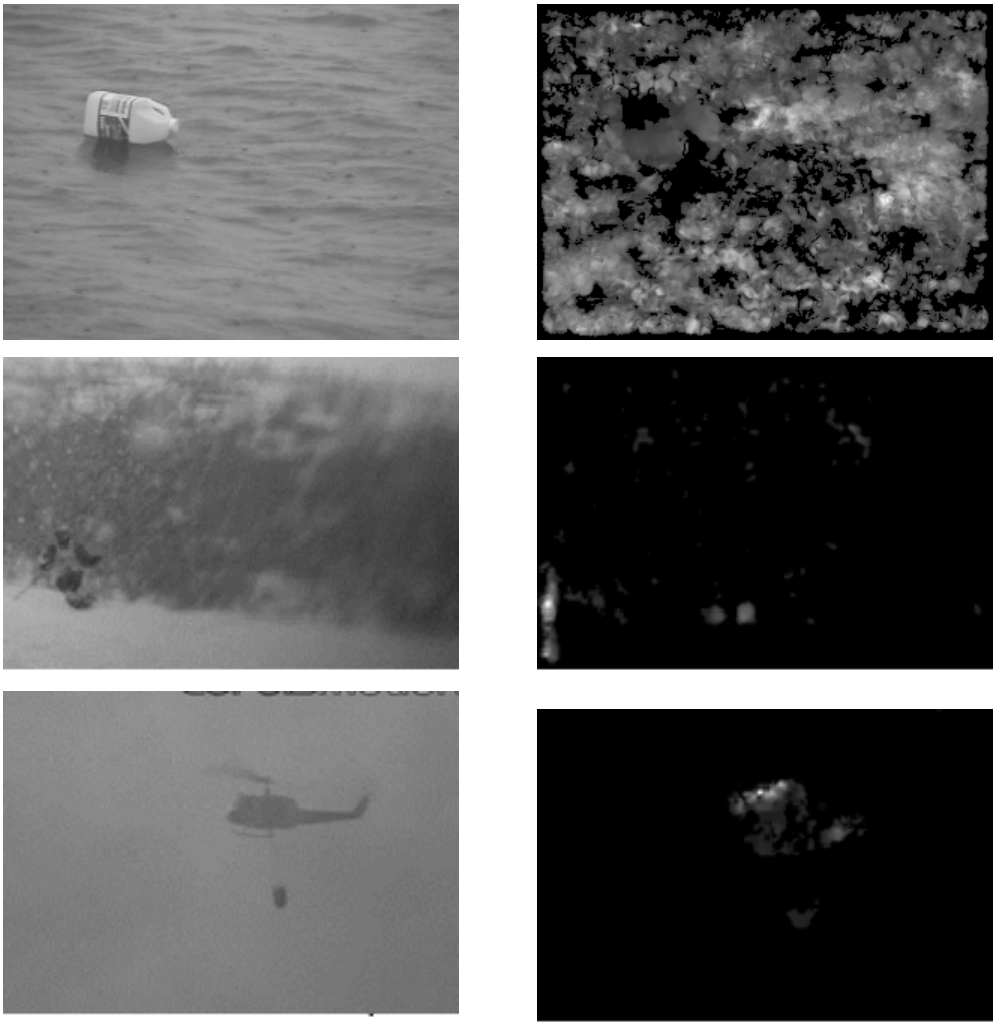


FIGURE 3.7 – Examples of dynamic saliency maps failures using optical flow in complex dynamic background scenes.

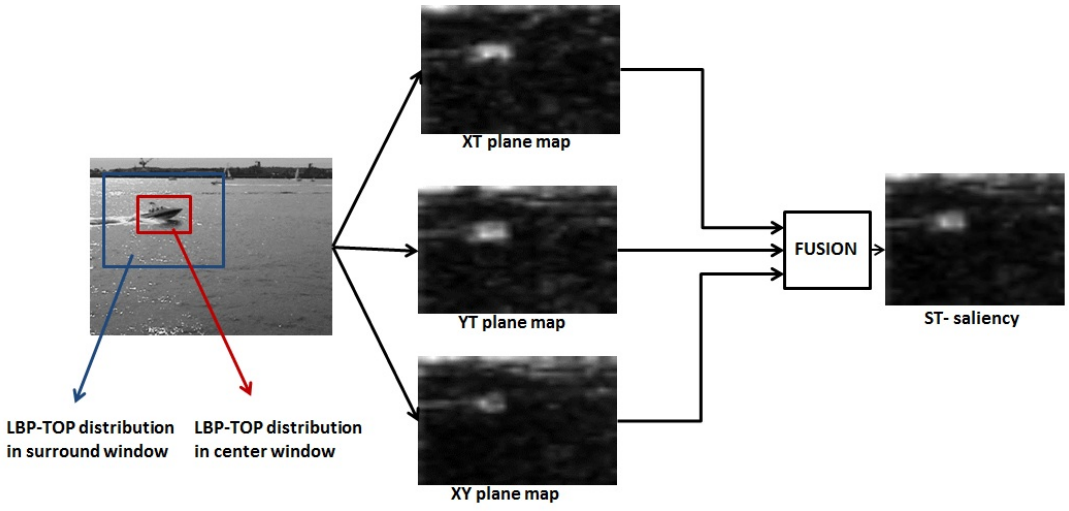


FIGURE 3.8 – Spatio-Temporal saliency overview.

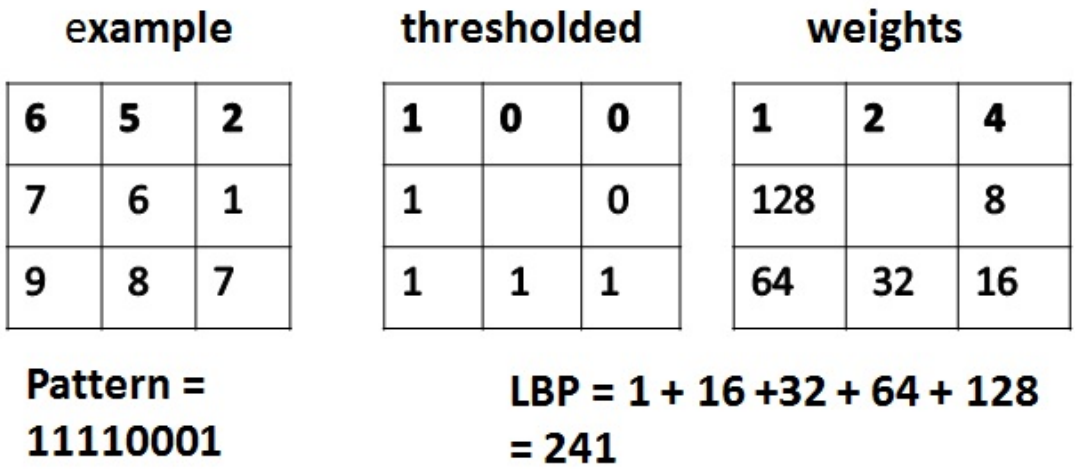


FIGURE 3.9 – Basic LBP description.

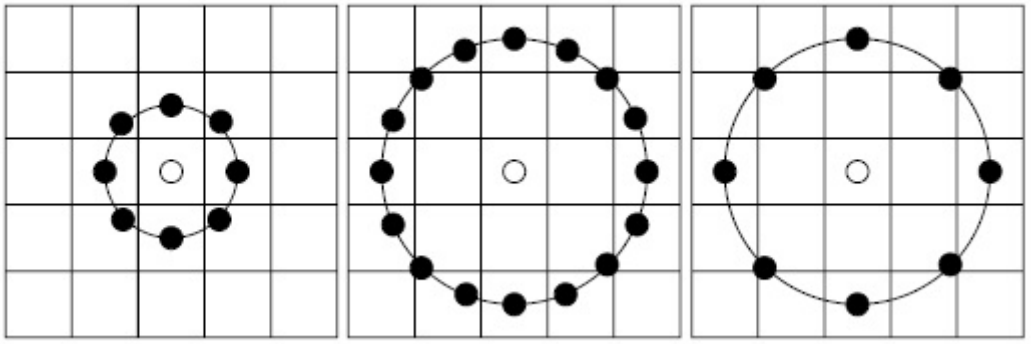


FIGURE 3.10 – Circular sampling.(source : [58])

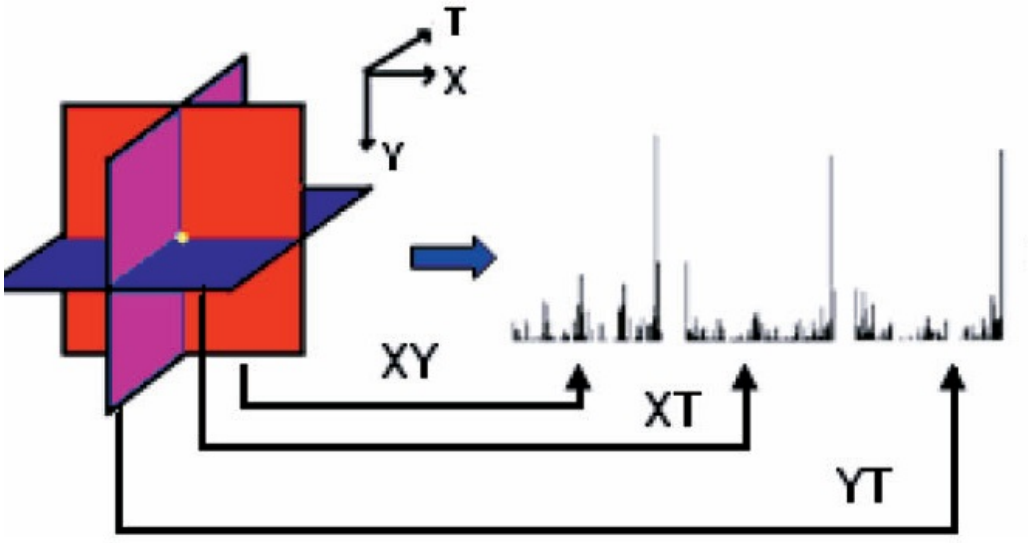


FIGURE 3.11 – LBP-TOP computation using three orthogonal planes(source : [58]).

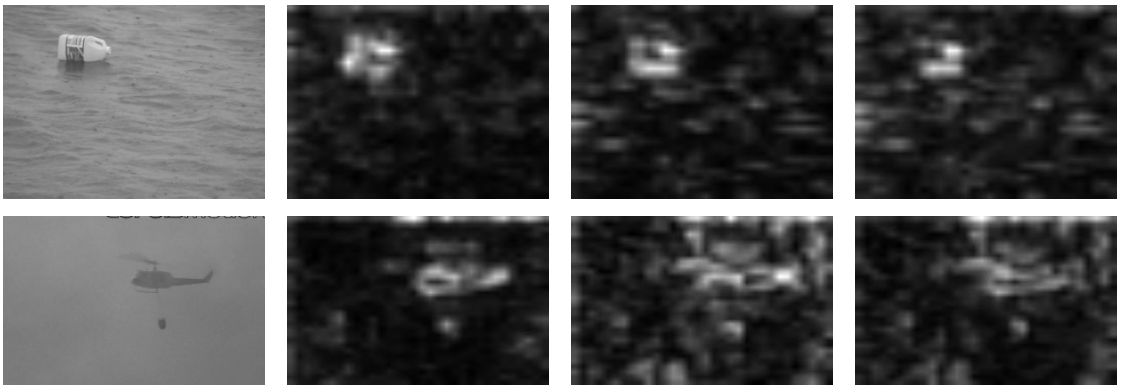


FIGURE 3.12 – Examples of saliency maps using LBP-TOP. From left to right : original video frame, saliency map in XY plane, in XT plane, in YT plane, and final spatio-temporal saliency map.

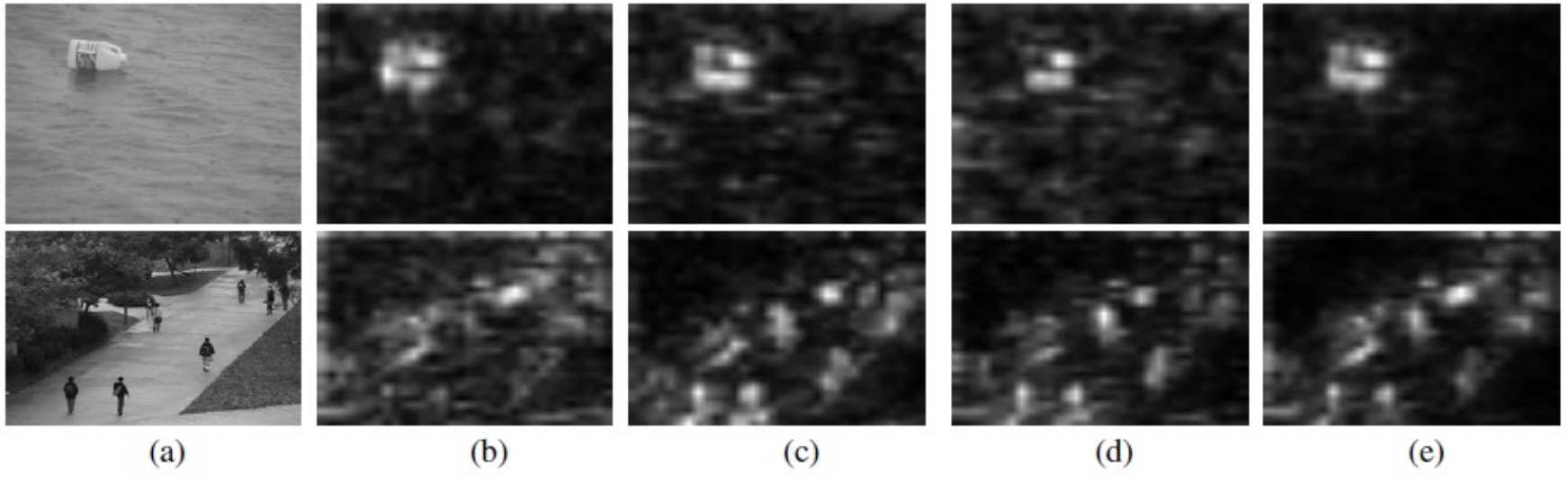


FIGURE 3.13 – Examples of spatio-temporal saliency detection with LBP-TOP features. (a) Original frame ; (b) saliency map in XY plane ; (c) saliency map in XT plane ; (d) saliency map in YT plane ; (e) fused spatio-temporal saliency map.

