VISUAL SALIENCY ANALYSIS AND APPLICATIONS

NGUYEN VAN TAM

NATIONAL UNIVERSITY OF SINGAPORE

2013

VISUAL SALIENCY ANALYSIS AND APPLICATIONS

NGUYEN VAN TAM

B.Sc., University of Science, Vietnam, 2005

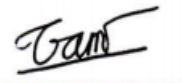
M.E., Chonnam National University, South Korea, 2009

A THESIS SUBMITTED
FOR THE DEGREE OF DOCTOR OF PHILOSOPHY
DEPARTMENT OF ELECTRICAL AND
COMPUTER ENGINEERING
NATIONAL UNIVERSITY OF SINGAPORE
2013

DECLARATION

I hereby declare that the thesis is my original work and it has been written by me in its entirety. I have duly acknowledged all the sources of information which have been used in the thesis.

This thesis has also not been submitted for any degree in any university previously.



NGUYEN VAN TAM

23 July 2013

Acknowledgements

I wish to thank many people who have in one way or another helped me throughout my Ph.D. study.

First and foremost, I thank my advisor, Assoc. Prof. Shuicheng Yan for his constant encouragement and patient guidance throughout the research carried out in this thesis. I benefited tremendously from his vision, his right questions, and his passion for conducting research that matters. Advice and support from Prof. Mohan Kankanhalli, Assoc. Prof. Loong Fah Cheong and Dr. Qi Zhao will be always remembered.

This work would have been impossible without the support of my best friends, Vu Le, Jiashi Feng, Luoqi Liu, Jian Dong, Bin Cheng, and Mengdi Xu. You are always there for me, when I need help with my research. You have given me the courage to make the next transitions in my life. For all of this, I thank you. I would also like to express my sincere thanks to all of my labmates from Learning and Vision research group. I have learned so much from all of you, from figuring out what research is, to choosing a research direction, to learning how to present my work. Your constructive criticism and collaboration have been precious assets throughout my study.

Importantly, I owe immense gratitude to my family. I thank my parents for their love and precious support throughout my life. Finally, this thesis is dedicated to my beloved wife, Kim Anh and my daughter, Minh Thao. You are by my side when I turn to in good times and in bad. You spent much time waiting for me to complete my research work. You have always had such tremendous faith in me and never failed to remind me that I can do anything I set my mind to even when I most doubted myself. Without your endless sacrifice, understanding and love, I would not finish my study.

Contents

Summary

Visual saliency refers to the preferential fixation on conspicuous or meaningful regions in a scene that have also been shown to correspond with important objects and their relationships. It is naturally built into the complex biological system to rapidly detect potential prey, predators, or mates in the real world. Visual saliency is also crucial for human visual experience and also relevant to many applications. Visual attention - particularly stimulus-driven, saliency-based attention - has been an active research field over the past decades. Many attention models are now available, which aside from lending theoretical contributions to other research areas, have given rise to successful applications in computer vision, mobile robotics, and cognitive systems. Here in this thesis, we analyze the visual saliency and its applications in image re-attentionizing, depth matters, video captioning and action recognition.

In the first work, we propose a computational framework, called *Image Re-Attentionizing*, to endow the target region in an image with the ability of attracting human visual attention. In particular, the objective is to recolor the target superpixels by color transfer with naturalness and smoothness preserved yet saliency augmented. We propose to approach this objective within the Markov Random Field (MRF) framework and an extended graph cuts method is developed as a solution. The input image is first segmented into superpixels, and those within the target region as well as their neighbors are used to construct the consistency graphs. Within the MRF framework, the unitary potentials are defined to encourage each target superpixel to match with the patches with similar shapes and textures from a large patch database, each of which corresponds to a high-saliency region in one image, while the spatial and color coherences are reinforced as pairwise potentials. We evaluate the proposed method on the collected *Forbes Ad Dataset*, and the user studies demonstrate that for the recolored images, the target region(s) successfully attract human attention and in the meantime both spatial and color coherences are well preserved.

In the second work, we study the saliency in 3D scenes. In literature, most previous studies on visual saliency have only focused on static or dynamic 2D scenes. Since the human visual system has evolved predominantly in the natural three dimensional environments, it is important to study whether and how depth information influences visual saliency. For this task, we first collect a large human eye fixation database compiled from a pool of 600 2D-vs-3D image pairs viewed by 80 subjects, where the depth information is directly provided by the Kinect camera and the eye tracking data are captured in both 2D and 3D free-viewing experiments. We then analyze the major discrepancies between 2D and 3D human fixation data of the same scenes, which are further abstracted and modeled as novel depth priors. Finally, we evaluate the performances of several state-of-the-art saliency detection models over 3D images, and propose solutions to enhance their performances by integrating the depth priors.

In the third work, we conduct comparative studies between the static saliency and dynamic saliency. We construct the datasets of human fixation on both images and videos for the comparison purpose. Then we make several observations of the relationship of static and dynamic saliency. Inspired by these observations, we propose the noval CMASS learning framework to fuse static saliency into dynamic saliency estimation to improve the video saliency prediction.

Furthermore, we also investigate the application of visual saliency in recognizing human actions in realistic videos. Many works have been devoted to this challenging problem, and breakthroughs have been made gradually. Therefore, we propose transferring the visual saliency based models to such the human action recognition task.

To summarize, our work has outperformed the state-of-the-art methods in different problems and validated the effectiveness of visual saliency. Beyond the aforementioned directions, we foresee more applications of visual saliency in image classification, video summarization and avatar thumbnailing.

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Motivation

Human visual exploration and selection of specific regions for detailed processing are permitted by the visual attention mechanism. The eyes remain nearly stationary during fixation events as humans look at details in selected locations, which makes eye movements a valuable proxy to understand human attention. Visual saliency refers to the preferential fixation on conspicuous or meaningful regions in a scene that have also been shown to correspond with important objects and their relationships. Since visual saliency is believed to drive human fixation during free viewing [116], it is crucial for human visual experience and also relevant to many applications, such as automatic image collection browsing, image segmentation and image decolorization.

Over the last several decades, many research efforts have been devoted toward the further understanding of the mechanisms that underlie visual sampling, either through observing fixational eye movements, or considering the control of focal cortical processing. The consideration of fixational eye movements necessarily involves two distinct components, one being the top-down task-dependent influence on these behaviors, and the second characterized by bottom-up stimulus-driven factors caused by the specific nature of the visual stimulus. The concept of saliency has been extensively studied by psychologists [136, 50, 103, 118, 129]. Later, the proposal for saliency computation within the visual cortex is put forth based on

the premise that localized saliency computation serves to maximize the information sampled from one's environment. It is demonstrated that a variety of visual search behaviors appear as emergent properties of the model such as information coding, and probability [51, 18, 127]. Visual saliency also benefits other research works [22, 58, 8].

1.2 Thesis Focus and Main Contributions

In this thesis, we will explore several different areas in multimedia and computer vision related to visual saliency. In particular, we will introduce the applications of saliency in the image re-attentionizing, depth matters, video saliency prediction and action recognition. Figure 1.1 shows the foci of the thesis.

1. Image re-attentionizing. We propose a novel computational framework to endow the target region in an image with the ability of attracting human visual attention. In particular, the objective is to recolor the target patches by color transfer with naturalness and smoothness preserved yet visual attention augmented. We propose an approach within the Markov Random Field (MRF) framework and an extended graph cuts method is developed. In our work, the input image is first over-segmented into patches, and the patches within the target region as well as their neighbors are used to construct the consistency graphs. Within the MRF framework, the unitary potentials are defined to encourage each target patch to match the patches with similar shapes and textures from a large salient patch database, each of which corresponds to a high-saliency region in one image, while the spatial and color coherences are reinforced as pairwise potentials. We evaluate the proposed method on the AdSaliency dataset. The results demonstrate that the target region(s) successfully attract human attention and in the meantime both spatial and color coherences are well preserved.

2. Depth matters. We investigate the impact of depth in visual saliency. Most previous studies on visual saliency have only focused on static or dynamic 2D scenes. In this work, we first collect a large human eye fixation database

compiled from a pool of 600 2D-vs-3D image pairs viewed by 80 subjects, where the depth information is directly provided by the Kinect camera and the eye tracking data are captured in both 2D and 3D free-viewing experiments. We then analyze the major discrepancies between 2D and 3D human fixation data of the same scenes, which are further abstracted and modeled as novel depth priors. Finally, we evaluate the performances of several state-of-the-art saliency detection models on 3D images, and propose solutions to enhance their performances by integrating the depth priors.

3. Video saliency prediction. We conduct comprehensive comparative studies of dynamic saliency (video shots) and static saliency (key frames of the corresponding video shots), and two key observations are obtained: 1) video saliency is often different from, yet quite related with, image saliency, and 2) camera motions, such as tilting, panning or zooming, affect dynamic saliency significantly. Motivated by these observations, we propose a novel camera motion and image saliency aware model for dynamic saliency prediction. The extensive experiments on two static-vs-dynamic saliency datasets collected by us show that our proposed method outperforms the state-of-the-art methods for dynamic saliency prediction. Finally, we also introduce the application of dynamic saliency prediction in dynamic video captioning, and assisting people with hearing impairments to better enjoy videos with only off-screen voices, *e.g.*, documentary films, news videos and sports videos.

4. Action recognition. We further study the application of video saliency in human action recognition. Human action recognition is useful for many practical applications, e.g., gaming, video surveillance, and video search. We hypothesize that the classification of activities can be improved by smartly designing a feature pooling strategy in the prevalently used bag-of-words classification scheme. We utilize the feature pooling driven by video saliency and propose the Spatial-Temporal Attention-aware Pooling (STAP) scheme. Firstly, we detect salient visual semantics using bio-inspired visual saliency models, and then a spatial temporal feature pooling is performed according to the saliency levels. The kernels later match different levels of video foreground (salient areas) and background (non-salient areas). Finally the kernels are fed into popular support vector machines for classification. Extensive experiments on the evaluated datasets show that our proposed method

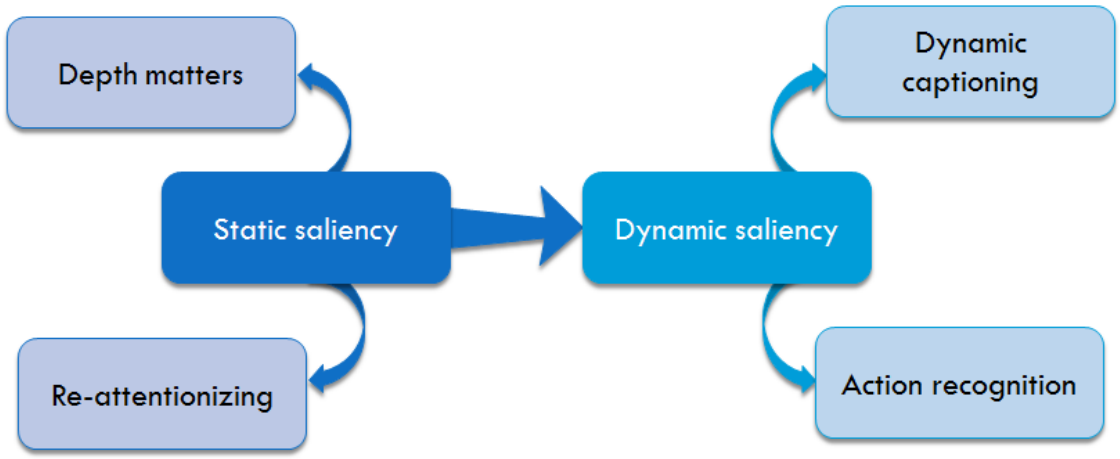


Figure 1.1: The organization of our thesis. The first two works explored two aspects of static saliency, 3d depth matters and image re-attentionizing. The last two works focus on dynamic saliency and its application on dynamic captioning and action recognition.

outperforms state-of-the-art bag-of-words based methods, namely 62.5% on Hollywood2 (better by 4.2%), 87.9% on YouTube dataset (better by 3.7%), and 95.3% on UCF Sports (better by 0.3%).

1.3 Organization of the Thesis

The structure of the thesis is as follows. In Chapter 2, we give a brief review of visual saliency research. In Chapter 3 we introduce the proposed methodologies for human visual attention retargeting. Then, the work on 3D Saliency is given in Chapter 4. The comparative studies of video saliency are presented in Section 5. We further investigate the application of video saliency in action recognition in Section 6. Finally, Chapter 7 concludes this thesis with discussions for future exploration.

Chapter 2

Visual Saliency - Literature Review

In this chapter, we summarize the research works on visual saliency. Specifically, we introduce and discuss the experiment setups, available datasets, various saliency computational models and the applications.

2.1 Experimental setups

The visual saliency research has been originated from the psychophysical area in the 20^{th} century when the modern eye tracker was not even invented yet. Eye movements were first studied in the 1950s and 1960s by Yarbus [136]. He pioneered the study of saccadic exploration of complex images, by recording the eye movements performed by observers while viewing natural objects and scenes. In his work, Yarbus showed that the eye gazes depend on the task that the observer has to perform. Figure 2.1 shows the stimulus and corresponding eye gaze to different task. The gaze tends to jump back and forth between the same parts of the scene, for example, the eyes and mouth in the picture of a face. If the participant was asked specific questions about the images such as human age, position, his/her eyes would concentrate on areas of the images relevant to the questions.

Later, the research on capturing eye fixation data grows rapidly based on the invention of eye trackers. Frank Schumann et al. introduced their work about

recording and analyzing a large amount of video data to compare the spatial distribution of stimulus features in head and gaze centered coordinates during free natural exploration behavior [108]. Their system, EyeSeeCam, including gaze camera, eye trackers and head camera is depicted in Figure 2.2(a). The results point out for a realistic assessment of the role of eye-movements relative to head-centered coordinates, stimuli should be biased. Also, the results show that for a truthful recording of natural human input, head-fixed recordings are not sufficient, and gaze-centered stimuli should be recorded in a situation where eyes, body, and head can freely move. The authors discovered the relationship between indoor environments and some outdoor ones, however, the classification can be done based on the finding of similarity of features in those environments. One common issue in such experiment is that the authors did not mention about the information and prior knowledge of the participating observers.

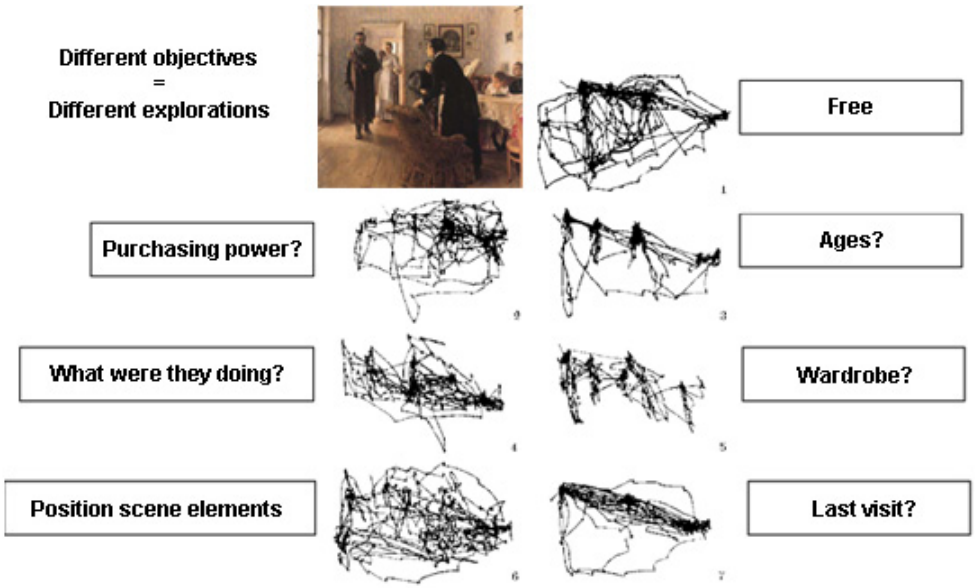
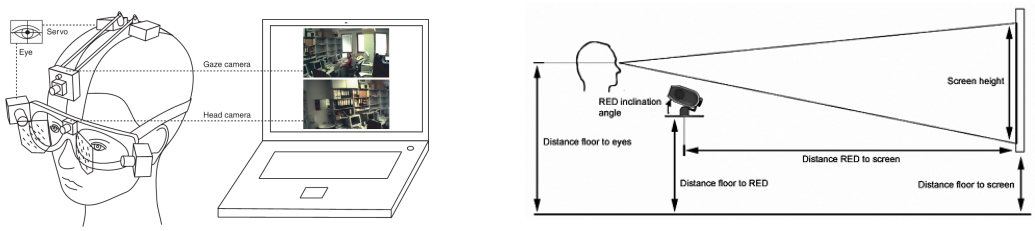


Figure 2.1: Yarbus experiment. Seven records of eye movements by the same subject. Each record lasted 3 minutes. The eye movements are different according to the given question.



(a) (b)

Figure 2.2: Eye gaze tracking system. (a) The schematic view of a head mounted display eye tracker, (b) Infrared eye tracker bar

2.2 Datasets for saliency computation

Eye fixation are an excellent modality to learn semantics-driven human understanding of images, which is vastly different from feature-driven approaches employed by saliency computation models. Following the earlier works like Yarbus's experiment and the emergence of eye trackers, a number of fixation datasets have been constructed for visual saliency research such as Bruce's dataset [18], FIFA [21], NUSEF [102] or MIT [59]. The data set from Bruce and Tsotsos contains data from 11 subjects across 120 color images of outdoor and indoor scenes. Participants were given no particular instructions except to observe the images, 4 seconds each. For FIFA data set, fixation data were collected from 8 subjects performing a 2-s-long free-viewing task on 180 color natural images. They were asked to rate, on a scale of 1 through 10, how interesting each image was. Scenes were indoor and outdoor still images in color. Images include faces in various skin colors, age groups, gender, positions, and sizes.

The eye-tracking data set from MIT is the largest one to date. It includes 1003 images collected from Flickr and LabelMe. Eye movement data were recorded from 15 users who free-viewed these images for 3 seconds. The most recent built NUSEF, an eye fixation database compiled from a pool of 758 images and 75 subjects, aims to learn the preferential visual attention. The database comprises fixation patterns acquired using an eye-tracker, as subjects free-viewed images corresponding to many semantic categories such as faces (human and mammal), and actions (look,

Figure 2.3: Exemplar images from various semantic categories (top) and corresponding gaze patterns (bottom) from NUSEF. Darker circles denote earlier fixations while whiter circles denote later fixations. Circle sizes denote fixation duration.

read and shoot). Figure 2.3 depicts the exemplar images with corresponding gaze patterns of NUSEF dataset. The consistent presence of fixations clusters around specific image regions confirms that visual attention is not subjective, but is directed towards salient objects and object-interactions. As stated in [102], detection of visually salient image regions is useful for applications like object segmentation, adaptive compression, and object recognition. The authors already utilized fixation data to perform tasks, *e.g.* applying mean-shift to cluster eye fixation data and then performing the segmentation task.