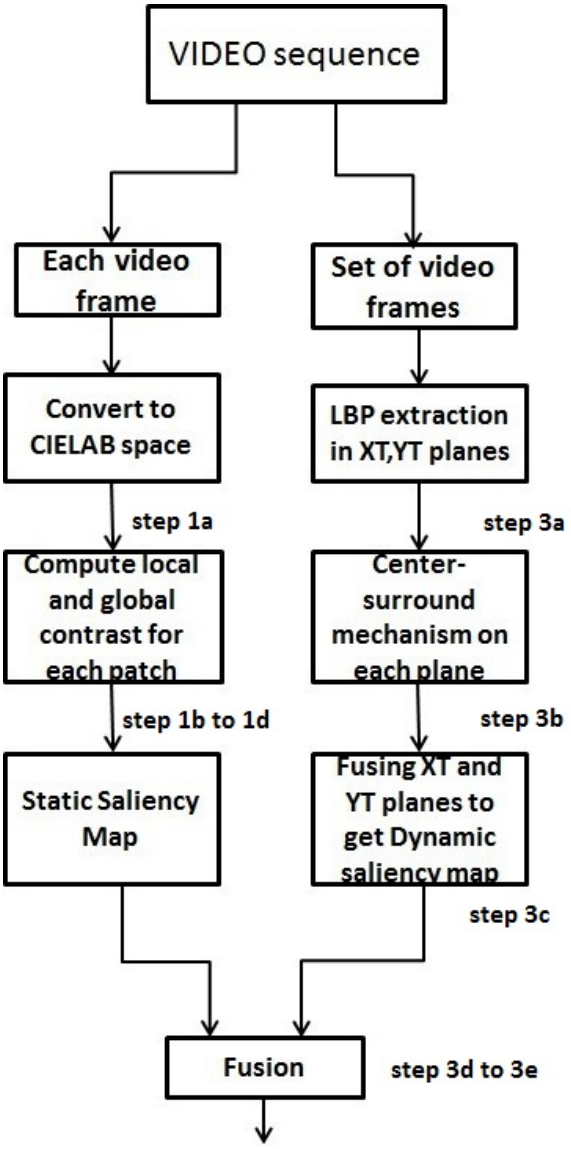


FIGURE 3.14 – Combined model with color and texture features.

3.4/ About Fusion

As mentioned in the introduction, a spatio-temporal saliency map is often obtained by fusing a static saliency map with a dynamic saliency map. The fusion balances the information from both maps, and it is a crucial step and it plays a major role in getting accurate spatio-temporal saliency map.

In the bottom-up visual attention process, low-level features are processed separately to produce feature maps, which are then fused into a master saliency map that shows the most salient regions among all feature maps computed spatially and temporally. The fusion step is an important component in bottom-up spatio-temporal saliency modeling. Different fusion methods have been used by authors in literature. Here we briefly described the most common ones below. In the following, the static saliency map, the dynamic saliency map and the fused spatio-temporal saliency map are referred to as M_S, M_D and M_F respectively.

3.4.1/ Different Fusion Techniques

Mean fusion [44] : this fusion method takes the pixel average of both static and dynamic saliency maps.

$$M_F = (M_S + M_D)/2. \tag{3.38}$$

Max fusion [51] : this is a "winer takes all" (WTA) strategy in which the maximum value between the two saliency maps is taken for each pixel.

$$M_F = max(M_S, M_D). \tag{3.39}$$

Multiplicative fusion [51] : a pixel by pixel multiplication is done, corresponding to a logical *AND*.

$$M_F = M_s \times M_D. \tag{3.40}$$

Maximum skewness fusion [51] : this fusion technique takes advantage of the characteristics of the static and the dynamic saliency maps. The static pathway is modulated by its maximum value α. The dynamic saliency map is modulated by its skewness value β. The reinforcement term γ gives more importance to the areas that are salient both in a static and dynamic way.

$$M_F = \alpha M_S + \beta M_D + \gamma(M_S \times M_D), \tag{3.41}$$

with $\alpha = max(M_S)$, $\beta = skewness(M_D)$ and $\gamma = \alpha\beta$.

Binary thresholded fusion [45] : first, a binary mask M_B is generated by thresholding the static saliency map (the mean value of M_S is used as threshold). The binary mask is used to exclude spatiotemporal inconsistent areas and to enhance the robustness of the final saliency map when the global motion parameters are not estimated properly.

$$M_F = max(M_S, M_D \cap M_B). \tag{3.42}$$

Motion priority fusion [55] : this fusion technique is based on *motion priority* which states that a viewer might pay more attention to the motion caused by a moving object even when the static background is more attractive [55]. The perception of moving objects saliency increases nonlinearly with motion contrast and shows significant saturation and threshold effects.

$$M_F = (1-\alpha)M_S + \alpha M_D, \tag{3.43}$$

with $\alpha = \lambda e^{1-\lambda}$ and $\lambda = max(M_D) - mean(M_D)$.

Dynamic weight fusion [72] : in this fusion, the weights of the static and dynamic saliency maps are determined by the ratio between the means of both maps for each frame. The final saliency map M_F is given by :

$$M_F = \alpha M_D + (1-\alpha)M_S, \tag{3.44}$$

Where $\alpha = \frac{mean(M_D}{(mean(M_D + mean(M_S)}$.

Shannon fusion [29] : this fusion technique is based on information theory. The feature maps will represent the sparse distribution of conspicuous locations over the visual field and they are taken as informative events. The probability $p(M)$ of a map is computed by taking the ratio between the number of pixels with their values above a certain threshold τ and the total number of pixels :

$$p(M) = \frac{\#(M(i,j) > \tau)}{\#(M(i,j))}.$$

Then the importance of a map is given by $I(M) = -log(p(M))$.

The maps are finally fused as :

$$M_F = \alpha_S I(M_S) M_S + \alpha_D I(M_D) M_D, \tag{3.45}$$

where the weights α_S and α_D are given by $\alpha_i = max(M_i)I(M_i)$, $I(M_i)$ being the importance of the saliency map M_i.

Scale invariant fusion [38] : in this fusion technique, the input images are analyzed at three different scales from 32×32 to 128×128 to original size of image. The smaller map represents a coarser resolutions. In this coarser resolution, the gloabal features are emphasized and the detailed local features are suppressed. The local features are retained in the scene in finer resolution. The scale invariant saliency map is obtained through multi-scale analysis [38]. Three fused maps are obtained which are finally combined linearly into the final spatio-temporal saliency map.

$$M_F^l = alphaM_S + (1-alpha)M_D \tag{3.46}$$

$$M_F = \sum_{l=1}^{3} w_l M_F^l, \tag{3.47}$$

where M_F^l is the fused map at scale l and the coefficients of the linear combination are $w_1 = 0.1$, $w_2 = 0.3$ and $w_3 = 0.6$.

To see the performance of the these different fusion techniques, we did the performance evaluation on the dynamic dataset [9]. The performance evaluation is explained in the next chapter.

(a) (b)

FIGURE 3.15 – Examples of spatio-temporal saliency detection map using color and texture features. (a) Original frame (b) Spatio-temporal saliency map.

4

Experiments and Results

This chapter describes and discusses various experiments performed to evaluate the proposed spatio-temporal saliency detection methods. To validate the proposed method, we performed experiments in two ways. Firstly, we use a publicly available dataset of dynamic scenes. This dataset contains ground truth segmentation of the salient objects for each frame of sequences, thus allowing us to evaluate the ability of the method in locating interesting foreground objects in complex scene. Secondly, we evaluated our model on another dataset in which the ground truth is given as eye tracking data, i.e. human observers fixations. This evaluate the performance of the model in predicting human fixation when viewing a video. The performance of the proposed method are compared with other state-of-the-art methods.

4.1/ Experimental Setup

To evaluate the different spatio-temporal saliency models, we select two publicly available complex video scenes datasets : SVCL dataset [9] and ASCMN dataset [63]. The former, SVCL, is composed of twelve video sequences containing dynamic entities such as fog, moving trees, snowing, smoke, and moving water. This dataset is used to evaluate the ability of the method in locating interesting foreground objects in a complex scene. The later, ASCMN, contains video sequences divided into different categories such as abnormal, surveillance, crowd and moving noise. It contains humans fixation data which are recorded using eye tracking equipment. The main objective of using this dataset is to detect the salient regions which tell us where the human eyes are looking. Both datasets are described in section 4.2.

Several evaluation metrics are used and are described in section 4.3. Then, in section 4.4 we evaluate different techniques for the fusion of static and dynamic saliency maps into one single spatio-temporal saliency map. Section 4.5 describes evaluation of salient object detection of the proposed method and its comparison against state-of-art methods. Finally, the evaluation of eye tracking data to detect the salient regions is explained in section 4.6.

TABLE 4.1 – ASCMN dataset video types

Video classes	Description
1) ABNORMAL	Some moving blobs have different speed or detection compared to the main stream
2) SURVEILLANCE	Classical surveillance camera with no special motion event
3) CROWD	Motion of more or less dense crowds
4) MOVING	Videos taken with a moving camera
5) NOISE	No motion during several seconds followed by sudden important motion

TABLE 4.2 – Details of ASCMN dataset

Video classes	**Number of Video per class**	**Total frames per class**
1) ABNORMAL	5	589
2) SURVEILLANCE	4	3220
3) CROWD	5	1539
4) MOVING	4	462
5) NOISE	6	4370

4.2/ EVALUATION DATASET

This section gives a brief description of the two datasets used in our evaluation :
– The ASCMN (Abnormal-Surveillance-Crowd-Moving-Noise) dataset [63] which contains humans fixation obtained using eye tracking equipment.
– The SVCL (Statistical Visual Computing Lab) dataset [9] which results of manual segmentation of salient objects.

4.2.1/ ASCMN DATASET

The ASCMN dataset is a collection of videos from various sources such as Itti's CRCNS database [36], SVCL database [9] and a standard complex-background video surveillance database [43]. This dataset provides data which cover a wider spectrum of video types, than the existing databases, and accumulates previously published videos which are suitable for dynamic saliency evaluation. It contains totally 24 videos, together with eye tracking data collected from 13 human observers using eye tracking apparatus. The ASCMN dataset is divided into 5 classes of sequences which are described in Table. 4.1. Some sample frames of these 5 different classes of videos are shown in Figure. 4.1. The gaze positions of the observers are recorded and superimposed on the initial video for all the videos as can be seen in Figure 4.2. The figure also shows the heat maps of the gaze positions. The details of the video per class and the total number of frames per class are illustrated in Table 4.2.

4.2.2/ SVCL DATASET

The SVCL dataset, from the Statistical Visual Computing Lab at University of California San Diego, contains natural videos which are usually composed of dynamic entities such

Video classes	Total frames per video	video characteristics
Birds	71	camera motion and moving water
Boats	31	moving water
Bottle	31	moving water and camera motion
Cyclists	30	camera motion and swaying trees
Chopper	100	camera motion and fog
Freeway	44	complete fog
Peds	170	multiple peds walking
Ocean	176	moving water background
Surfers	41	waves and camera motion
Skiing	111	snowing and camera motion
Jump	81	background with smoke
Traffic	190	noisy with fog

TABLE 4.3 – Details of SVCL dataset

as swaying trees, crowd, moving water, waves, snow and smoke filled environments. The authors in [48] created this dataset for different applications such as background subtraction, salient objects detection, dynamic texture modelling and object tracking. The ground truths used in this dataset are manually segmented objects for each frames. The details of each video and its characteristics are illustrated in Table 4.3

Each frame has a ground truth which is in binary image. The object in the ground-truth frame is indicated by 1 and the background is indicated by 0. This dataset is used to evaluate the robustness of saliency models in correctly detecting salient objects in complex dynamic environments. Some sample frames of different types of videos and the corresponding ground truths are shown in Figure 4.3.

4.3/ EVALUATION MEASURES

The usual metrics used in state-of-the-art methods for evaluation of saliency detection methods are Normalized Scanpath Saliency (NSS), Area Under ROC Curve (AUC) and Kullback-Leibler Divergence (KL). In our experimental evaluation, we choose these three measures. The main reason for choosing more than one evaluation measure is to ensure that the discussion about the results is as independent as possible from the choice of the metrics. The results of the different evaluation metrics is not necessarily the same, but when two metrics show similarities, then it is easy to interpret the robustness of the methods.

These metrics are used do evaluate the performance of the models relatively to the two kinds of experiments proposed :

– The detection of salient objects in videos using SVCL dataset.
– The prediction of human eyes fixation in videos using ASCMN dataset.

4.3.1/ Receiver Operating Characteristics (ROC)

The Receiver Operating Characteristics (ROC) analysis [22] is one of the most popular and most widely used method in the community for assessing the degree of similarity of two saliency maps. It is a graphical plot which illustrates the performance of a binary classifier system as its discrimination threshold is varied. It is created by plotting the fraction of true positives out of the total actual positives (TPR = true positive rate) versus the fraction of false positives out of the total actual negatives (FPR = false positive rate), at various threshold settings. TPR is also known as *sensitivity or recall*, whereas FPR is also known as the fall-out and can be calculated as one minus the more well known *specifity*.

Lets consider a two-class prediction problem, or a binary classification problem, in which the outcomes are labeled either as positive (p) or negative (n). There is four possible outcomes for a binary classifier :
– **true positive** (TP) : the predicted outcome is p and the actual value is also p ;
– **false positive** (FP) : the predicted outcome is p and the actual value is n ;
– **true negative** (TN) : the predicted outcome is n and the actual value is also n ;
– **false negative** (FN) : the predicted outcome is n and the actual value is p.

From the Fig. 4.4, the TPR and FPR are given by :

$$TPR = \frac{TP}{P} = \frac{TP}{(TP + FN)}. \tag{4.1}$$

$$FPR = \frac{FP}{N} = \frac{FP}{(FP + TN)}. \tag{4.2}$$

A good prediction method would give a TPR of 1 at a FPR of 0, yielding a point in the upper left corner of the ROC space that corresponds to a perfect classification. A completely random guess would give a point along a diagonal line from the left bottom to the top right corners. The diagonal divides the ROC space. Points above the diagonal represent good classification results (better than random), points below the line poor results (worse than random). Thus, a measure of performance derived from the ROC curve is the AUC (Area Under Curve) which is equal to the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one (assuming 'positive' ranks higher than 'negative').

4.3.2/ Normalized Scanpath Salience (NSS)

The Normalized Scanpath Saliency (NSS) [57] is a metric that involves a saliency map and a set of fixations. The main idea behind this metric, is to measure the saliency values at fixation locations along subject's scanpath. Each saliency map is linearly normalized to have zeros mean and unit standard deviation :

$$Z_{SM}(x) = \frac{SM(x) - \mu}{\sigma}, \tag{4.3}$$

where Z_{SM} is the standardized saliency map, $SM(x)$ is the saliency value at pixel location x, and $\mu = \frac{1}{|I|} \sum SM(x)$, where $|I|$ is the number of pixels in the image.

Next the normalized saliency values are extracted from each point corresponding to the fixation locations along a subject's scanpath and the mean of these values or normalized scanpath salience (NSS) is taken as a measure of the correspondence between the saliency map and scanpath.

Then NSS value for a given human fixation location x_k is computed on a small neighbourhood centred on that location :

$$NSS(x_k) = \sum_{j \in \pi} K_h(x_k - x_j) Z_{SM}(x_j), \tag{4.4}$$

where K is a kernel with a bandwidth h and π is a neighborhood.

The NSS score of a frame is the average value of all fixations of an observer and is given by :

$$NSS = \frac{1}{M} \sum_{k=1}^{M} NSS(x_k), \tag{4.5}$$

with M the number of fixations points.

4.3.3/ Kullback-Leibler Divergence (KL)

The Kullback-Leibler Divergence is a metric which estimates dissimilarity between two probability density functions [40]. Given two discrete distributions R and P with the corresponding probability density functions r_k and p_k, the KL divergence is computed as the relative entropy of P towards R :

$$KL(R, P) = \sum_k p_k \log \frac{r_k}{p_k}. \tag{4.6}$$

In the evaluation of attention models, the distribution are given by the eye fixations points and the saliency maps produced by the model. Let FM be the probability distribution of the heat map from eye tracking data, and SM be the probability distribution of the saliency map. The distributions are normalized and they are given by :

$$SM(x) = \frac{SM(x)}{\sum_{x=1}^{X} SM(x) + \epsilon}, \tag{4.7}$$

$$FM(x) = \frac{FM(x)}{\sum_{x=1}^{X} FM(x) + \epsilon}, \tag{4.8}$$

where X is the number of pixels and ϵ is a small constant to avoid division by zero.