Recently, Mathe et al. have collected, and made available to the research community, a set of comprehensive human eye-tracking annotations for Hollywood-2 and UCF Sports, some of the most challenging, recently created action recognition datasets in the computer vision community [84].

2.3 Saliency computational models

Following the psychological experiments, some emerging research focused on computational model from computer science community. Saliency estimation methods can broadly be classified as bottom-up or top-down models.

2.3.1 Bottom-up saliency models

In general, most methods employ a low-level approach of determining contrast of image regions relative to their surroundings using one or more features of intensity, color, and orientation. One of the pioneer is Itti with the well-known work using winner-take-all model [51]. The model combines different visual sub-modalities like colors, intensity and orientations into an overall saliency map. Subsequently, a winner-take-all network defines which spatial position on this map will be considered as the next focus of attention. The weights of the different sub-modalities can be adjusted in the process. This gives the opportunity to steer those weights by a top-down attention mechanism. Meanwhile, Hou et al. presented a spectra residual method to compute visual saliency [47]. The spectral residual resolves the problem of weighting features from different channels (for example, shape, texture, and orientations).

In the other work, Bruce and Tsotsos introduced the research about saliency, attention and visual search based on an information theoretic approach [18]. This work was inspired from Attneave's experiment. Unlike the previous models offering little in explaining why the operations involved in the model have the structure that is observed, the authors focus on explaining why certain components implicated in visual saliency computation behave as they do. They proposed a new framework, AIM, which maximizes the information to compute saliency map. The authors mentioned sparse coding when using ICA to generate basis coefficients. Sparse coding problem is a very interesting topic in visual saliency. Xiaodi Hou et. al in [46] use 192-dimension sparse features, but they still achieve the similar results to this work. Another approach using Bayesian inference theory proposed by Chikkerur et al [26]. Their model resembles the interaction between the parietal and ventral streams mediated by feedforward and feedback connections. One issue is the assumption from the authors. They assumed "To achieve this goal, the visual system selects and localizes objects, one object at a time". However, as discussed in [131], the human visual system does not only do serial work, but also it applies parallel work. In addition, is it true to say "the object location and object identity are independent"? Actually, in some cases, object location and object identity are

not independent. The authors compare their work with other works from Itti et al, Bruce and Tsotsos, etc. Mahadevan et al. proposed a spatiotemporal saliency algorithm based on a center-surround framework [80]. The algorithm is inspired by biological mechanisms of motion-based perceptual grouping and extends a discriminant formulation of center-surround saliency previously proposed for static imagery. The paper offers new insight and gives clear comparison to other methods. It has, however, some shortcomings as follows. This work from Mahadevan et al. extended the work from Gao et al [37]. The original work proposed the usage of discriminant saliency in static scene. In the extended work, Mahadevan et al. applied the original work to dynamic scene. One common issue for such models is the processing time. The processing time to compute the saliency map is very slow. For example, for one testing frame with size 340×256, the average processing time is about 7.1 seconds. Since the target of detecting salient region is to boost the processing time for the other tasks, it is impossible to compute the saliency map in the real-time manner.

2.3.2 Top-down saliency models

Some top-down factors in free-viewing are already known although active investigation still continues to discover more semantic factors. For instance, Einhauser et al. proposed that objects are better predictors of fixations than bottom-up saliency [31]. Elazary et al. showed that interesting objects (annotations from LabelMe dataset [105]) are more salient [32]. Cerf et al. showed that the meaning objects such as faces and text attract human attention [21]. Similarly , Judd et al., observed that humans, faces, cars, text, and animals attract human gaze increasingly [59] by plotting image regions at the top salient locations of the human saliency map (built from eye fixations). These objects convey more information in a scene. Alongside, some personal characteristics such as experience, age, and culture change the way humans look at images.

The basic idea is that a weighted combination of features, where weights are learned from a large repository of eye movements over natural images, can enhance

saliency detection compared with unadjusted combination of feature maps. Kienzle et al. [60], Judd et al. [59] and Peters et al. [100], used image patches, a vector of several features at each pixel, and scene gist, respectively for learning saliency. Zhao and Koch learned optimal weights for saliency channel combination separately for each eye-tracking dataset [141]. While they show tunning weights for each dataset results in high accuracies, learned weights sometimes do not agree over datasets. It is also unclear how this approach generalizes to unseen images.

2.4 Applications

There are many applications of saliency prediction models in computer vision, mobile robotics, and other systems [12].

2.4.1 Computer Vision

Avidan et al. introduced a method for content-aware resizing of images using seam carving [8]. Seams are computed as the optimal paths on a single image and are either removed or inserted from an image. This method can be used for a variety of image manipulations including: aspect ratio change, image retargeting, content amplification and object removal. Figure 2.4 illustrates the use of seam carving. Wang et al. presented Picture Collage which is a kind of visual image summary to arrange all input images on a given canvas, allowing overlay, to maximize visible visual information [126]. They formulated the picture collage creation problem in a Bayesian framework. The salient regions of each image are firstly extracted and represented as a set of weighted rectangles. Then, the image arrangement is formulated as a Maximum a Posterior (MAP) problem such that the output picture collage shows as many visible salient regions (without being overlaid by others) from all images as possible. Sadaka et al. propose an attentive super-resolution technique that exploits the available saliency information of the active pixels to further reduce the computational complexity while achieving the desired

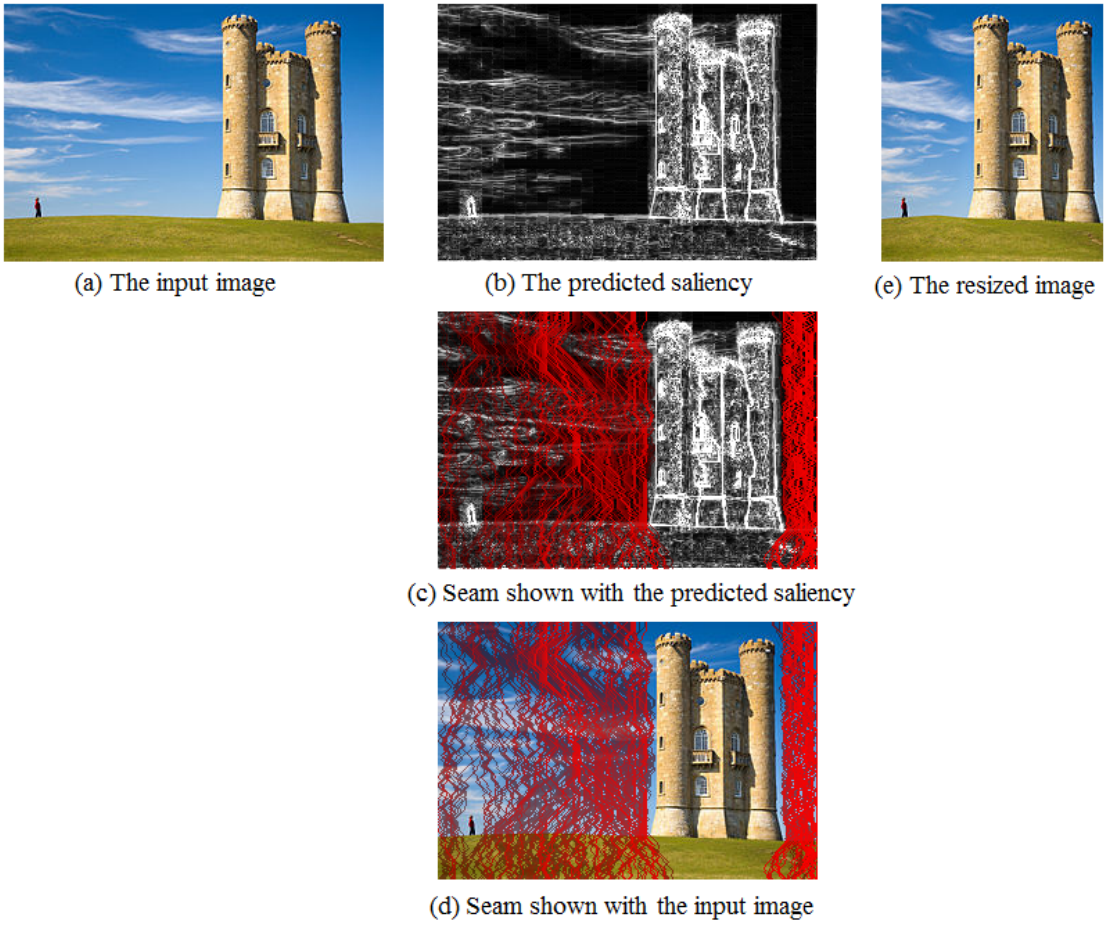


Figure 2.4: The illustration of seam carving using saliency to resize the input image.

visual quality of the high resolution image [106]. Later, Liu et al. proposed the new method to detect salient objects automatically with the supposition that a salient object exists in an image [78].

2.4.2 Robotics

Regarding the mobile robotics, Dankers et al. have developed a synthetic active visual system capable of detecting and reacting to unique and dynamic visual stimuli, and of being tailored to perform basic visual tasks [30]. For example, when driving a car, we tend to keep our eyes on the road, and as such we bias the lower

half of the mosaic where we would expect to find the road. The system can be preempted for regions not in the current view frame. Meanwhile, Siagian et al. proposed using salient regions as the localization cues. The proposed method is efficient since it only process on the salient regions instead on the whole scene [111]. Muhl et al. proposed a human-robot interactive system using saliency [92]. In their work, the robot gazed at the most salient location in each video frame. The robot's eyes were controlled so that human partners could perceive that it was responding to their action and was looking at an interesting location for it. Gadde et al. built a robot to recapture a better photograph by assessing the visual quality of the captured photo [35]. The strength of their approach is the computational efficiency which can be applied in autonomous robots. The accuracy can be improved further by adding symmetry in the subject region as mandatory since images with some symmetry are rated higher than the rest and with more complicated composition guidelines of professional photography. Courty et al. proposed using saliency for video surveillance application [27]. The principle of this application is quite simple: each frame acquired by the camera is processed and a feature map that includes both spatial and spatio-temporal information is created. The global maximum of the map is determined through a simple scanning of the feature map and given as input of the visual servoing process. The pan/tilt camera is then focused at this point.

2.4.3 Other applications

For intelligent advertisement, Mei et al. introduced ImageSense [86] and VideoSense [85] which is able to automatically decompose the Web page into several coherent blocks, utilize salient regions to select the suitable images from these blocks for advertising, detect the nonintrusive ad insertion positions within the images and videos. Hong et al. have developed a segmentation method to detect salient regions in mammograms [44]. Salient regions correspond to distinctive areas that may include the breast boundary, the pectoral muscle, candidate masses and some other dense tissue regions. Wong et al. introduced a saliency-enhanced method for the classification of professional photos and snapshots [132]. They extract the salient

regions from an image by utilizing a visual saliency model. Then, in addition to a set of discriminative global image features, they extract a set of salient features that characterize the subject and depict the subject-background relationship. Liu et al. presented a generic virtual content insertion system [75] which determines insertion time by detecting higher attentive shot with temporal attention analysis, and determines insertion place by detecting lower attention region with spatial attention analysis. By inserting virtual content into the attentive shots at lower attention region, the system balances between the notice of the virtual content by audience and disruption of viewing experience to the original content.

Chapter 3

Image Re-Attentionizing

In this chapter, we introduce a new application of visual saliency, namely, image attention retargeting. We propose the new framework in order to modify some image regions so that those regions attract more human attention than the original ones. We also introduce a new dataset which is served for the evaluation of this task.

3.1 Introduction

Retargeting the human attention to certain part(s) of an image benefits many applications, *e.g.*, intelligent advertisement, image editing, and image assessment. For example, people tend to skim through the photos when they read magazines and newspapers, especially in the advertisement columns. Therefore to implicitly attract readers to where the advertisers want is important. From the view-point of an advertiser, placing the advertisement at the right spot is a critical task. As another example, people often take amateurish photos that have the wrong object being the objects of interest. Therefore, it is desirable to emphasize the intended objects during photo editing.

There exist few attempts for this task, and these methods [112] [133] simply alter the global features or local features based on neighborhood information. As admitted in [133], maintaining naturalness is a challenging issue. For example,

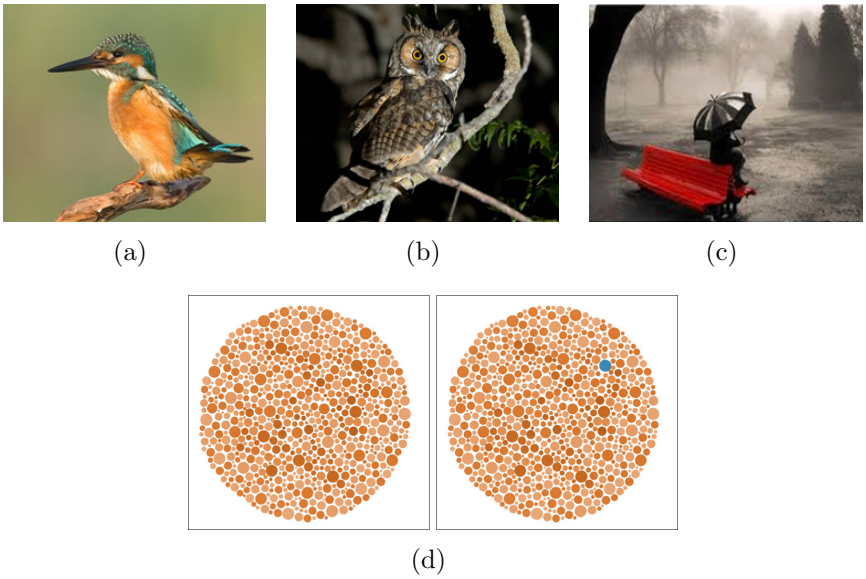


Figure 3.1: What is human visual attention drawn to? Visual dominance of the subject can be achieved using (a) acutely sharp focus, (b) lighting contrast, and (c) color contrast. (d) The blue dot in the right image receives the higher attention than the original dot in the left image because it is different from the rest.

blurring is often utilized in saliency retargeting to force focus to the main subjects. However, in reality, there is no blur when people perceive the scene. The blur effect is essentially caused by human visual system. When people fixate at an interesting area, points at fixation are sharply focused, but points away from fixation are increasingly blurred. In fact, those effects which simulate the human biological blur are not favored by humans [97].

In literature, some research works focused on computational saliency model from computer vision, human vision and psychology community. In general, most methods employ a low-level approach of determining contrast of image regions relative to their surroundings. Itti et al. combined different visual sub-modalities like colors, intensity and orientations into an overall saliency map [51]. Meanwhile, Hou et al. presented a spectral residual method to compute visual saliency [47]. Bruce et al. proposed the usage of information maximization in order to predict saliency [18]. Judd et al. applied the learning method in order to compute the saliency value of the pixel from the low level features. Recently, many researchers

proposed various computational models to compute saliency maps from images such as Saliency by Induction Mechanisms (SIM) [93], Region-based Contrast (RC) [24], Frequency Tune (FT) [5] or co-saliency models [70]. However such these works are for saliency prediction, rather than for saliency retargeting.

In this chapter, we propose a novel computational framework, called *Image Re-Attentionizing*, to recolor an image so that human attention may be relocated to the target region and also the image naturalness may be well preserved. Following Oxford English Dictionary [1], we adopt the definition of 'naturalness' as 'lack of artificiality in conduct or bearing'. To facilitate such an objective, we formulate the problem within the Markov Random Field (MRF) framework [57]. Instead of dealing with individual pixels, we follow the patch-based approach inspired from [33, 135]. The input image is first over-segmented into patches, and each patch is considered as a node of MRF. Then, the unitary potentials of MRF encourage the target image patches to match the high-saliency patches in a training image patch dataset, and the image smoothness and coherence between patches are reinforced as pairwise potentials. The solution is effectively sought by a refined graph cuts method.

The contributions of our approach are as follows.

1. Our method considers both spatial coherence and color coherence, and thus the recolored image is natural.
2. We utilize a *salient patch dataset* that includes the recorded human fixation data. The intention of using salient patch dataset is based on the assumption that the patch with larger saliency ratio gets the higher chance to attract the user attention (namely in a supervised way or using priors).

3.2 Related Works

Targeting the main subject is a classical topic in photography. As stated in [36], the basic rule of *ABC* (an abbreviation of ***A***cutely sharp, ***B***right, and ***C***olorful) can

be applied to attract human attention. Figure 3.1.(a-c) show examples of the *ABC* rule. First, acutely sharp elements mean that the whole picture is blurry except there is one sharp object. Second, bright elements refer the dark picture with one bright object. Last, colorful elements indicate that the whole picture is black and white or monochrome and there is only one color saturated object. Actually, photography rules are not the only cues to locate human attention. Attneave's psychological experiment [7] showed that redundant regions of an image are often skipped and instead the viewers focus on the rare regions in the image. As shown in Figure 3.1.d, the blue dot in the right image receives the higher attention than the original dot in the left image since it is different from the rest. The case that certain stimuli seem to be found effortlessly from others is called the *pop-out* phenomenon in psychology study.

There exist few works on saliency retargeting [112, 133, 132, 114]. Su et al. [112] first proposed the idea of altering the predicted saliency of an image by reducing the background saliency to redirect attention to the main subject. Their method utilizes texture power maps to de-emphasize texture variations to decrease the saliency of distracting regions. This method preserves key features, however since adding white noise maintains the overall graininess, the resulting images appear too noisy and do not seem to be as natural as their originals. Wong et al. proposed the concept of applying saliency retargeting for enhance image aesthetics [133]. The method modifies only the low-level image features that correspond directly to the features used in the saliency computation model of [51]. In the cases that photos compose only one subject, saliency retargeting can make them more acceptable than before; in other cases, saliency retargeting cannot help improve the visual attention if there exist many subjects and the main subjects are not salient.

The above mentioned methods can be regarded as passive methods, which change the rest of image to force the focus to the main subject. For example, in order to drive human attention to certain area(s), Gaussian blur [41] is applied to the image region except the target region. However, this passive approach is questionable. First, the blur unavoidably causes information loss in those complementary regions (the regions which are not the target regions in the image), which sometimes are important. Second, blur might not effectively change the predicted

saliency map. As can be seen in Figure 3.2, the generated saliency maps are similar in both cases for images before and after processing.