The KL-divergence measure is computed between these two distributions to know whether the saliency map which is computed from the saliency model matches human fixations. The KL-divergence is non-linear and varies in the range of zero to infinity. A zero value indicates that the two probability density functions are strictly equal.

4.4/ Evaluation of Fusion Techniques

In this section, we evaluate the performances of different fusion approaches described in Section 3.4 to compute spatio-temporal saliency maps. For a quantitative evaluation, we use the SVCL dataset of complex dynamic scenes [48]. The dataset contains twelve video sequences captured with different challenges such as dynamic background scenes with moving trees, snow, smoke, fog, pedestrians, waves in the sea and moving cameras.

For each sequence, a manual segmentation of the salient objects is available for every frame and served as ground truth. We can therefore evaluate the different fusion techniques by generating Receiver Operating Characteristic (ROC) curves and evaluating the Area Under ROC Curve (AUC). For each fusion technique, the obtained spatio-temporal saliency map is first normalized to the range $[0, 1]$, and then binarized using a varying threshold $t \in [0, 1]$. With the binarized maps, we compute the true positive rate (TPR) and false positive rate (FPR) with respect to the ground truth data.

Table 4.4 summarizes the results obtained with all sequences by the different fusion techniques whereas the visual comparison of these fusion techniques are shown in Figure. 4.5. We observe that the best performances are obtained by the *Mean* [44], *Scale Invarian*t [38], *Max* [51] and *Dynamic Weight* [72] fusion methods respectively. In particular, the *Mean* fusion technique achieves an average AUC value of 0.9325 for all twelve sequences. Those fusion methods take the best of each saliency map (static and dynamic) in the final spatio-temporal map : the static saliency value is given more importance if it is higher at a given position and vice-versa. On the contrary, the *Motion Priority* [55] and the *Multiplication* [51] fusion techniques give the lower performances. In particular, the *Motion Priority* method gives an average AUC value of 0.7943 for all sequences, which is 17% less than the value obtained by the *Mean* fusion technique. This can be explained by the fact that this fusion approach gives more importance to motion information. Therefore, when the motion contrast is not estimated properly, the final saliency map is not accurate. This problem can be observed with the *Skiing* sequence for which the *Binary Threshold* and the *Motion Priority* fusion methods achieve AUC values of 0.9807 and 0.4905 respectively. The salient object segmentation results for those two fusion methods are shown in Fig. 4.6. As can be seen, for this sequence with low motion contrast *Motion Priority* fusion method fails to localize the target due to the incorrect estimation of the dynamic saliency map. The red box shows the ground truth location of the salient object while the green box is the output of the estimated spatio-temporal saliency detection method.

Analyzing the sequences individually, we see that the best and worst performances are obtained with the *Boats* and *Freeway* sequences, respectively, with average AUC values of 0.9826 and 0.6688 for all fusion techniques. The *Boats* sequence shows good color and motion contrasts, so both static and dynamic maps are estimated correctly (as shown in Fig. 4.9). As a consequence, all fusion techniques perform well. On the other hand, the color contrast of the *Freeway* sequence is very limited. So fusion methods such as *Binary Threshold* (BTF) and *Dynamic Weight* which give high importance to the static map perform poorly, with AUC values of 0.5087 and 0.5456 respectively. For instance, in the BTF technique, the mean value of the static map is used to generated a binary mask which is then combined with the dynamic map. It is clear that if the static map is not accurate, the final spatio-temporal saliency map will be inaccurate as well.

The results show the consistency of fusion approaches that base decision on the scene's

characteristics as the final spatio-temporal saliency map takes the best of each individual saliency map (static and dynamic). This include *Mean*, *Scale Invariant*, *Max* and *Dynamic Weights* fusion methods. On the other hand, fusion techniques which are based on a strong a priori such as *Motion Priority* fusion achieve good results only when the underlying assumption is satisfied. Thus, they performances vary depending on the sequence.

It is clear that the accuracy of a spatio-temporal saliency map depends on the quality of both static and dynamic maps, which are based on the scene's contents. The ROC curves comparing performances of the different fusion techniques on the two sequences are shown in Figure. 4.8 and Figure. 4.7

FIGURE 4.1 – Sample videos of the 5 different classes in ASCMN database. From top to bottom : ABNORMAL motion with bikes and cars moving faster than people ; SURVEILLANCE videos ; CROWD motion with increasing density from left to right ; MOVING camera videos ; and NOISE.

FIGURE 4.2 – Sample images of ground truth data. The left column shows aggregated eye tracking results : each dot is the position of the eye gaze of video. The middle column shows the heatmaps and in the right column thresholded version of the heatmap is shown.

FIGURE 4.3 – Examples of videos and ground truth from SVCL dataset.

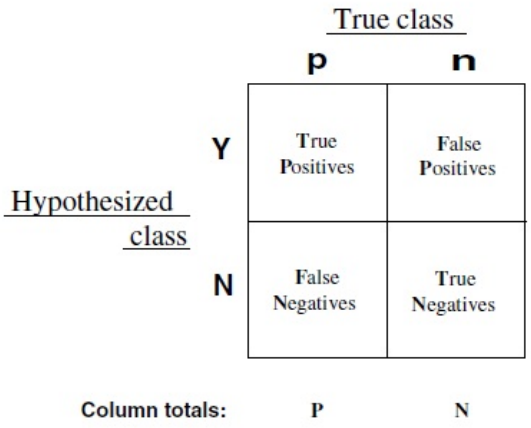


FIGURE 4.4 – Confusion and Common performance metric calculated from it

Sequence	Mean	Max	AND	MSF	BTF	DWF	MPF	ITF	SIF	Avg AUC
Birds	0.9713	0.9794	0.9023	0.9563	0.9852	0.9669	0.7639	0.9097	0.9245	0.9288
Boats	0.9891	0.9745	0.9867	0.9881	0.9695	0.9827	0.9808	0.9889	0.9829	**0.9826**
Cyclists	0.9628	0.9497	0.8862	0.9418	0.9533	0.9602	0.8394	0.9248	0.9498	0.9298
Chopper	0.9784	0.9847	0.6891	0.6956	0.9852	0.9850	0.6791	0.9628	0.9711	0.8812
Freeway	0.7128	0.6633	0.7023	0.7614	0.5087	0.5456	0.7581	0.6218	0.7452	**0.6688**
Peds	0.9608	0.9435	0.8984	0.9380	0.9441	0.9512	0.8852	0.9400	0.9558	0.9352
Jump	0.9395	0.9314	0.8949	0.9212	0.9459	0.9479	0.8535	0.8804	0.9197	0.9149
Ocean	0.8273	0.7465	0.8108	0.8126	0.7535	0.7810	0.8032	0.8063	0.8412	0.7980
Surfers	0.9453	0.9782	0.7993	0.9208	0.9844	0.9545	0.6251	0.9334	0.8757	0.8907
Skiing	0.9678	0.9784	0.5195	0.6491	0.9807	0.9796	0.4905	0.9394	0.9365	0.8268
Landing	0.9701	0.9524	0.9718	0.9703	0.9521	0.9579	0.9047	0.9353	0.9720	0.9541
Traffic	0.9645	0.9566	0.8860	0.9540	0.8736	0.9615	0.9477	0.9640	0.9593	0.9408
Avg AUC value	**0.9325**	0.9199	0.8289	0.8758	0.9030	0.9145	**0.7943**	0.9006	0.9200	

TABLE 4.4 – Fusion techniques evaluation results. Mean (Mean fusion), Max (Max fusion), AND (Multiplication fusion), MSF (Maximum skewness fusion), BTF (Binary thresholded fusion), DWF (Dynamic weight fusion), MPF (Motion priority), ITF (Information theory fusion), SIF (Scale invariant fusion).

FIGURE 4.5 – Visual comparison of different fusion techniques. From left to right : original video frame followed by Mean (Mean fusion), Max (Max fusion), AND (Multiplication fusion), MSF (Maximum skewness fusion), BTF (Binary thresholded fusion), DWF (Dynamic weight fusion), MPF (Motion priority), ITF (Information theory fusion), SIF (Scale invariant fusion).

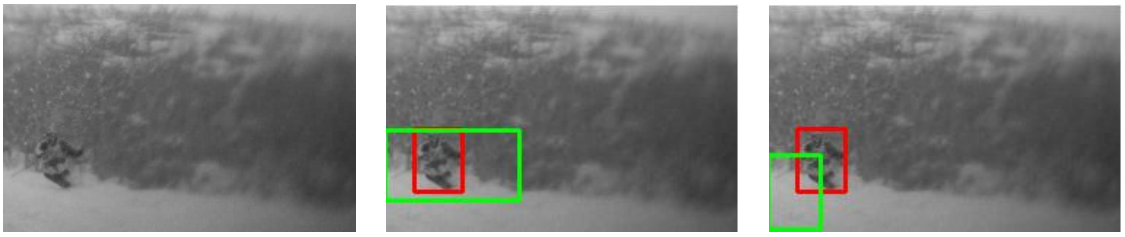


FIGURE 4.6 – Example of salient region segmentation with the *Skiing* sequence. From left to right : input frame ; detection with *Binary Threshold* and with *Motion Priority* fusion techniques. Red box indicates ground truth and green box indicates the detected salient region.

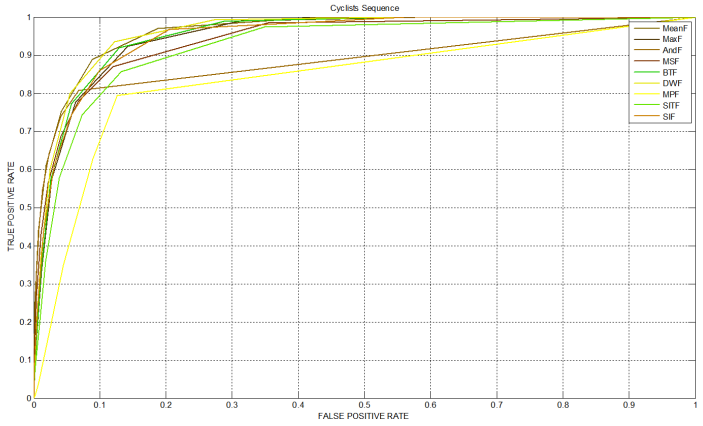


FIGURE 4.7 – Quantitative fusion comparision with *Cyclists* sequence.

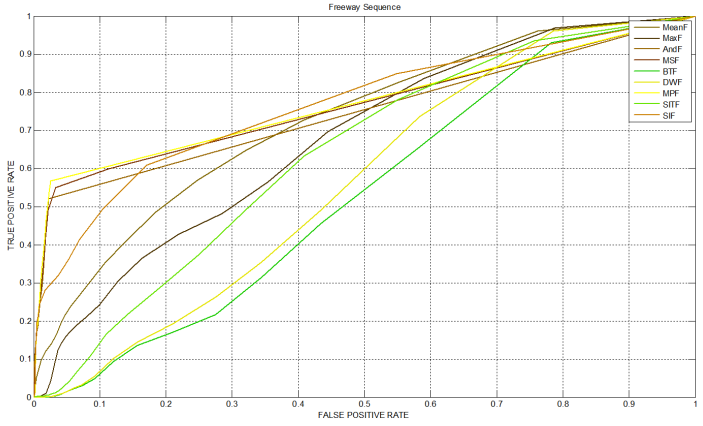


FIGURE 4.8 – Quantitative fusion comparision with *Freeway* sequence.

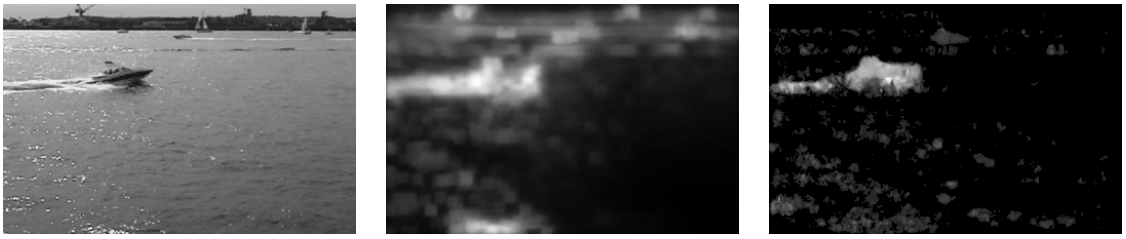


FIGURE 4.9 – Examples of static and dynamic saliency detection on good color contrast and motion contrast boats sequence. From left to right : *Original frame*, *Static saliency* and *Dynamic saliency*.

4.5/ SPATIO-TEMPORAL SALIENCY DETECTION WITH LBP FEATURES

In this section, we evaluate the performance of the proposed spatio-temporal saliency detection method using LBP features described in Chapter 3. We used the SVCL dataset for evaluating the performance of our method in detecting salient objects in dynamic scenes, and also for comparison with other techniques. We compare our spatio-temporal saliency detection method combining color features and LBP-TOP features (PROPOSED), with the method using LBP features only (LBP-TOP) and three state-of-art methods : a method using optical flow to compute motion features (OF) [52], the self-resemblance method (SR) [64] and the phased discrepancy based saliency detection method (PD) [77]. For the last three methods, we use codes provided by the authors. For LBP-TOP based saliency, we use center-surround mechanism described in Section 3.3.3 with a center region of size 17×17 and a surround region of size 97×97, and we extract LBP features from a temporal volume of six frames.

We evaluate the different spatio-temporal saliency detection methods by generating Receiver Operating Characteristic (ROC) curves and evaluating the Area Under ROC Curve (AUC).

The post-processing step described in Section 3.3.4 is important in order to obtain good final saliency maps. It basically lower the final saliency value of pixels far away from all pixels with saliency value above a defined threshold. The parameter λ in Eq. (**??**) controls the importance of the attenuation. The effect of this parameter can be observed in Fig. 4.10 for three sequences. As can be seen, optimal values are between 0.2 and 0.6 for the three sequences. We have selected the value $\lambda = 0.5$ as it is, in average, the best value for all tested sequences.

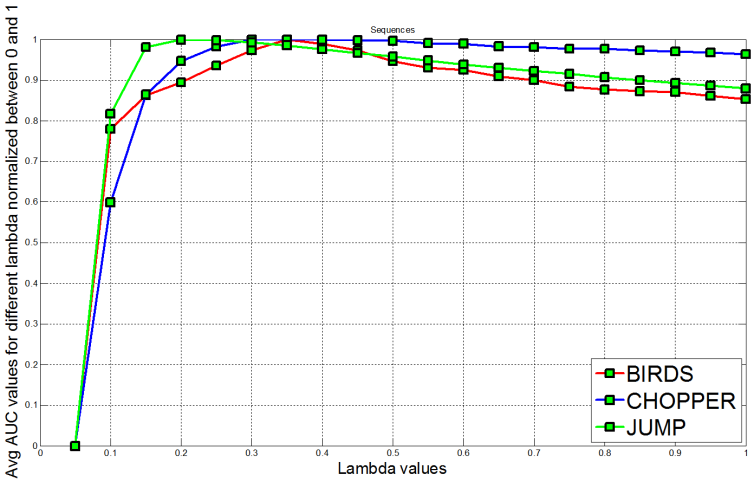


FIGURE 4.10 – Influence of λ on the performance of the proposed method.