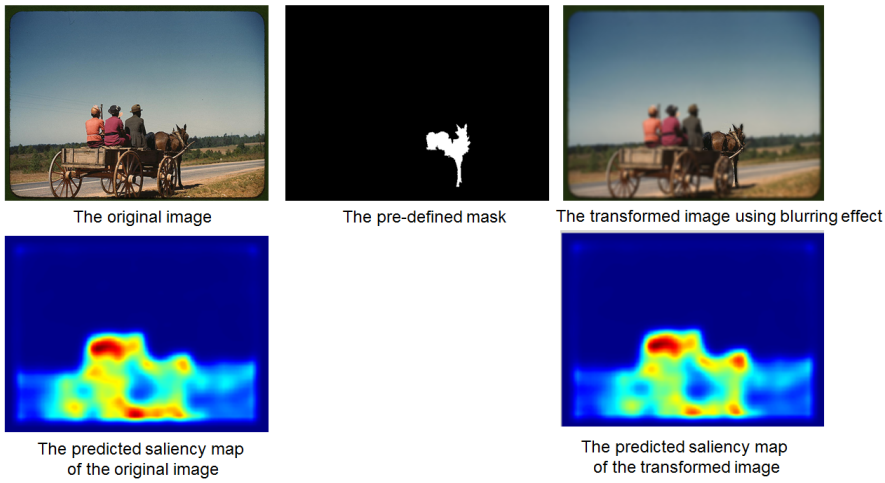


Figure 3.2: The comparison results show the blurring effect does not change much the saliency map. The transformed image (top-right) is achieved by applying the Gaussian blur [41] on input image (top-left) with the pre-defined mask (top-mid), and their corresponding predicted saliency maps are computed by [18] (bottom row). Note that red values in saliency map represent higher saliency, while blue values mean lower saliency.

3.3 Image Attention Retargeting

In this work, we propose a new computational framework which actively recolors only the main subject to make it stand out, in both local and global sense. In this way, the information of the complementary regions is well preserved. We utilize a *salient patch dataset* that includes the recorded human fixation data. We compile most existing eye fixation datasets. The used fixation datasets include Bruce's dataset [18], MIT [59], NUSEF [102] and FIFA [21]. Totally $2,165$ images with $630,288$ patches have been extracted. The patches in the salient patch dataset are indexed as $\{1, 2, 3, ...\}$. Figure 3.3 illustrates our framework with an example, and from the result, we may notice that the coherence of the recolored image is maintained, *e.g.*, the colors of the pair of gloves are consistent. The main superiority

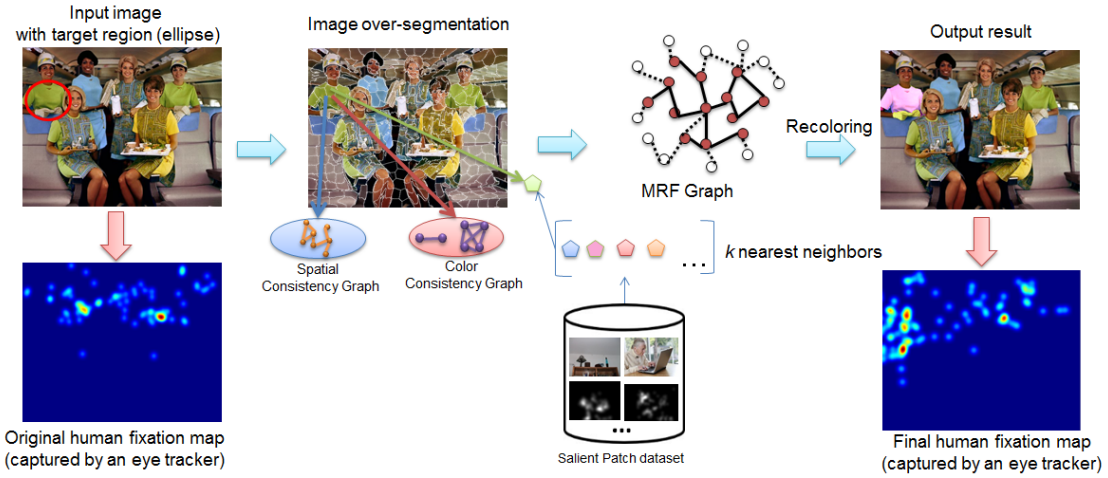


Figure 3.3: Exemplar illustration of image re-attentionizing framework. Note that the human fixation map has been redirected to the target regions. For better viewing, please see original color pdf file.

over related works is that we simultaneously consider image naturalness and the attention retargeting within a unified framework. In the following subsections, we introduce our MRF-based image re-attentionizing framework, including 1) how to build the graph from the user input, 2) the detailed MRF framework, and 3) how to perform the image recoloring.

3.3.1 Consistency Graph Construction

In order to maintain the naturalness, our framework considers the consistency in terms of spatial and color coherence. As shown in Figure 3.3, two consistency graphs, namely spatial consistency and color consistency graphs, are built from the input image with pre-defined target regions. There are two types of nodes: the modified nodes represent the patches in the target regions, and the fixed nodes represent the patches in the complementary regions, which shall remain unchanged after the re-color transformation.

- Spatial consistency graph $G_1 = (V_1, E_1)$ consists of the nodes V_1, which represent patches from the target regions T (modified nodes) and their spatial

neighbors F (fixed nodes). The edges in the graph G1 connect only the neighboring nodes.

This graph is motivated by the fact that to achieve natural inter superpixel appearance transition, superpixels that are transformed should be close to each other, *i.e.*, spatial smoothness.

- Color consistency graph $G_2 = (V_2, E_2)$ consists of the nodes V_2, which represent only patches decomposed from the target regions T. Meanwhile, the edge set E_2 contains all the connections among the nodes in node set V_2. Intuitively, E_2 is the subset of E_1. Two nodes have a connection if their original color information, *i.e.*, color moments, is similar and the Euclidean distance is smaller than a pre-defined threshold (25 in our implementation). Note that color moment is a low-level color measurement and consists of the first order (mean of color values) and the second order moments (variance of color values) of the input image patch. After the image recoloring, two connected patches should again maintain similar colors. This graph is driven by the fact that originally color-coherent patches should still preserve the color consistency in the transformed image, therefore the underlying structure of the original image is not significantly changed.

In order to facilitate the problem formulation, we combine two consistency graphs into one unified graph, $G = (V, E)$. The set of nodes V includes the patches decomposed from F and T. Modified nodes are from the user input while fixed nodes are the spatial neighbors of the modified nodes. Meanwhile, the set of edges E contains two types of edges: the spatial neighboring edges connect the neighboring nodes, and the coherence edges connect the nodes having the similar color information. Figure 3.4 shows a sample graph constructed from the over-segmented image.

3.3.2 Problem Formulation for Image Re-Attentionizing

The target region(s) consist of n patches after image over-segmentation [6]. For a patch indexed by i, f_i^t denotes the feature vector concatenating the sub-features

Figure 3.4: The exemplar MRF graph built on the over-segmented image. For the sake of clarity, only some edges are drawn. Please see original color pdf file for better viewing.

including shape moment [43], texton histogram [110], patch height-to-width ratio (height and width of the patch bounding box) and patch fixation ratio, where each sub-feature is normalized to unit-norm. x_i is the index of the solution patch (in the aforedmentioned salient patch dataset) to recolor patch i (in the test image). However, it is infeasible to consider all the patches in the salient patch dataset for each patch in constructing MRF model. Therefore, we consider only k candidates from the salient patch dataset which closely match the local patch in feature similarity. The similarity between patch i in the test image and patch j in the dataset is calculated as below.

$$d(f_i^t, f_j^d) = \|f_i^t - f_j^d\|_2, \tag{3.1}$$

where f_j^d is the feature vector of the patch j in salient patch dataset, and $\|.\|_2$ is the ℓ_2 norm. The dimensionality of feature f_i^t for each patch is listed as follows, shape moment [43] - 9 dimensions, texton histogram [110] - 200 dimensions, and patch height-to-width ratio and patch fixation ratio - 2 dimensions. We utilize this combination since it covers both shape and texture. Note that the patch fixation ratio in f_j^t is set as 1 in order to encourage the patch with higher saliency ratio to be returned. We utilize kNN-GPU [38] which exploits the speedup of GPU and returns the exact nearest neighbors. Denote the indices of kNN salient patches as π_i, then $x_i \in \pi_i \cup \{0\} = \tilde{\pi}_i$, where 0 means the patch remains unchanged. We compute the color histogram H of the image regions except the target region for the later usage.