Table 4.5 summarizes the results obtained with all sequences by the different saliency detection methods whereas the visual comparison of these methods are shown in Fig. 4.22. We can observe that the proposed method achieves the best overall performance with an average AUC value of 0.914 for all twelve sequences. The optical flow based method (OF) achieves an average AUC value of 0.907, whereas as self-resemblance (SR), phase

Sequence	PROPOSED	LBP-TOP	OF [52]	SR [64]	PD [77]	Avg AUC
Birds	0.9586	0.7680	0.9664	0.9379	0.8221	0.8906
Boats	0.9794	0.8358	0.9827	0.9227	0.9765	**0.9394**
Bottle	0.9953	0.9413	0.8787	0.9961	0.8285	0.9279
Cyclists	0.9317	0.6737	0.9602	0.8682	0.9551	0.8777
Chopper	0.9717	0.9427	0.9850	0.7447	0.6470	0.8582
Freeway	0.7775	0.8684	0.5456	0.7760	0.7318	**0.7398**
Peds	0.9552	0.7376	0.9512	0.8603	0.8548	0.8718
Ocean	0.9271	0.8513	0.7810	0.8016	0.8235	0.8369
Surfers	0.9674	0.7489	0.9545	0.9455	0.9352	0.9103
Skiing	0.8389	0.3787	0.9796	0.8872	0.9367	0.8042
Jump	0.8957	0.6960	0.9481	0.8321	0.6616	0.8067
Traffic	0.7693	0.6088	0.9615	0.5491	0.8720	0.7521
Avg AUC	**0.9140**	**0.7453**	**0.9079**	**0.8434**	**0.8371**	

TABLE 4.5 – Evaluation of spatio-temporal saliency detection methods. PROPOSED (with color and LBP features), LBP-TOP (LBP features only), OF (Optical Flow based), SR (Self-Resemblance) and PD (Phase Discrepancy).

discrepancy (PD) and LBO-TOP achieve lower average AUC values, respectively 0.843, 0.837 and 0.745. These results confirm the observation from Section 3.3.4 that the combination of color features with LBP features produces better saliency map. In fact, the proposed method fusing color and LBP features gives an average AUC value which is 22% higher that the value wih LBP-TOP features.

Analyzing the sequences individually, we see that the best and least performances are obtained with the *Boats* and *Freeway* sequences, respectively, with average AUC values of 0.9394 and 0.7398 for all five saliency detection methods. The *Boats* sequence shows good color and motion contrasts, so both static and dynamic maps are estimated correctly, and all spatio-temporal saliency detection methods perform well. Note however that the LBP-TOP based method gives slightly lower accuracy than other techniques. On the other hand, the color contrast of the *Freeway* sequence is very limited. So getting a correct static saliency map is difficult with this sequence whereas the quality of the final spatio-temporal saliency map relies on the dynamic map. The best performing method with this sequence is the LBP-TOP based technique with an average AUC value of 0.868, while optical-flow based technique achieves an average AUC value of only 0.545. This example illustrates that using LBP features to represent dynamic textures (and to compute the dynamic saliency map) gives very good results. The ROC curves comparing performances of the different methods on three sequences are shown in Fig. 4.12,4.13 and 4.14.

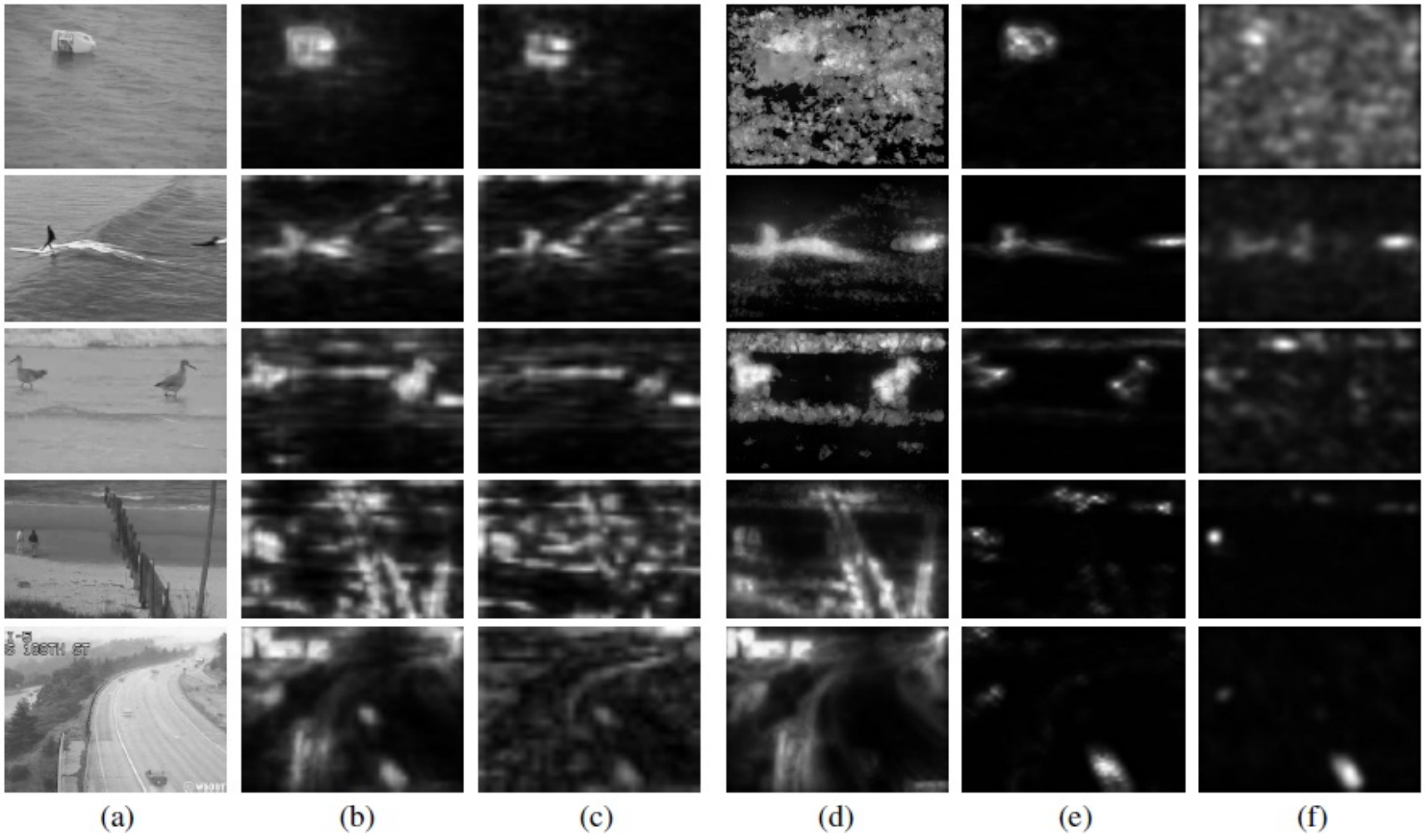


FIGURE 4.11 – Visual comparison of spatio-temporal saliency detection of our methods and state of art methods. (a) Original frame ; (b) PROPOSED ; (c) LBP-TOP ; (d) OF [52] ; (e) SR [64] and (f) PD [77]

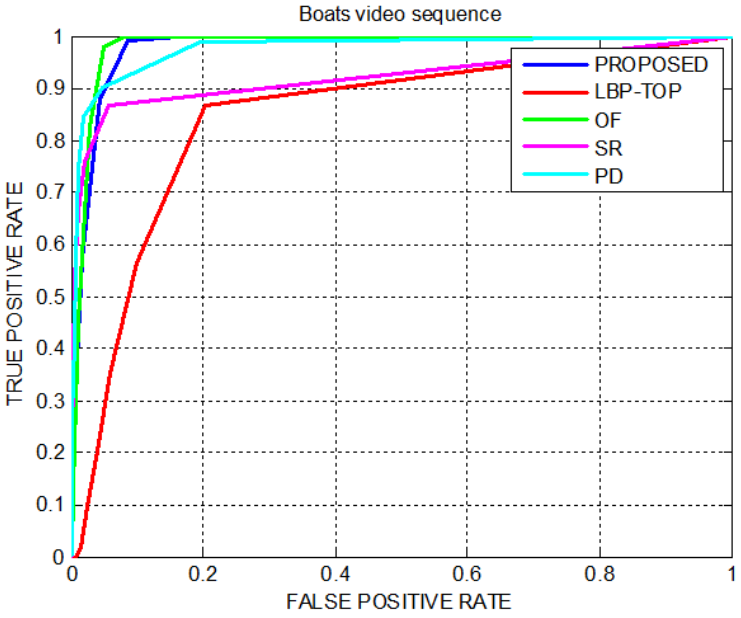


FIGURE 4.12 – Quantitative comparision with *Boats* sequence.

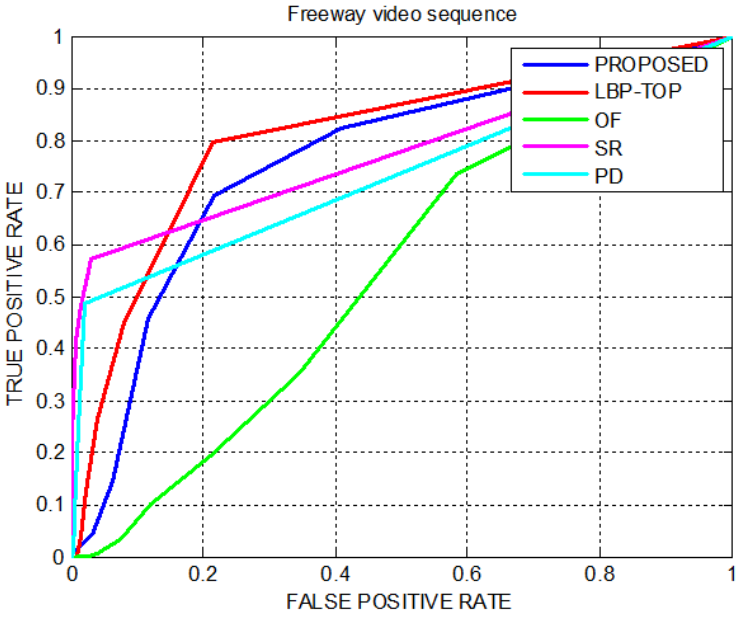


FIGURE 4.13 – Quantitative comparision with *Freeway* sequence.

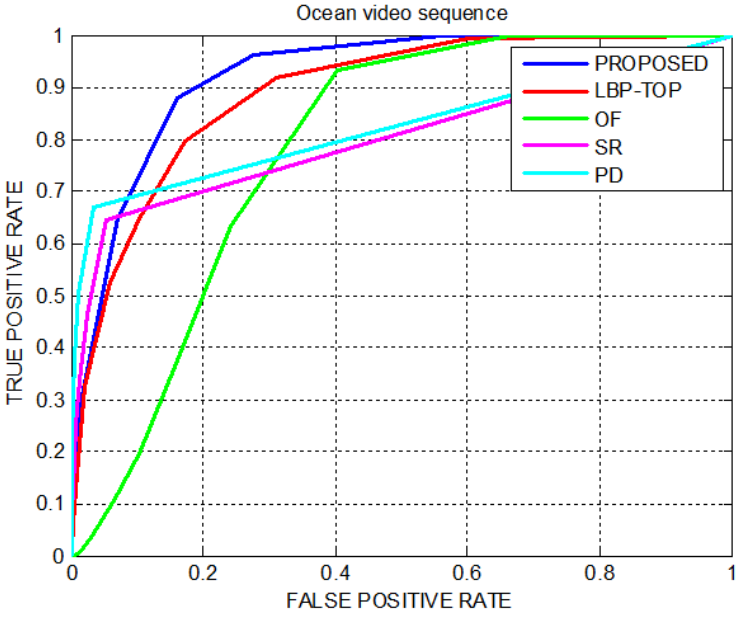


FIGURE 4.14 – Quantitative comparision with *Ocean* sequence.

4.6/ Comparison with eye-tracking data

In this section we evaluate the performance of the proposed method in predicting human fixations. We use the ASCMN dataset [63], which contains 24 videos divided into five classes : ABNORMAL,SURVEILLANCE,CROWD, MOVING, NOISE. In the ABNORMAL class, videos contain objects whose motion can be seen as abnormal compared with other objects (different speed or orientation), and thus attract an observer's attention. The SURVEILLANCE class contains videos which are taken from fixed surveillance camera. In the CROWD class contains videos with more or less dense crowd. The MOVING class contains videos taken with a moving camera, and finally, videos in the NOISE class show sudden appearances of salient objects. The dataset also comes with human fixations data obtain using an eye tracking equipment. This eye tracking data serve as ground truth for a quantitative analysis of methods performance.

We compare our proposed spatio-temporal saliency detection method combining color features and LBP-TOP features (PROPOSED), with the method using LBP features only (LBP-TOP) and four other state of the art methods which are the incremental coding length (ICL) method [31], the method based on natural images statistics (SUN) [74], the self-ressemblence method (SR) [64], and the method of Mancas et al. [50].

The post processing step described in Section 3.3.4 is important to obtain a good saliency maps. It basically attenuates the lower saliency values of pixels which are far away from all pixels with saliency values above a define threshold. The parameter λ in Eq. (3.37) controls the importance of the attenuation. The effect of this parameter can be observed in Fig. 4.10 for three sequences. As can be seen, optimal values are between 0.2 and 0.6 for the three sequences. We have selected the value $\lambda = 0.2$ as it is the best value for all tested sequences which are used for eye tracking data experiments.

To have a fair relative comparison of the different dynamic saliency models, we propose the use of a novel saliency map preprocessing methodology. The initial saliency maps of the six dynamic models can be seen in the left column of Figure 4.15. We, therefore, performed mathematical morphology operations can be observed in same figure middle column and Gaussian smoothing on the saliency maps to make them all have the same object size. The results are visible in the third column.

The morphological operations and smoothing parameters were chosen in order to maximize each model's results in terms of NSS, KL and AUC metrics. Those parameters ensure both the optimal evaluation for each model and their relative fairness. In addition, as substantial shifts between the tracked target and the real gaze position are observed, using the precise fixation position has much less meaning than in the case of still images. Thus, we dilated the fixation points to larger disks before producing the heat maps.

4.6.1/ Comparison based on evaluation metrics

Table 4.6, Table 4.7 and Table 4.8 summarize the results obtained by the different saliency detection methods for all the twenty four video sequences of the dataset, using AUC, NSS and KL metrics respectively. First of all, we can see that the relative performances of the different methods depends on the evaluation metric used. This justify our idea of using more than one metric to ensure that the discussion about the results is as independent as possible from the choice of the metrics.