After constructing the unified graph G, our task is to find the label set x for all nodes from all kNN of all the nodes, which minimizes the energy function E, consisting of data energy term E_d (unitary potential) and smoothness energy term E_s (pairwise potential).

$$\min_{\{x_i \in \tilde{\pi}_i\}} \{E(x) = \sum_{i \in V} E_d(x_i) + \lambda \sum_i \sum_{j \in N(i)} E_s(x_i, x_j)\}, \tag{3.2}$$

where $N(i)$ denotes the set of spatial neighboring patches of the i^{th} patch. Data cost E_d indicates the cost of the selected patch x_i in the whole image sense. λ, a weight to balance the data term and smoothness term, is empirically set as 0.1. When i is a fixed node,

$$\forall i \in F,\ E_d(x_i = 0) = 0, E_d(x_i \neq 0) = \infty, \tag{3.3}$$

where F is the fixed node set as introduced above. Eqn. (3.3) means that the fixed node will not be changed and the solution label is always 0. If i is not in the fixed node set, E_d is calculated as below,

$$\forall i \notin F,\ E_d(x_i = z) = \begin{cases} \frac{\varphi(x_i = z)}{\sum_{k \in \pi_i} \varphi(x_i = k)} & , z \neq 0, \\ \infty & , z = 0, \end{cases} \tag{3.4}$$

where $\varphi(x_i = z)$ is defined as

$$\varphi(x_i = z) = \frac{1}{l} \sum_{m=1}^{l} \sum_{p=1}^{N_l} H_{pm} \delta(\mu_{zm}, p), \tag{3.5}$$

which measures the redundancy of candidate patch z. l is the number of color channels and N_l is the number of color bins in channel l. μ_{zm} is the mean of the m^{th} color channel of the patch z and H_{pm} is the histogram value of bin p in the m^{th} color channel of the image regions except the target region(s). $\delta(a, b)$ is the binary function, returning 1 if $|a - b| < \gamma$, and 0 otherwise; γ is set as 1 in our implementation. Note that Eqn. (3.4) captures the insight of Attneave's experiment: our eyes are drawn to things that are of the most importance to us, or that will give us the most information [7]. Since the patch z is oversegmented,

most pixels in patch z share the similar color. Therefore, the mean color is used to represent the patch color. Should the histogram value of the color bin similar to the patch mean color is large, the corresponding $E_d(x_i = z)$ is large accordingly. In other words, z is not a good candidature patch. Equation 3.5 encourages the patches to be different from the image regions which exclude the target regions.

Meanwhile, the smoothness energy E_s is defined as

$$E_s(x_i, x_j) = \psi(x_i, x_j)\eta(x_i, x_j). \tag{3.6}$$

The term $\psi(x_i, x_j)$ is defined as

$$\psi(x_i, x_j) = \begin{cases} -1 & \text{if } j \in N(i) \setminus R(i) \\ +1 & \text{if } j \in R(i) \end{cases}, \tag{3.7}$$

where $R(i)$ is the set of neighboring patches connecting to patch i by color consistent edge and $\eta(x_i, x_j)$ is computed as

$$\eta(x_i, x_j) = \|C_{xi} - C_{xj}\|_2, \tag{3.8}$$

where C_{xi} and C_{xj} are the color moment features of the possible solutions for patch i and patch j, respectively. On the one hand, the objective of the smooth term is to emphasize the dis-similarity between the patch and the neighborhood regions in the local context. On the other hand, the smooth term aims to minimize the dis-similarity between two neighboring nodes having a color coherence edge. In other words, the proposed energy function makes the target region salient in both global and local sense. Within the MRF framework, $E(x)$ can be approximately optimized by using graph cuts [15] [62] [14].

3.3.3 Image Recolorization

When transferring the color, we aim to maintain the color consistency. Hence, we extract connected components in color consistency graph G_2 instead of using G,

and apply color transfer [104] for each connected component. Note that the new color $I_T(c, r)$, applied for the pixel at column c and row r, is computed based on the original color $I_O(c, r)$ as follows. The color was first converted from RGB to $l\alpha\beta$ color space [104]. The significant advantage in this space is that changes in one color channel will have minimal influence to other channels. The new color $I_T(c, r)$ is computed as follows.

$$I_T(c, r) = \frac{\sigma_t}{\sigma_s}(I_O(c, r) - \mu_s) + \mu_t, \tag{3.9}$$

where σ_t, μ_t are standard deviation and mean color of the target components, respectively. Note that the target components are the original patches. Meanwhile, σ_s, μ_s are standard deviation and mean color of the source components, respectively. The source components are from the patches obtained from the solution of graph cuts method. The new color I_T for the pixel (c, r) is updated based on the current color I_O, the deviation and the mean color of the original patch contain that pixel, the deviation and the mean color of the solution patch. The intuition of Eqn. (3.9) is to enforce two components to have the same color distribution in the sense of mean and variance statistics, by subtracting the mean value of the target component and scaling the color values based on the ratio of variances.

3.4 Experimental Results

3.4.1 Dataset Collection

As intelligent advertisement is one of the potential applications, for the evaluation, we build up the *AdSaliency Dataset* [1] which includes the crawled advertisement-related images of the top commercial brands [2]. For each brand, we download the advertisement-related images based on the keywords *"ad + brand name"*. Then we intentionally select the images containing multiple objects or humans. Totally

[1]The *AdSaliency* dataset along with the fixation data is available at:
https://sites.google.com/site/vantam/image_re_attentionizing

31 images are used for the evaluation. All images are resized to 640 × 480 or 480 × 640 pixels, according to the aspect ratio of original images. In order to define the target region(s) in the image, one user selects the corresponding patches. The region(s) are randomly selected to avoid the center bias.

We evaluate the proposed approach in two perspectives: how it actually retargets the human visual attentions and the naturalness of the resulting images.

3.4.2 Implementation Settings

In literature, the traditional graph cuts method utilizes the 4-neighbor grid graph along with only one smooth cost matrix for the whole graph, which is not suitable for our approach. That definition is not flexible, especially when we utilize the superpixels. Therefore, instead of using the 4-neighbor grid graph, we employ the general graph, which allows arbitrary neighbors for one node. We modified the implementation of [25] by adding the edges between nodes. In addition, we add the smooth cost matrix for every connection between two nodes in the graph. Again, we change the smooth cost computation function pointing to the newly defined smooth cost matrix.

Figure 3.5: The comparison of the results of our proposed method with different k values. For better viewing, please see original color pdf file.

Meanwhile, k is set as 40 in our implementation for kNN search mentioned in Section 3.3.2 as it is the maximum value for the modified graph cuts implementation to run without memory overflow for our task. Figure 3.5 illustrates the comparison

of the results of our proposed method with different k values. The results with $k = 1$ reveal heavy color inconsistency as aforementioned. The results are different with different k values. However, when k is large enough ($k = 20$ or $k = 40$), the corresponding results are quite similar.

3.4.3 Attention Retargeting Evaluation

60 participants (students and staff members of a university) ranged from 21 to 36 years old (μ=26.9, σ=3.1), with normal or corrected-to-normal vision, volunteered to participate in the experiments. All participants are naive to the purpose of the study and have no prior exposure to experiments on vision. The participants have been split into six groups. Each group performs free-viewing only one of six following categories.

- Original images.
- Blurry effects: Gaussian blurring [41] is applied to the original image except the target region(s).
- Monochrome effects: the original image is applied grayscale filtering [41] except the target region(s).

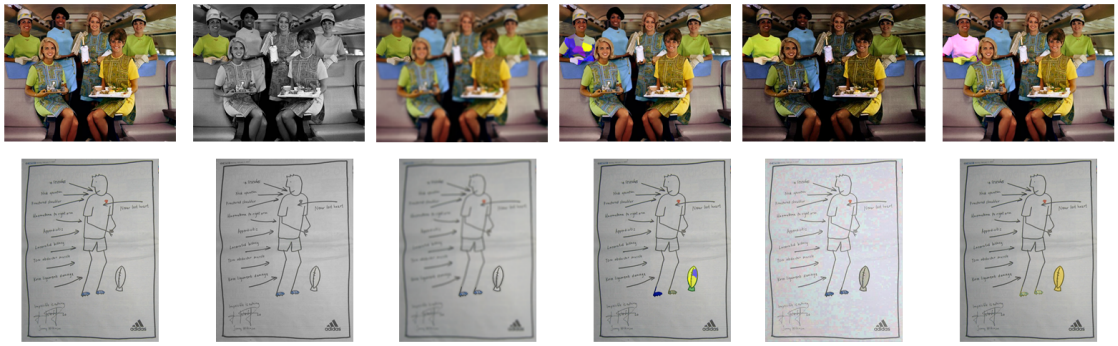


Figure 3.6: Comparison results from different methods. Left to right: Original image, transformed images using monochrome effects, blurring effects, 1 nearest neighbor, Wong et al. method [133] and our approach. For better viewing, please see original color pdf file.

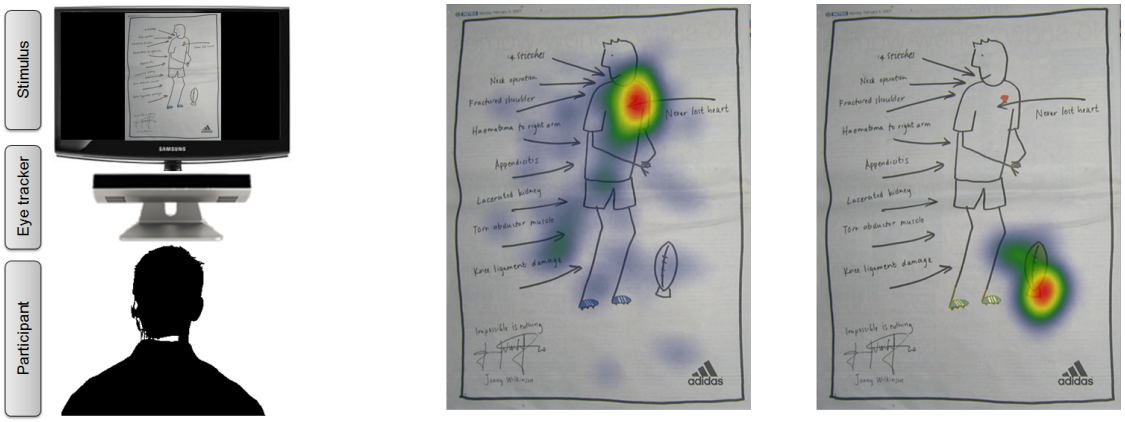


Figure 3.7: (a) The setting of fixation collection with an eye-tracker, (b) Two heatmaps: for the original image (left) and for the recolored image (right). Note the redirection of human fixation.

- 1NN: the resulting image is obtained by color transfer based on the nearest neighbor patch of each target patch.
- Wong et al. [133].
- Ours: the resulting image is obtained through our proposed framework.

Figure 3.6 shows transformed results from different methods. In order to record participants' eye gaze data, we use an infra-red based remote eye-tracker. The eye-tracker gives less than 1^o error on successful calibration. The eye tracker was calibrated for each participant using a 9-point calibration and validation method. Then images were presented in random order for 4 seconds followed by a gray mask for 2 seconds. Similar to [123], in order to produce a fixation map of an image, we convolve a Gaussian filter across all corresponding viewers's fixation locations, similar to the "landscape map" of [123]. Figure 3.7.a sketches the setting of fixation collection and Figure 3.7.b illustrates the heatmaps, the images with the embedded corresponding fixation maps, before and after applying *Image Re-Attentionizing*. More resulting images with the corresponding heatmaps are shown in Figure 3.8. Our experiments are carefully designed and conducted to evaluate different algorithms. The evaluation on the average fixation maps shows how much

Figure 3.8: Some results with human fixation data. For each pair of rows: images (top) and their corresponding heatmaps (bottom). For each row from left to right: the original, blurring effect, monochrome effect, 1nn result, Wong et al. [133], and our result. The target regions are highlighted as ellipses in the original images. For better viewing, please see original color pdf file.

the transformation methods really change the fixation center-bias existing in the original images. In next subsections, we compute Hit Rate and then Cumulative Score to show how much fixation is drawn on the target regions. In order to compare the naturalness of the resulting images, we conduct the comprehensive comparison user study, since automatically predicting the naturalness level of the image is difficult. Moreover, it is worth noting that the problem explored in this work is essentially novel, and its evaluation can only be based on human fixation data.

3.4.4 Center Bias Evaluation

We compute the average fixation maps of original images, masks of target regions and transformed images across various methods. Due to different sizes of testing images, the average fixation maps have cross-like shape. As can be seen in Figure 3.9, the center bias remains strong in the average fixation map of original images which agrees with the finding in [59]. The average map for target region mask is not center-biased due to the intentional purpose of the user. Meanwhile, the average fixation map of transformed images using monochrome effects still has a center bias. It can be explained that the color of some target regions are close to monochrome, which cannot be well distinguished after applying monochrome effects. The center bias is not strong in the average fixation maps of transformed images using remaining methods.

3.4.5 Attention Retargeting Quantitative Comparison

We utilize two measurements to evaluate the saliency retargeting performance: (1) the Hit Rate (HR) and (2) the cumulative score (CS). For each input image I, the corresponding human fixation map obtained by the eye tracker is denoted as H. Given pre-defined mask is denoted as M, HR can be obtained:

$$HR = \frac{\sum_k H_k \times M_k}{\sum_k H_k}, \tag{3.10}$$

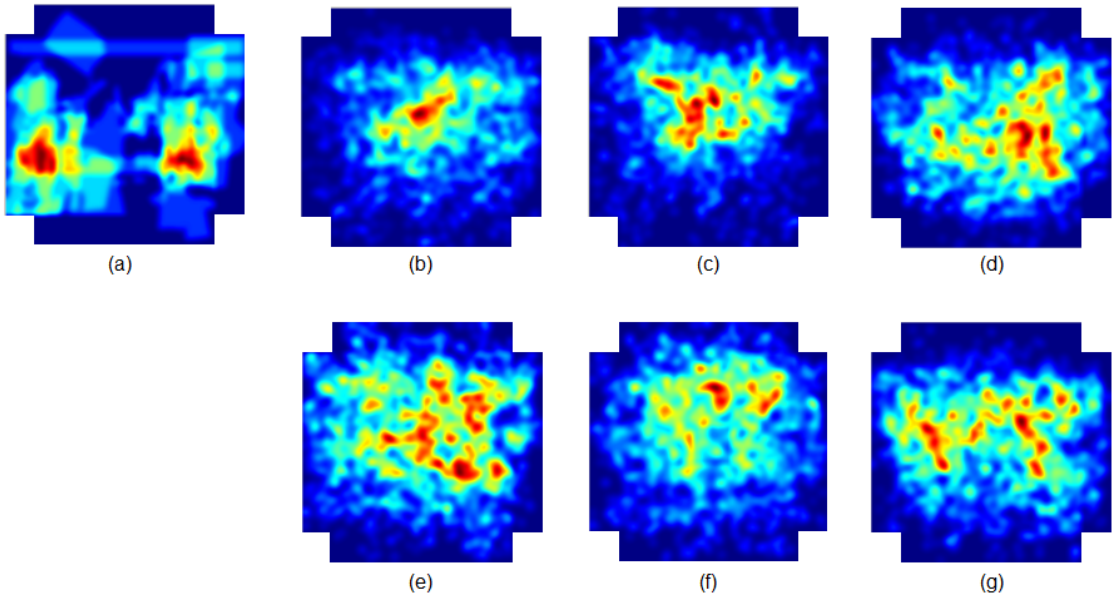


Figure 3.9: (a) The average pre-defined mask map and fixation maps of (b) the original images and the transformed images across (c) monochrome effect, (d) blur effect, (e) 1 nearest neighbor, (f) Wong et al. [133], and (g) our approach.

where H_k, in the interval of $[0, 1]$, is the fixation value of point k in H, and M_k is the corresponding binary value in M. This criterion indicates the proportion of fixation data placed on the target region. We conduct another evaluation on predicted saliency retargeting. Here we utilize Saliency by Induction Mechanisms (SIM) [93], Region-based Contrast (RC) [24], and Frequency Tune (FT) [5] models, which are the recent state-of-the-art saliency prediction models. Those predicted saliency models are applied on the original images, the results of Wong et al. [133], and the results of our method. Figure 3.10 illustrates the results of our method and [133] and their corresponding heatmaps from saliency prediction results and human fixation.

In terms of attention retargeting, Table 3.1 shows that the proposed framework outperforms other methods. The HR of the original images is the lowest. HR increases four times after applying our approach. Not surprisingly, HR values of blurring effects is the second highest one. Meanwhile, 1NN is also a good method to attract visual attention. The monochrome effect does not increase HR as much as other methods. The reason is that the original color of some target regions

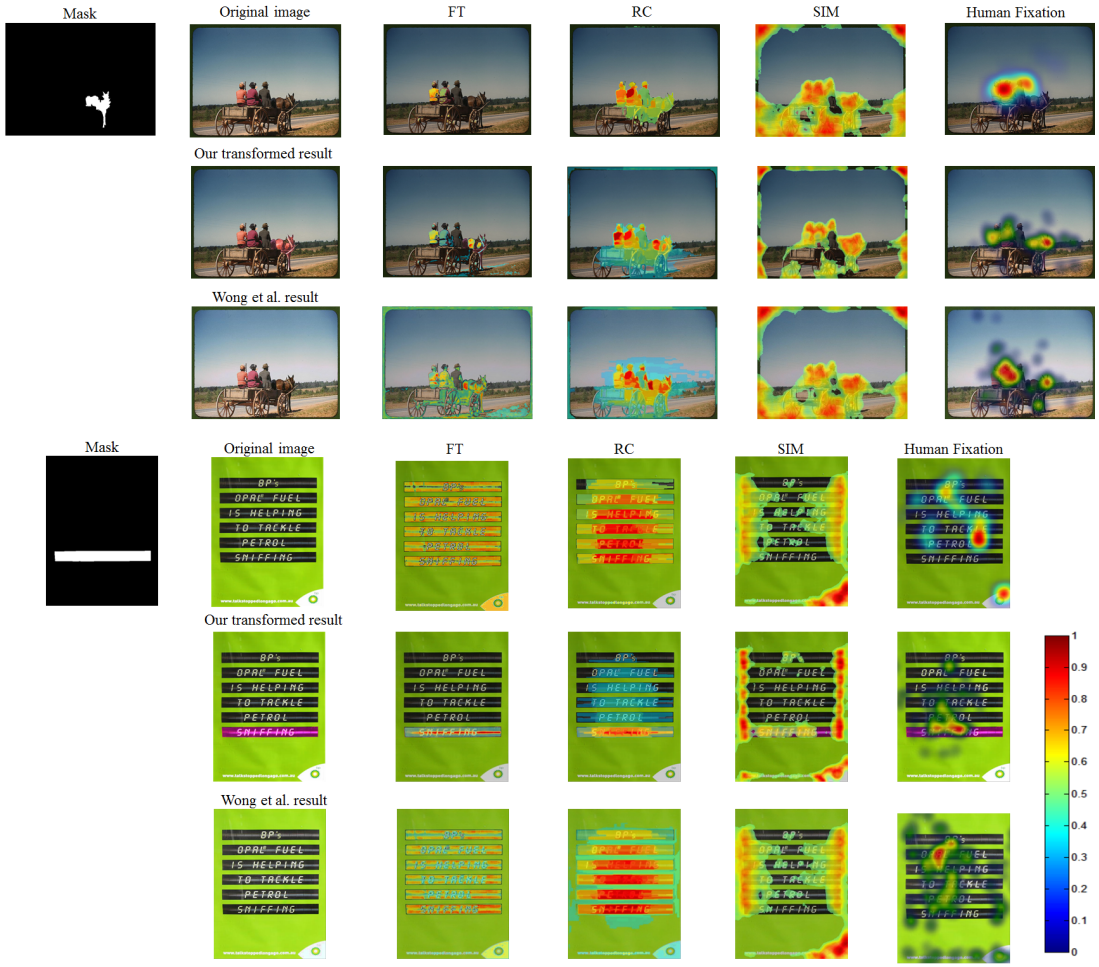


Figure 3.10: The exemplar heatmaps of our results and Wong et al.[133] and their corresponding saliency maps from state-of-the-art predicted saliency models (the reddish pixels are salient, the blue ones are not). Please view in high 200% resolution.