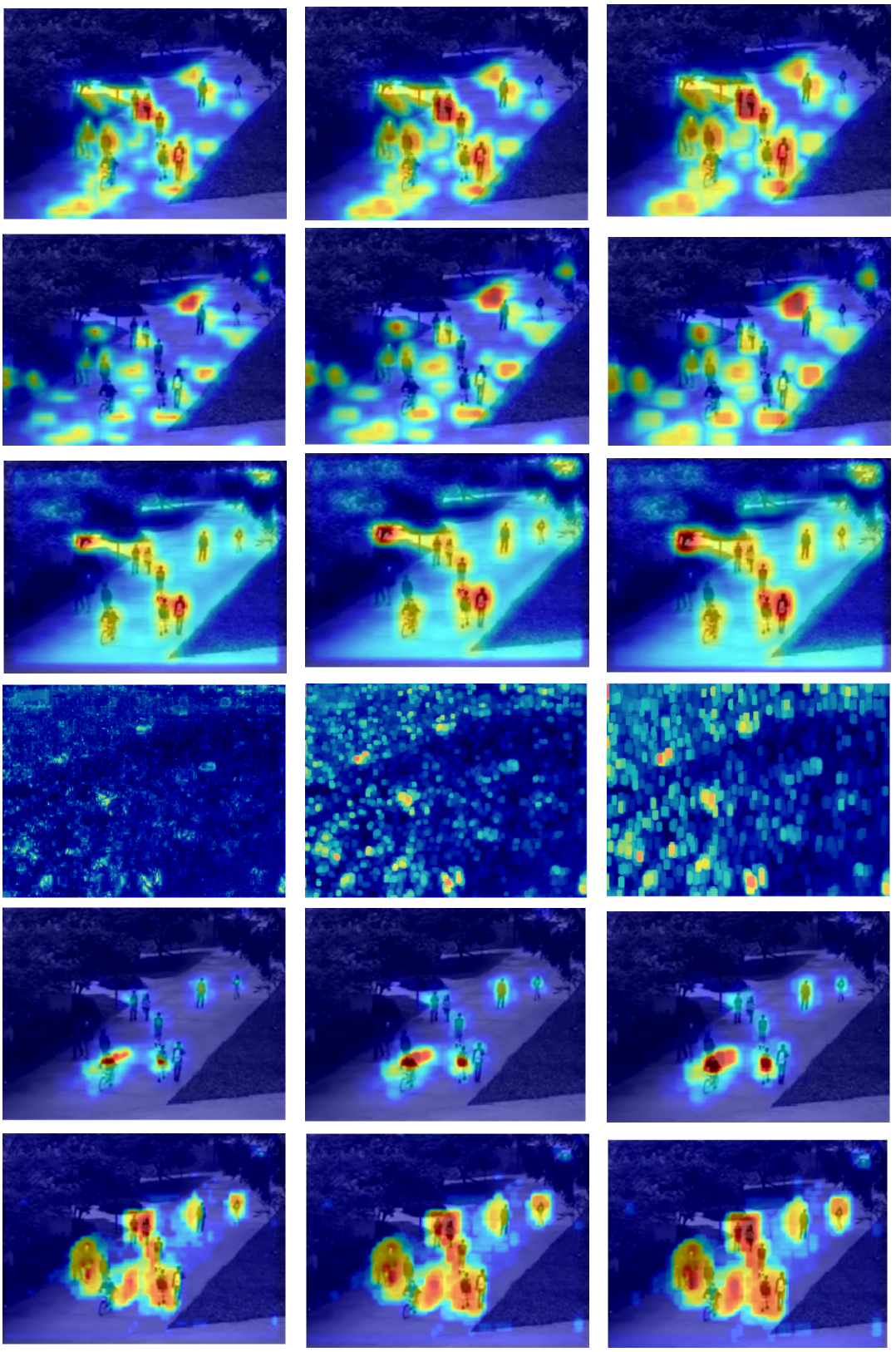


FIGURE 4.15 – Preprocessing of saliency maps using morphology and Gaussian filtering. From top to bottom, saliency maps obtained with PROPOSED, LBP-TOP, ICL, SUN, SR and MANCAS methods.

In terms on evaluation metrics, for AUC and NSS the higher the value the better is the performance of a method. On the contrary, for the KL measure, the lower the value the better the performance of a method. Table 4.6 show that the proposed method achieves an average AUC value of 0.64, which is higher that the performance of ICL, LBP-TOP and SUN methods which achieve average AUC values of 0.63, 0.53 and 0.62 respectively. However, the PROPOSED method using LBP features combined with color information has a lower performance compared to MANCAS and SUN method which achieve average AUC value of 0.68 and 0.66 respectively.

Regarding the performances in term of NSS metric as shown in Table 4.7, we can see that the method using only LBP features (LBP-TOP) achieves very poor results, whereas the PROPOSED method gives satisfactory results. The best method for this metric is MANCAS method, followed by SR and SUN.

Finally, when using KL metric, Table 4.8, the PROPOSED method achieves the second best result, being outperformed only by SR. However, we can also see that all saliency methods gives comparable results in terms of KL measure.

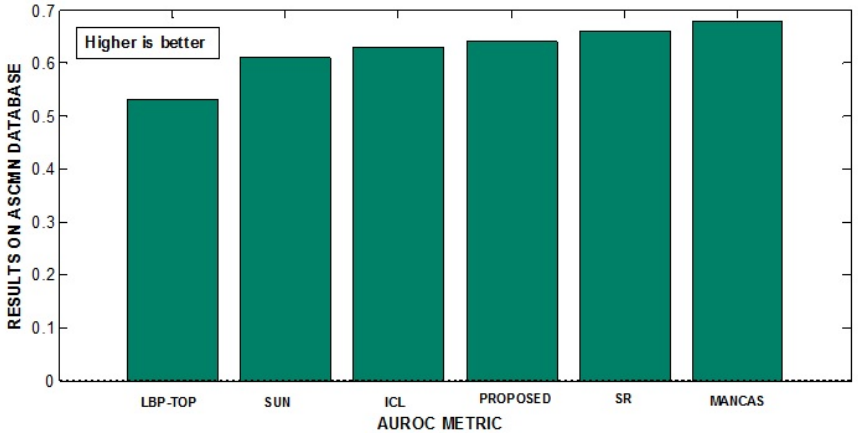


FIGURE 4.16 – Results on the ASCMN dataset using AUC metric.

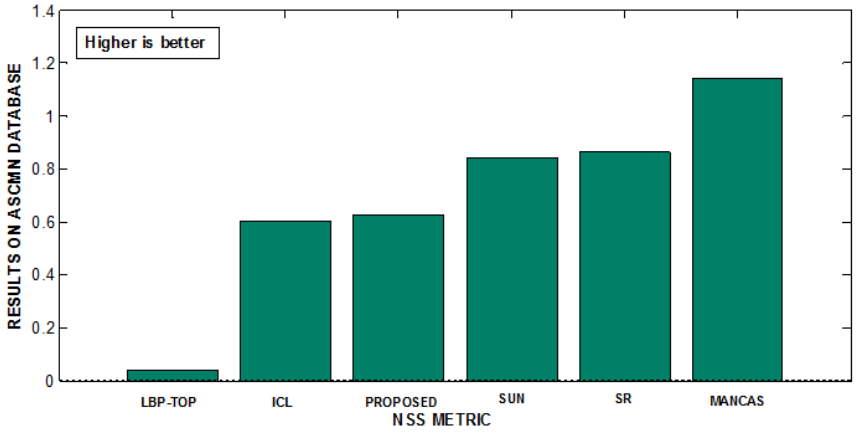


FIGURE 4.17 – Results on the ASCMN dataset using NSS metric.

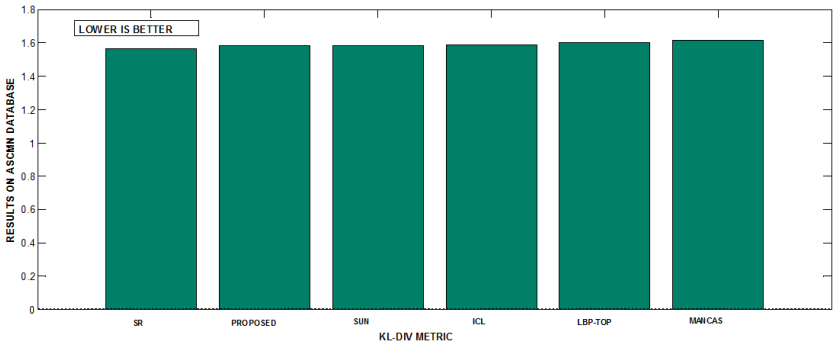


FIGURE 4.18 – Results on the ASCMN dataset using KL metric.

4.6.2/ COMPARISON BASED OF VIDEO CLASSES

Table 4.11 and Table 4.9 summarize the results obtained by the different methods for different videos classes. From this table, we can see that the PROPOSED method performs best for the MOVING class of videos and gives comparable results with SR and MANCAS for other classes in terms of AUC measure. It is worth mentioning that for all five classes of videos, using LBP features alone produces worst results. The most difficult class seems to be the CROWD class. For this class of videos, the PROPOSED method gives better performance than LBPTOP and ICL, but lower performance than SR, SUN and MANCAS.

Figure 4.19, Figure 4.20 and Figure 4.21 show the visualization of the results presented in Table 4.11 . The visual comparison of the evaluated methods are shown in Figure 4.22.

4.7/ COMPUTATION TIME DETAILS

The proposed algorithm is computed in three parts. The first part is computing the static saliency map and the second one is computing the dynamic (temporal) saliency map and finally they are fused. The proposed algorithm is implemented using MATLAB on a PC with Intel processor of 2.20 GHZ and with RAM of 8.GB. The computation time for each frame of size 640×480 for the static map takes in between 80-100 sec. Whereas for the dynamic saliency map it takes even more of about 300-400 sec.These maps are fused and for the fusion it takes approximately 1 sec to get. The details are summarized in Table 4.10

Sequence	PROPOSED	LBPTOP	ICL	SUN	MANCAS	SR
Video1	0.6347	0.5312	0.6027	0.6252	0.6590	0.6936
Video2	0.7634	0.6286	0.7552	0.7036	0.8084	0.8131
Video3	0.6848	0.4801	0.5999	0.6318	0.6637	0.7096
Video4	0.7771	0.6851	0.7647	0.6494	0.7653	0.7933
Video5	0.7766	0.6343	0.7952	0.7580	0.7464	0.7835
Video6	0.6908	0.5599	0.6956	0.6219	0.7089	0.6994
Video7	0.6177	0.4916	0.6290	0.6196	0.6195	0.6107
Video8	0.5440	0.4310	0.5850	0.5726	0.6913	0.6165
Video9	0.5316	0.4259	0.6529	0.6225	0.6503	0.6625
Video10	0.6478	0.6461	0.5547	0.6786	0.7294	0.6482
Video11	0.5996	0.6524	0.6253	0.5950	0.5981	0.6090
Video12	0.5866	0.3842	0.5075	0.6363	0.6379	0.6087
Video13	0.5783	0.5031	0.5757	0.5863	0.6522	0.6119
Video14	0.3259	0.3434	0.4373	0.4714	0.7109	0.5432
Video15	0.5580	0.4419	0.6170	0.6425	0.6864	0.6418
Video16	0.7141	0.5208	0.6840	0.6205	0.7658	0.7138
Video17	0.4259	0.2948	0.6183	0.7237	0.7445	0.6204
Video18	0.6971	0.5265	0.7142	0.6680	0.6920	0.7084
Video19	0.7333	0.5988	0.6769	0.6105	0.6878	0.6879
Video20	0.7428	0.5611	0.7462	0.6702	0.7286	0.7386
Video21	0.6294	0.4647	0.5950	0.5962	0.6566	0.6132
Video22	0.6530	0.6362	0.5886	0.4619	0.4207	0.5491
Video23	0.6768	0.6643	0.5579	0.5292	0.6542	0.5966
Video24	0.7010	0.5517	0.5722	0.5607	0.7181	0.6134
Avg val	**0.6371**	**0.5274**	**0.6313**	**0.6190**	**0.6828**	**0.6628**

TABLE 4.6 – Evaluation of spatio-temporal saliency detection methods using Arear Under Curve (AUC). PROPOSED (with color and LBP features), LBP-TOP (LBP features only), ICL [31], SUN [74], MANCAS [50] and SR [64].

Sequence	PROPOSED	LBPTOP	ICL	SUN	MANCAS	SR
Video1	0.4942	-0.0646	0.3143	0.7559	0.8754	1.0435
Video2	1.0668	0.0431	1.200	1.7176	2.3333	1.8060
Video3	1.0745	-0.4418	0.4324	1.4967	0.7524	1.1226
Video4	0.9921	0.2972	1.3432	1.1940	1.5964	1.7366
Video5	0.9629	0.1528	0.8652	1.1381	1.3817	0.9927
Video6	0.7005	0.0885	1.5841	0.7207	2.0196	1.2076
Video7	0.6184	-0.0042	0.5314	0.8845	0.4069	0.5231
Video8	0.1620	-0.1377	-0.0110	0.5028	1.0538	0.2182
Video9	0.0394	-0.3933	0.3405	0.8875	0.9314	0.6294
Video10	0.8349	0.7220	0.1400	0.8189	1.2022	0.8298
Video11	0.2554	0.2754	0.6319	0.8596	0.8828	0.5367
Video12	0.0189	-0.5390	-0.1877	0.6176	0.5423	0.1488
Video13	0.4770	0.0212	0.3736	0.8650	1.1854	0.7080
Video14	-0.4474	-0.3224	-0.3999	0.0226	0.9408	0.0238
Video15	0.3262	-0.1713	0.5936	0.8256	0.7947	0.8182
Video16	1.1478	0.0594	1.3420	1.0892	1.6465	1.1229
Video17	-0.3396	-0.6171	0.0537	1.2417	1.2571	0.8113
Video18	1.5664	0.9084	1.5604	1.4209	1.6057	1.5960
Video19	0.8340	0.0675	0.8990	0.5671	0.8774	1.1475
Video20	1.5877	-0.0668	1.9037	1.5841	1.8044	1.5533
Video21	0.5808	-0.0305	0.3283	0.5249	0.8920	0.4147
Video22	0.5258	0.4348	0.4051	-0.1429	-0.5654	0.1410
Video23	0.6184	0.6049	0.2948	0.2647	1.0349	0.5313
Video24	0.9234	0.1121	-0.0202	0.3748	1.5866	0.2655
Avg val	**0.6258**	**0.0416**	**0.6049**	**0.8430**	**1.1420**	**0.8623**

TABLE 4.7 – Evaluation of spatio-temporal saliency detection methods using Normalized Scanpath Saliency (NSS). PROPOSED (with color and LBP features), LBP-TOP (LBP features only), ICL [31], SUN [74], MANCAS [50] and SR [64].

Sequence	PROPOSED	LBPTOP	ICL	SUN	MANCAS	SR
Video1	1.6491	1.6555	1.6509	1.6452	1.6521	1.6199
Video2	1.5209	1.5605	1.5370	1.5081	1.5647	1.4536
Video3	1.6508	1.6899	1.6642	1.6431	1.6720	1.6262
Video4	1.6144	1.6277	1.6189	1.6140	1.6653	1.5516
Video5	1.6193	1.6456	1.6184	1.6188	1.6512	1.5914
Video6	1.4986	1.5619	1.4976	1.5146	1.6147	1.4785
Video7	1.6087	1.6294	1.6146	1.6095	1.6360	1.5917
Video8	1.6357	1.6464	1.6419	1.6301	1.6241	1.6332
Video9	1.6407	1.6564	1.6302	1.6194	1.6427	1.6095
Video10	1.4822	1.4761	1.5065	1.4987	1.4994	1.4866
Video11	1.5304	1.5275	1.5189	1.5173	1.5382	1.5083
Video12	1.6516	1.6768	1.6582	1.6439	1.6362	1.6473
Video13	1.5260	1.5387	1.5337	1.5250	1.5410	1.5143
Video14	1.5163	1.5114	1.5060	1.4965	1.5429	1.4968
Video15	1.5714	1.5914	1.5671	1.5671	1.5823	1.5432
Video16	1.4929	1.5528	1.4895	1.4961	1.5226	1.4764
Video17	1.5909	1.6002	1.5791	1.5556	1.5761	1.5474
Video18	1.6494	1.7016	1.662	1.6764	1.7471	1.6271
Video19	1.6207	1.6482	1.6191	1.6367	1.6418	1.6027
Video20	1.6803	1.7427	1.6792	1.6953	1.7280	1.6631
Video21	1.5862	1.6033	1.5948	1.5918	1.5974	1.5870
Video22	1.5529	1.4941	1.5520	1.5760	1.6635	1.5670
Video23	1.5105	1.5158	1.5259	1.5268	1.5426	1.5014
Video24	1.6633	1.6879	1.6927	1.6849	1.6795	1.6836
Avg val	**1.5860**	**1.6059**	**1.5899**	**1.5871**	**1.6158**	**1.5662**

TABLE 4.8 – Evaluation of spatio-temporal saliency detection methods using Kullback Leibler-Divergence (KL). PROPOSED (with color and LBP features), LBP-TOP (LBP features only), ICL [31], SUN [74], MANCAS [50] and SR [64].

MODELS	mean AUC	mean NSS	mean KL
PROPOSED	64%	0.6214	1.5860
LBP-TOP	53%	0.0416	1.6059
ICL	63%	0.6049	1.5899
SUN	61%	0.8430	1.587
MANCAS	68%	1.1420	1.6158
SR	66%	0.8623	1.5662

TABLE 4.9 – Results of whole dataset of videos with three evaluation metrics

saliency map types	PC details	Time in sec
Static	2.20GHz, 8 GB RAM	80-100sec
Dynamic	2.20GHz, 8 GB RAM	300-400sec
Fusion	2.20GHz,8 GB RAM	1 sec

TABLE 4.10 – Computation time details metrics

Eyetracking results on each classes			
Videoclasses	mean AUC	mean NSS	mean KL
ABNORMAL	PROPOSED : 74%	PROPOSED : 1.2722	PROPOSED : 1.5916
	LBPTOP :58%	LBPTOP :0.2483	LBPTOP :1.6371
	ICL :73%	ICL :1.4699	ICL :1.5973
	SUN :66%	SUN :1.4012	SUN :1.5980
	SR :75%	SR :1.5630	SR :1.5544
	MANCAS :75%	MANCAS :1.7973	MANCAS :1.6455
SURVEILLANCE	PROPOSED : 66%	PROPOSED : 0.6428	PROPOSED : 1.6400
	LBPTOP :52%	LBPTOP :0.1867	LBPTOP :1.6618
	ICL :67%	ICL :0.5154	ICL :1.6444
	SUN :66%	SUN :1.2111	SUN :1.6326
	SR :71%	SR :0.9471	SR :1.6118
	MANCAS :68%	MANCAS :0.9852	MANCAS :1.6545
CROWD	PROPOSED : 55%	PROPOSED : 0.2298	PROPOSED : 1.5744
	LBPTOP :45%	LBPTOP :−0.0615	LBPTOP :1.5820
	ICL :53%	ICL :−0.0261	ICL :0.4974
	SUN :59%	SUN :0.4974	SUN :1.5722
	SR :60%	SR :0.3271	SR :1.5702
	MANCAS :68%	MANCAS :0.9262	MANCAS :1.5800
MOVING	PROPOSED : 69%	PROPOSED : 0.7459	PROPOSED : 1.5839
	LBPTOP :59%	LBPTOP :0.1757	LBPTOP :1.5980
	ICL :63%	ICL :0.7170	ICL :1.5903
	SUN :56%	SUN :0.3799	SUN :1.6030
	SR :64%	SR :0.6904	SR :1.5830
	MANCAS :63%	MANCAS :0.9262	MANCAS :1.5800
NOISE	PROPOSED : 57%	PROPOSED : 0.3260	PROPOSED : 1.5563
	LBPTOP :50%	LBPTOP :0.0181	LBPTOP :1.5672
	ICL :60%	ICL :0.4132	ICL :1.5565
	SUN :61%	SUN :0.8235	SUN :1.5502
	SR :61%	SR :0.6548	SR :1.5344
	MANCAS :66%	MANCAS :0.9270	MANCAS :1.5694

TABLE 4.11 – Results for the 5 classes of videos with the three evaluation metrics.

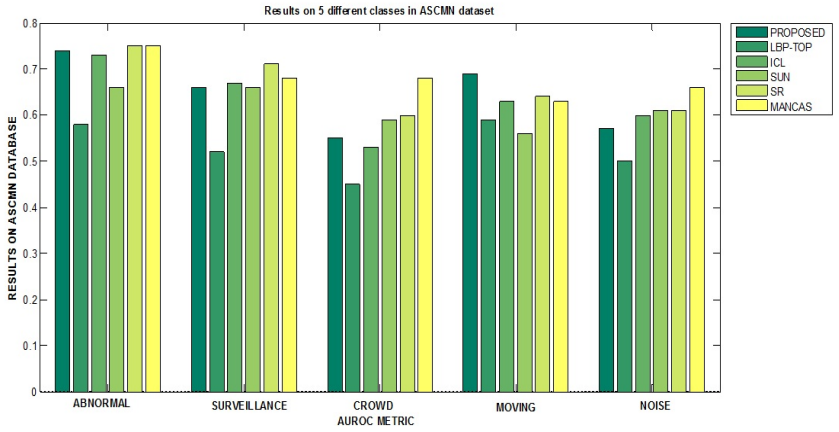


FIGURE 4.19 – Results for different classes using AUC metric

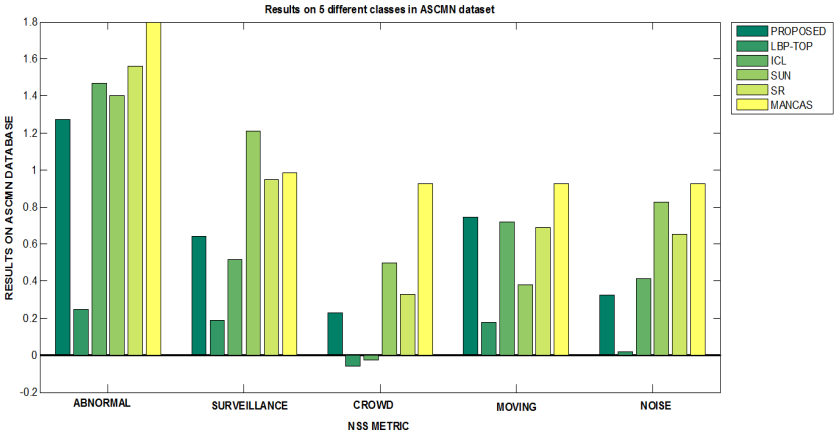


FIGURE 4.20 – Results for different classes using NSS metric

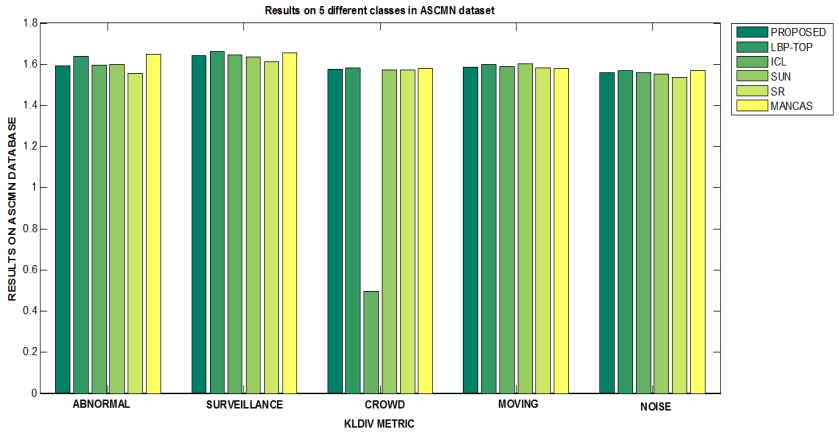


FIGURE 4.21 – Results for different classes using KL metric

FIGURE 4.22 – Visual comparison of spatio-temporal saliency detection of our methods and state of art methods on ASCMN dataset. (a) Original frame ; (b) PROPOSED ; (c) LBP-TOP ; (d) ICL [31] ; (e) SUN [74] ; (f) MANCAS [50] and (g) SR [64]

5

Conclusion and Future perspectives

This chapter presents a summary of the thesis and conclusions drawn on it based on experimental evaluations. It also provides some future perspectives that need to be carried out in order to overcome the limitations of the proposed solution.

5.1/ Conclusion of the thesis

Saliency detection is an active domain of research that has received a tremendous amount of attention from the computer vision community and most of the existing techniques are limited to static images. However, videos have different characteristics than images and the perception of videos is different from images due to its additional temporal information. Furthermore, a video contains strong spatial-temporal correlations between the regions of consecutive frames. The main objective of this thesis was to propose a new spatio-temporal saliency detection algorithm for dynamic scenes and to overcome the challenges identified in complex dynamic scenes.

In chapter 2 we reviewed several important state-of-the-art-methods of spatio-temporal saliency detection and these methods are classified into two categories : biological inspired and computational methods. We observed that these methods suffer from several limitations and are not able to satisfy all the challenges of natural video sequences. Let us stress out that natural video scenes have many dynamic entities such as moving water, snow, smoke etc. Along with these dynamic entities, if the video has additional camera motion it becomes very complex. The perception of videos is different than for static images due to its additional temporal information. So, dealing with temporal information is important for saliency detection in videos. The major conclusion drawn from this chapter is that the state-of-the-art methods fail to address the motion information properly. Moreover most of these methods are restricted to detect the salient regions only and use either Kullback-Leibler divergence (KL) or Normalized Scanpath Saliency (NSS) or Receiver Operating Curve (ROC) evaluation metric for their evaluation. A fair comparison should be done in order to compare the methods with common evaluation metrics. This led us to propose a new approach for saliency detection in dynamic scenes.

In chapter 3, first we proposed to use an optical flow based motion estimation method to compute the dynamic saliency map and to use color features to compute the static

saliency map. But the optical flow estimation method fails in complex scenes with low motion contrast. So we have proposed another spatio-temporal saliency detection method for dynamic scenes based on local binary patterns. The method combines in a first step color features for static saliency detection, and dynamic textures for temporal dynamic saliency detection. The obtained two saliency maps are in a second step fused into a spatio-temporal map. Since the fusion step is an important component in bottom-up saliency modeling, in this chapter we also reviewed some of the fusion strategies used in literature. The experimental evaluation of the proposed methods are carried in chapter 4.

In chapter 4 we performed three different experimental evaluation. First, a performance evaluation of different fusion techniques for spatio-temporal saliency detection in dynamic scenes was presented. Nine fusion techniques were evaluated on the SVCL dataset of twelve complex dynamic scenes. The evaluation was done using Receiver Operating Curve (ROC) and the results showed the consistency of fusion approaches that base their decision on the scenes characteristics as the final spatio-temporal saliency map takes the best of each of the individual saliency map (static and dynamic). This include Mean, Scale Invariant, Max and Dynamic Weights fusion methods. On the other hand, fusion techniques which are based on a strong a priori such as Motion Priority fusion achieve good results only when the underlying assumption is satisfied. Thus, they performances vary depending on the sequence. It is clear that the accuracy of a spatio-temporal saliency map depends on the quality of both static and dynamic maps, which are based on the scenes contents.

In a second experiment we evaluated the proposed spatio-temporal saliency, which is based on local binary patterns, using the SVCL dataset. The experimental results show that our method performs significantly better than using LBP features alone, and also better than other state-of-art methods. The proposed method is in particular, able to deal with complex dynamic scenes showing difficult background textures.

Finally we did an extensive experimental evaluation for the proposed method on the large ASCMN dataset which also contains eye-tracking data. This dataset contains 24 videos of different classes (ABNORMAL, SURVEILLANCE, CROWD, MOVING and NOISE) and covers real-life scenarios. We compared our proposed method against five other state-of-the art methods. The comparison was performed both separately for each of the 5 video classes, and globally for the whole database. Based on those comparisons several conclusions can be drawn. The proposed method perform better that three out of five methods using AUC metric (Area Under the ROC Curve) and with NSS (Normalized Scanpath Saliency) the proposed method shows satisfactory results. Whereas, with KL (Kullback-Leibler divergence) metric the proposed method achieves the second best results.

Overall, these experimental results show that combination of color and texture feature plays a major role in saliency detection. The proposed spatio-temporal saliency detection method performance on the two datasets, SVCL and ASCMN, is significantly better than using LBP features alone or using optical flow based motion estimation, and is comparable to other state-of-the-art methods.

Though our method is bit expensive in computation cost. Since, we mainly focussed on dealing with challenges of complex scenes. We have not so far devoted any attention to computational optimization, we do not expect the algorithm to be deployable in real time without further investigation. We intend to address this issue in future work.

5.2/ FUTURE PERSPECTIVES

Visual attention is a very complex task especially when dealing with natural scenes. Much still need to be known about visual attention structures and functionalities in the brain. Saliency detection in static images has been widely addressed in literature and very little research has been done for spatio-temporal saliency detection. Spatio-temporal saliency detection in complex dynamic scenes is a challenging task. In this thesis we have proposed a solution for spatio-temporal saliency detection in complex scenes. Our method combines color features for spatial saliency detection, and texture features for temporal saliency detection and fused them to get a final spatio-temporal saliency map. Despite our efforts in this thesis, there are still many questions that remain to be answered.

Human visual attention is guided by other aspects of saliency apart from color, intensity, texture, shape, orientation or motion. Usually a small number of this features are used for saliency detection which results in method that are not very robust. So, it would of the interest to know which combination of features gives robust results. This therefore remains a future opportunity to improve saliency detection.

Another interesting direction would be including depth feature. According to psychological studies and neuro-science research depth feature plays an important role in human visual attention. An object which have less depth will be visually more salient than the object with larger depth value. With the availability of RGB-D sensors today, a combination of color, texture and depth features can achieve robust results in complex scenes.

A lot of what we know about vision comes from psychological studies. Usually the proposed models explain only some of the observed visual capabilities. But there are newer discoveries made in this area, so it would be better if we can update our models with these discoveries and improve saliency detection. Finally one potential interesting research direction is to develop a spatio-temporal saliency model using top-down process, which is an task dependent process. The saliency map computed using bottom-up process is due to pop-out effect. So, not everything that is salient is important to us. This bottom-up process based attention is important in unknown environments. We have a strong familiarity with most of the environments that we encounter in our daily lives. In this cases we are able to ignore salient regions which grab our involuntarily attention for a very brief moment, to be able to find the things that matter us. So top-down process based salient detection helps to widen the range to real world application in day-to-day use.

Appendices

List of publications

Publications during the thesis period

- Satya M. Muddamsetty, Désiré Sidibe, Alain Trémeau and Fabrice Mériaudeau, "Spatio-Temporal Saliency Detection in Dynamic Scenes using Local Binary Patterns" (IEEE International Conference on Pattern Recognition-2014, Sweden).
- Satya M. Muddamsetty, Désiré Sidibé, Alain Trémeau, and Fabrice Mériaudeau, "A performance Evaluation of Fusion Techniques For spatio-temporal saliency detection in Dynamic Scenes" (IEEE International Conference on Image Processing-2013, Australia).

SAMPLE OF VIDEO DATASET

In this section we provide the frames of all the videos of ASCMN dataset and SVSCL dataset.

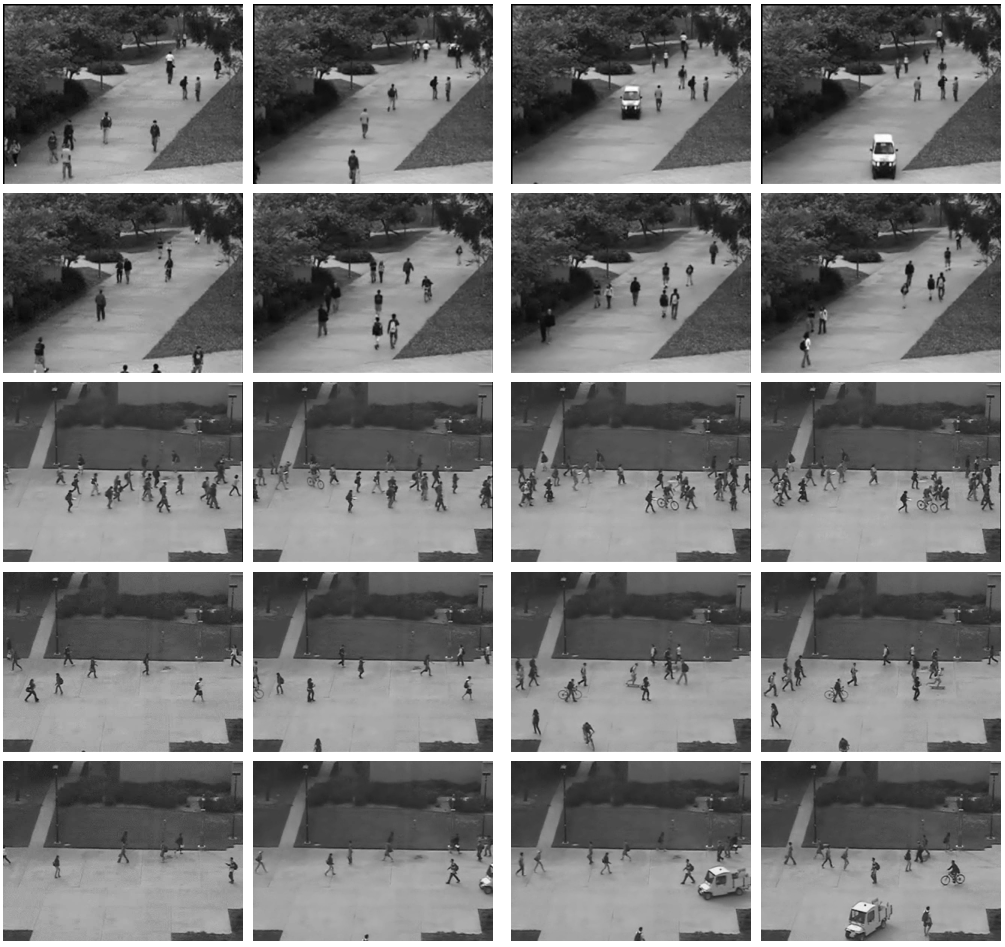


FIGURE 1 – Sample images of ABNORMAL class of ASCMN dataset is shown.

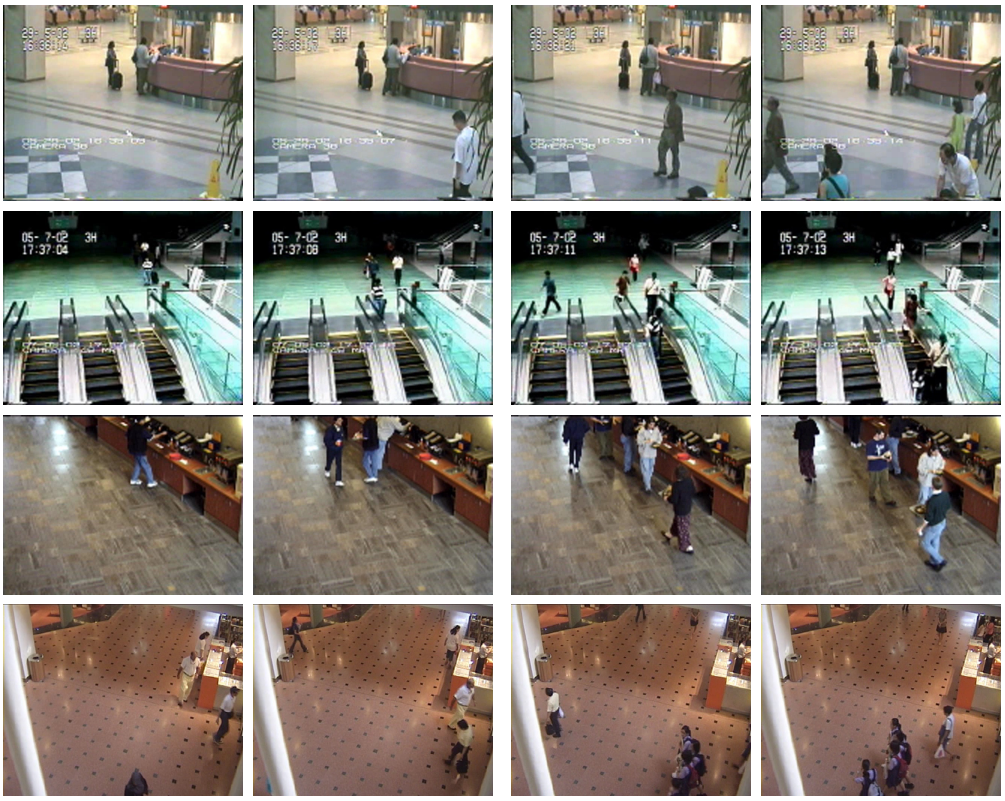


FIGURE 2 – Sample images of SURVEILLANCE class of ASCMN dataset is shown.

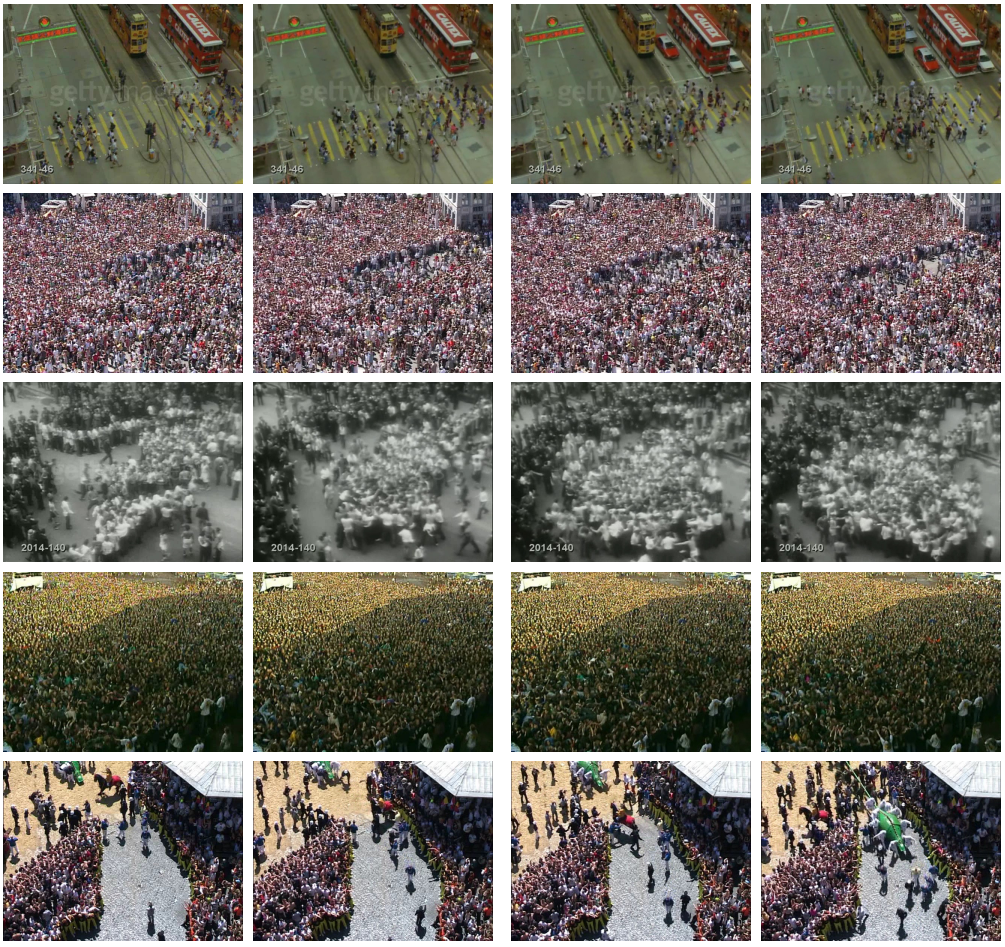


FIGURE 3 – Sample images of CROWD class of ASCMN dataset is shown.

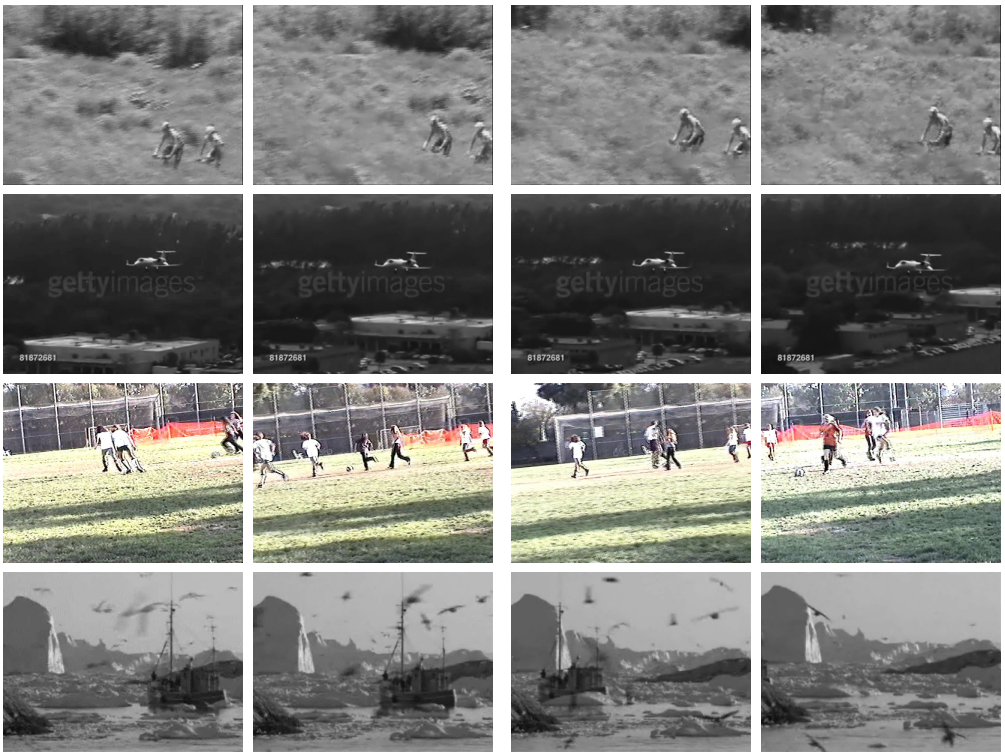


FIGURE 4 – Sample images of MOVING class of ASCMN dataset is shown.

FIGURE 5 – Sample images of NOISE class of ASCMN dataset is shown.

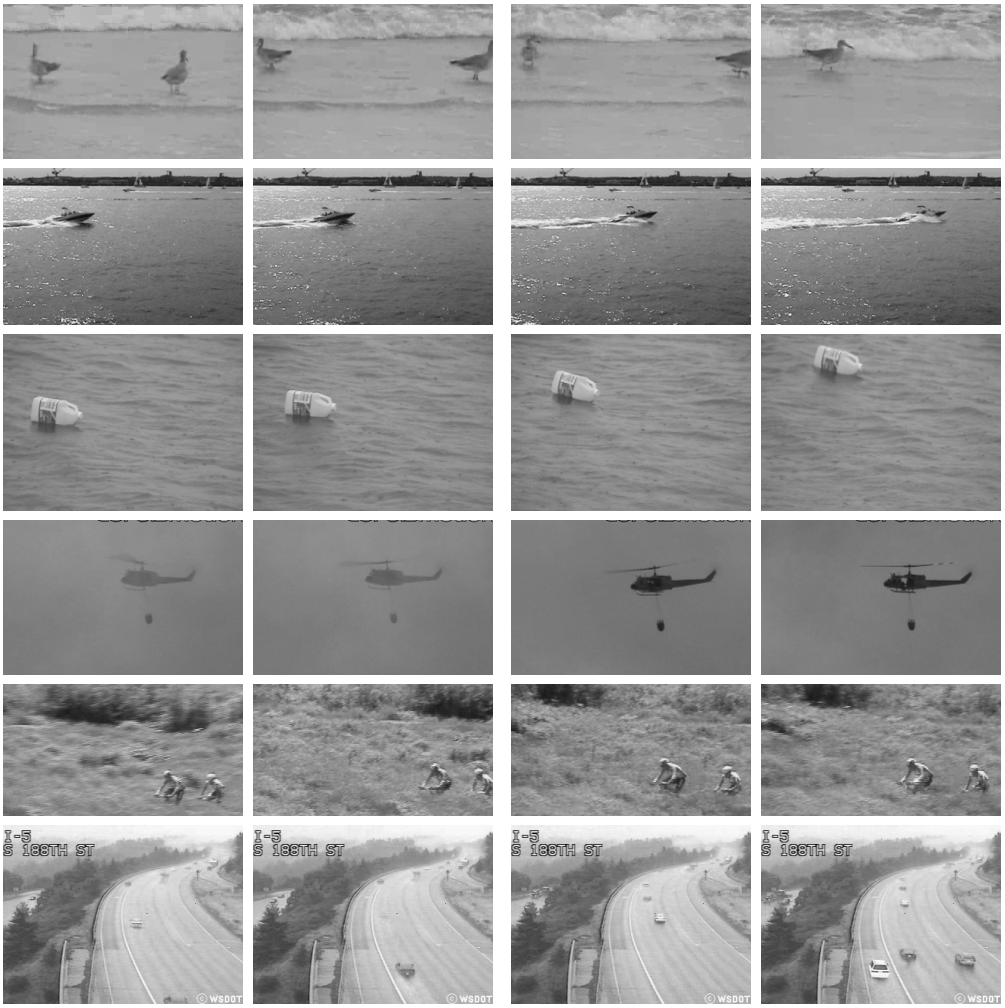


FIGURE 6 – Sample images of SVCL dynamic scenes dataset is shown.

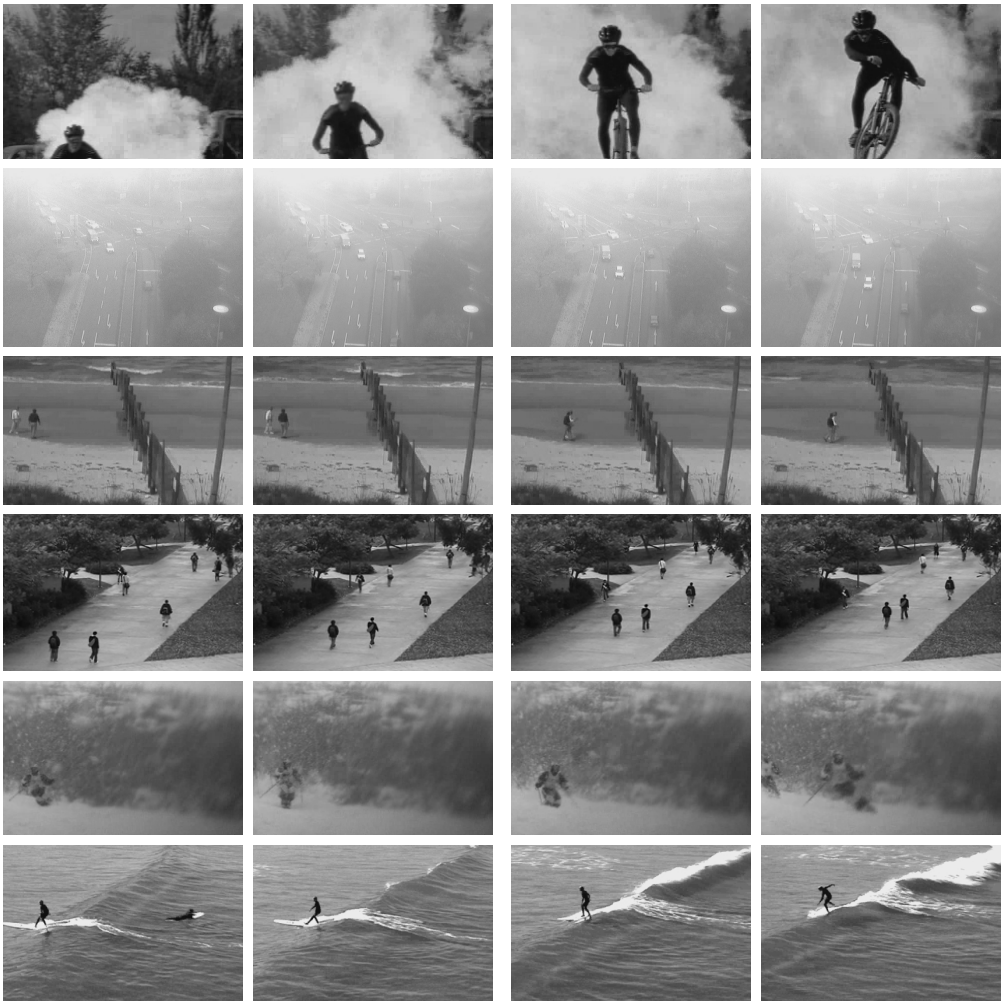


FIGURE 7 – Sample images of SVCL dynamic scenes dataset is shown.

BIBLIOGRAPHIE

[1] *Fundamentals of Neurosciences, http ://www.allpsych.uni-giessen.de/karl/teach/aka.htm*.

[2] *http ://www.techcyn.com/feature.php ?id=f2&issue=1*.

[3] R.. Achanta. *Finding objects of interest in images using saliency and superpixels*. PhD thesis, École Polytechnique Fédérale de Lausanne, 2011.

[4] R. Achanta, F. Estrada, P. Wils, and S. Süsstrunk. Salient region detection and segmentation. In *Computer Vision Systems*, pages 66–75. Springer, 2008.

[5] R. Achanta, S. Hemami, F. Estrada, and S. Süsstrunk. Frequency-tuned Salient Region Detection. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR 2009)*, pages 1597 – 1604, 2009.

[6] R. Achanta and S. Susstrunk. Saliency detection for content-aware image resizing. In *16th IEEE International Conference on Image Processing (ICIP)*, pages 1005–1008. IEEE, 2009.

[7] R. Achanta and S Susstrunk. Saliency detection using maximum symmetric surround. In *17th IEEE International Conference on Image Processing (ICIP)*, pages 2653–2656. IEEE, 2010.

[8] J. Antonio and S. Rodriguez. Attention visual search and object recognition. Technical report, Department of Computer science, York University.

[9] Statistical Visual Computing Lab at UCSD. *SVCL dataset, http ://www.svcl.ucsd.edu/projects/background*$_{s}ubtraction/, 2008(accesedOctober1st, 2012)$.

[10] S. Avidan and A. Shamir. Seam carving for content-aware image resizing. In *ACM Transactions on graphics (TOG)*, volume 26, page 10. ACM, 2007.

[11] M. Behrmann, J. J Geng, and S. Shomstein. Parietal cortex and attention. *Current opinion in neurobiology*, 14(2) :212–217, 2004.

[12] N. P. Bichot. Attention, eye movements, and neurons : Linking physiology and behavior. In *Vision and attention*, pages 209–232. Springer, 2001.

[13] A. Borji and L. Itti. State-of-the-art in visual attention modeling. *IEEE Trans on Pattern Analysis and Machine Intelligence,*, 35(1) :185–207, 2013.

[14] A. Borji, N. Dicky. Sihite, and L. Itti. Salient object detection : A benchmark. In *ECCV (2)*, pages 414–429, 2012.

[15] C. K. Chang, C. Siagian, and L. Itti. Mobile robot vision navigation & localization using gist and saliency. In *Intelligent Robots and Systems (IROS), 2010 IEEE/RSJ International Conference on*, pages 4147–4154. IEEE, 2010.

[16] L. Chang, Pong C. Yuen, and Guoping Qiu. Object motion detection using information theoretic spatio-temporal saliency. *Pattern Recogn. 2009*, 42(11) :2897–2906.

[17] M. M. Cheng, G. X. Zhang, N. J. Mitra, X. Huang, and S.M. Hu. Global contrast based salient region detection. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 409–416. IEEE, 2011.

[18] D. Chetverikov and R. Péteri. A brief survey of dynamic texture description and recognition. In *Computer Recognition Systems*, pages 17–26. Springer, 2005.

[19] G. Doretto, A. Chiuso, Y. N. Wu, and S. Soatto. Dynamic textures. *Int. J. Comput. Vision*, 51(2) :91–109, February 2003.

[20] J. Duncan and G. W. Humphreys. Visual search and stimulus similarity. *Psychological physics*, 57 :117–120, 1989.

[21] G. Farnebäck. Two-frame motion estimation based on polynomial expansion. In *Proceedings of the 13th SCIA,2003*. Springer.

[22] T. Fawcett. An introduction to roc analysis. *Pattern recognition letters*, 27(8) :861–874, 2006.

[23] S. Frintrop. *VOCUS : A visual attention system for object detection and goal-directed search*, volume 3899. Springer, 2006.

[24] S. Frintrop. *Computational Visual Attention*. Springer, 2011.

[25] S. Frintrop and P. Jensfelt. Attentional landmarks and active gaze control for visual slam. *Trans. Rob.*, 24(5) :1054–1065, October 2008.

[26] X. Gao, W. Lu, D. Tao, and X. Li. Image quality assessment and human visual system. volume 7744, pages 77440Z–77440Z–10, 2010.

[27] S. Goferman, L. Zelnik-manor, and A. Tal. Context-aware saliency detection. In *IEEE Conf. on Computer Vision and Pattern Recognition,2010*.

[28] C. L. Guo and L. M. Zhang. A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression. *IEEE TIP*, 19(1) :185 –198, 2010.

[29] B. Han and B. Zhou. High speed visual saliency computation on gpu. In *InProc of ICIP,2007*.

[30] J. Harel, C. Koch, P. Perona, et al. Graph-based visual saliency. *Advances in neural information processing systems*, 19 :545, 2007.

[31] X. Hou and L. Zhang. Dynamic visual attention : searching for coding length increments. In *NIPS*, volume 5, page 7, 2008.

[32] L. Huang and H. Pashler. A boolean map theory of visual attention. *Psychological review*, 114(3) :599, 2007.

[33] L. Itti. *Models of Bottom-Up and Top-Down Visual Attention*. PhD thesis, California Institute of Technology, 2000.

[34] L. Itti. Automatic foveation for video compression using a neurobiological model of visual attention. *Image Processing, IEEE Transactions on*, 13(10) :1304–1318, 2004.

[35] L. Itti. Quantifying the contribution of low-level saliency to human eye movements in dynamic scenes. *Visual Cognition*, 12(6) :1093–1123, 2005.

[36] L. : Crcns-orig video Itti and eye tracking database. *(http ://crcns.org/datasets/eye/eye-1)*.

[37] S. Walter K. R. Gegenfurtner and D. I. Braun. *Visual Inforamtion processing in the brain, http ://www.allpsych.uni-giessen.de/karl/teach/aka.htm.*, (acessed january, 2014).

[38] W. Kim, C. Jung, and C. Kim. Spatiotemporal saliency detection and its applications in static and dynamic scenes. *IEEE Trans. Circuits Syst. Video Techn,2011*, 21(4) :446–456.

[39] C. Koch and S. Ullman. Shifts in selection in visual attention :toward the underlying neural circuitry. *Human Neurobiology*, vol. 4, no. 4 :219–27, 1985.

[40] O. Le Meur and T. Baccino. Methods for comparing scanpaths and saliency maps : strengths and weaknesses. *Behavior research methods*, 45(1) :251–266, 2013.

[41] O. Le Meur, P. Le Callet, and D. Barba. Predicting visual fixations on video based on low-level visual features. *Vision research*, 47(19) :2483–2498, 2007.

[42] O. Le Meur, P. Le Callet, D. Barba, and D. Thoreau. A coherent computational approach to model bottom-up visual attention. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 28(5) :802–817, 2006.

[43] L. Li, W. Huang, I.H. Gu, and Q. Tian. Statistical modeling of complex backgrounds for foreground object detection. *IEEE Transactions on Image Processing*, 13(11) :1459–1472, 2004.

[44] L. Itti, C. Koch, and E. Neibur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Trans on Pattern Analysis and Machine Intelligence,1998*, 20 :1254–1259.

[45] T. Lu, Z. Yuan, Y. Huang, D. Wu, and H. Yu. Video retargeting with nonlinear spatial-temporal saliency fusion. In *ICIP,2010*.

[46] P. Anandan M. J. Black. The robust estimation of multiple motions : Parametric and piecewise-smooth flow fields. *CVIU*, Vol 63, No. 1, :74–104, 1996.

[47] V. Mahadevan and N. Vasconcelos. Saliency-based discriminant tracking. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1007–1013. IEEE, 2009.

[48] V. Mahadevan and N. Vasconcelos. Spatiotemporal saliency in dynamic scenes. *IEEE Transactions on Pattern Analysis and Machine Intelligence,*, 32(1) :171 –177, 2010.

[49] M. Mancas. Attention-based dense crowds analysis. In *Image Analysis for Multimedia Interactive Services (WIAMIS), 2010 11th International Workshop on*, pages 1–4. IEEE, 2010.

[50] M. Mancas, N. Riche, J. Leroy, and B. Gosselin. Abnormal motion selection in crowds using bottom-up saliency. In *Image Processing (ICIP), 2011 18th IEEE International Conference on*, pages 229–232. IEEE, 2011.

[51] S. Marat, T. H. Phuoc, L. Granjon, N. Guyader, D. Pellerin, and A. Guérin-Dugué. Modelling spatio-temporal saliency to predict gaze direction for short videos. *IJCV,2009*, 82(3) :231–243.

[52] S. M. Muddamsetty, D. Sidibé, A. Trémeau, and F. Mériaudeau. A performance evaluation of fusion techniques for spatio-temporal saliency detection in dynamic scenes. In *ICIP*, 2013.

[53] Society of Neuroscience. *Brain Facts : a primer on the brain and the nervous system*, 2008 (accessed April 5th, 2014).

[54] T. Ojala, M. Pietikäinen, and D. Harwood. A comparative study of texture measures with classification based on featured distributions. *Pattern recognition*, 29(1) :51–59, 1996.

[55] J. Peng and Q. Xiaolin. Keyframe-based video summary using visual attention clues. *IEEE on MultiMedia,*, 17(2) :64 –73, 2010.

[56] F. Perazzi, P. Krahenbuhl, Y. Pritch, and A. Hornung. Saliency filters : Contrast based filtering for salient region detection. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 733–740. IEEE, 2012.

[57] R. J Peters, A. Iyer, C. Koch, and L. Itti. Components of bottom-up gaze allocation in natural scenes. *Journal of Vision*, 5(8) :692–692, 2005.

[58] M. Pietikäinen, G. Zhao, A. Hadid, and T. Ahonen. *Computer Vision Using Local Binary Patterns*. Number 40 in Computational Imaging and Vision. Springer, 2011.

[59] Y. Pinto, A. R. van der Leij, I. G Sligte, Victor A. F. Lamme, and H. S. Scholte. Bottom-up and top-down attention are independent. *Journal of Vision*, 13(3) :16, 2013.

[60] M. I. Posner and S. E. Petersen. The attention system of the human brain. *Annaual Reviews of Neuroscience*, 13 :25–42, 1990.

[61] D. S. Kim R. Goebel, L. Mucklil. *The Human Nervous System (Second Edition)*, chapter Visual System, pages 1280–1305. Elsiveir, 2004.

[62] E. Rahtu, J. Kannala, M. Salo, and J. Heikkilä. Segmenting salient objects from images and videos. In *Computer Vision–ECCV 2010*, pages 366–379. Springer, 2010.

[63] N. Riche, M. Mancas, D. Culibrk, V. Crnojevic, B. Gosselin, and T. Dutoit. Dynamic saliency models and human attention : A comparative study on videos. In *Computer Vision–ACCV 2012*, pages 586–598. Springer, 2013.

[64] H. J. J. Seo and P. Milanfar. Static and space-time visual saliency detection by self-resemblance. *Journal of vision*, 9(12), 2009.

[65] D. Sidibé, D. Fofi, and F. Mériaudeau. Using visual saliency for object tracking with particle filters. In *EUSIPCO*, 2010.

[66] H. Thorsten. *A Neural Model of Early Vision :Contrast, Contours, Corners and Surfaces*. PhD thesis, University of Ulm,Faculty of Computer Science,Dept. of Neural Information Processing, 2003.

[67] K. Toyama and G. D Hager. Incremental focus of attention for robust vision-based tracking. *International Journal of Computer Vision*, 35(1) :45–63, 1999.

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

[69] S. K. Ungerleider and Leslie G. Mechanisms of visual attention in the human cortex. *Annual review of neuroscience*, 23(1) :315–341, 2000.

[70] Y.S. Wang, C. L. Tai, O. Sorkine, and T.Y. Lee. Optimized scale-and-stretch for image resizing. In *ACM Transactions on Graphics (TOG)*, volume 27, page 118. ACM, 2008.

[71] J M Wolfe. Guided search 2.0 : A revised model of visual search. *Psychonomic Bulletin & Review*, 1(2) :202–238, 1994.

[72] X. Xiao, C. Xu, and Y. Rui. Video based 3d reconstruction using spatio-temporal attention analysis. In *Multimedia and Expo (ICME), 2010*.

[73] A. Yilmaz, O. Javed, and M. Shah. Object tracking : A survey. *Acm computing surveys (CSUR)*, 38(4) :13, 2006.

[74] L. Zhang, M. H Tong, and G. W. Cottrell. Sunday : Saliency using natural statistics for dynamic analysis of scenes. *InProceedings of the 31st Annual Cognitive Science Conference*, pages 2944–2949, 2009.

[75] L. Zhang, M. H Tong, T. K. Marks, H. Shan, and G. W. Cottrell. Sun : A bayesian framework for saliency using natural statistics. *Journal of vision*, 8(7) :32, 2008.

[76] G. Zhao and M. Pietikäinen. Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(6) :915–928, 2007.

[77] B. Zhou, X. Hou, and L. Zhang. A phase discrepancy analysis of object motion. In *InProceeding of the 10th Asian Confernce of Computer Vision*, pages 225–238, 2011.

Document réalisé avec LaTeX et :
le style LaTeX pour ThÃÍse de Doctorat créé par S. Galland — http://www.multiagent.fr/ThesisStyle
la collection de paquets `tex-upmethodology` — http://www.arakhne.org/tex-upmethodology/

SPIM

■ École doctorale SPIM - Université de Bourgogne/UFR ST BP 47870 F - 21078 Dijon cedex
■ tél. +33 (0)3 80 39 59 10 ■ ed-spim@univ-fcomte.fr ■ www.ed-spim.univ-fcomte.fr

UNIVERSITÉ DE BOURGOGNE