is close to monochrome. Therefore, in order to evaluate the monochrome effects, we divide the *AdSaliency* dataset into 2 groups: gray and non-gray target-region groups. The target regions are considered as gray if the deviation of 3 channels of the mean color of the target regions is smaller than the threshold θ. Here θ is set as 10. The corresponding ratio of gray/non-gray target-region groups is 10/22 in our dataset. Then we compute HR on both groups. The HRs for gray and non-gray target-region groups are 0.066 and 0.22, respectively. HRs of gray target-region images are comparable to the ones of original images. Meanwhile, the higher HRs

Table 3.1: HR values computed across different saliency prediction methods and human fixation.

Method	Original images	Blurring effects	Monochrome effects	1NN	Wong et al.[133]	Ours
SIM [93]	0.08	0.13	0.26	0.28	0.14	0.18
RC [24]	0.11	0.14	0.18	0.23	0.22	0.25
FT [5]	0.13	0.26	0.34	0.28	0.24	0.37
Human fixation	0.07	0.28	0.16	0.27	0.11	0.30

of non-gray target-region images show the monochrome effect is a good attention retargeting method. Note that the HR of human fixation on [133] is even lower than the ones obtained from Monochrome and Blurring effects as aforementioned. In other words, Monochrome and Blurring effects have strong influence to human visual attention.

Regarding saliency retargeting evaluation, we compute HR on different saliency detection methods such as SIM, RC and FT. Table 3.1 also displays HR values computed across different saliency prediction methods. HRs obtained on the original images are still the lowest ones. Region-based saliency detection model like RC achieves higher HR than the pixel-based saliency detection models such as SIM. 1NN yields the best result on SIM. This is because SIM model is sensitive to color information, especially, color contrast [93]. Meanwhile, our method achieves the best performance on FT and RC. In short, the results demonstrate that our proposed method outperforms other methods in terms of visual saliency and in the meantime both spatial and color coherence is well preserved.

The cumulative score is defined as $CS(h) = N_{HR \geq h} \ / \ N \times 100\ \%$, where $N_{HR \geq h}$ is the number of testing images whose HR is not less than h, and N is the number of all testing images. Figure 3.11 depicts the cumulative scores from comparison methods. The CS curves of the original images and [133] drop rapidly, while blurring effects and our approach yield the similar CS curves. It is obvious and common sense that the resulting image from blurry method may lose useful overall information. Though the user is forced to focus on the selected area, she/he may (partially) lose the information in other areas, which is definitely not the expectation of saliency retargeting task. Also the resulting image from blurry

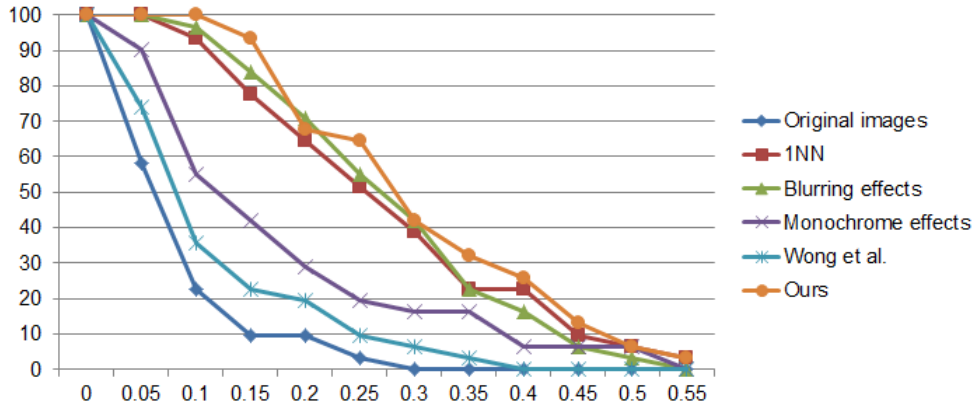


Figure 3.11: Comparison cumulative scores of different methods on *AdSaliency Dataset*.

method may be unnatural, e.g., the target region of the third example in Figure 3.8 is the dressing of the second woman from the left, and the blurred face makes the image unnatural as a whole. Therefore, we also take another important factor, *naturalness*, into consideration, which is introduced in the following subsection.

3.4.6 Naturalness Evaluation

To evaluate the naturalness of the transformed images, we conduct a user study by comparing our proposed method with three baseline methods: monochrome transform, blurring and 1NN. All participants from the earlier saliency retargeting evaluation joined this user study. The evaluators were requested to indicate their satisfaction with respect to the following perspectives:

1. Smoothness: How do you think about the smoothness of color alteration?
2. Experience: How does the color alteration help you experience the image?
3. Acceptance: Do you think the altered colors are reasonable?

For each sample, the participant rates each method on a 5-point scale from the best (5) to the worst (1). The image order is randomized. Figure 3.12 depicts

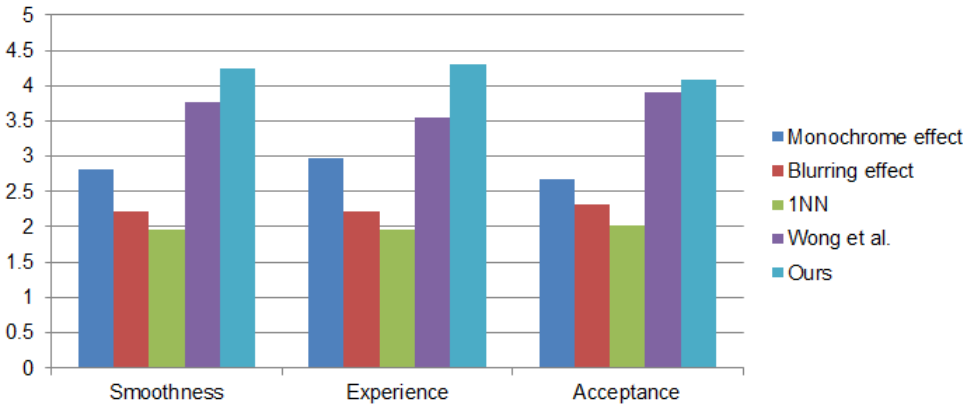


Figure 3.12: Results of naturalness evaluation on the *AdSaliency Dataset*. Our proposed method yields the best performance while 1NN performs the worst.

the results of naturalness evaluation. Generally, our method outperforms others in all aspects since it jointly optimizes the spatial and color coherence. Wong et al. method [133] has the second highest scores. 1NN yields the lowest score due to the heavy color inconsistency in the target region(s). We also observe that blurring is not a good transformation in terms of naturalness. This is because viewers are not willing to be forced to view specific regions, and instead they want to discover the image themselves. Monochrome effects offer the second best solution, which is usually used in advertisement to emphasize the target human(s) or object(s). However, this method sacrifices the rich appearance information brought by color. Overall, the comprehensive comparison results well demonstrate the effectiveness of the proposed method in terms of average fixation maps, hit rate, cumulative score and naturalness.

3.5 Discussion

In this chapter, we proposed a novel computational framework for image re-attentionizing task. Our work is based on a premise that human eyes tend to look at the unique area in the image in both global and local sense. The experiments demonstrated that the recolored images successfully attracted human attention to the target region(s) and in the meantime both spatial coherence and color coherence are well

preserved. Although the proposed method yields a better experience, it still has its limitations. The first is boundary artifact when selecting target regions from patches. To overcome this issue, interactive methods can be applied to provide better region selection [73], [91]. Another solution is to increase the number of superpixels in the image to provide finer over-segmentation. Another issue is the *unnatural* color for the objects which do not exist in the patch dataset. The remedy for this is to increase the dataset size.

Chapter 4

Depth Matters in Visual Saliency

In this chapter, we study the depth matters in visual saliency. We first introduce the construction of the new dataset, NUS3D-Saliency, for this problem. We then present some interesting observations. Finally, we propose using the depth priors in order to enhance the performance of existing saliency prediction methods.

4.1 Introduction

Human visual exploration and selection of specific regions for detailed processing is permitted by the visual attention mechanism [51]. The eyes remain nearly stationary during fixation events as humans look at details in selected locations, which makes eye movements a valuable proxy to understand human attention. Visual saliency refers to the preferential fixation on conspicuous or meaningful regions in a scene [116] that have also been shown to correspond with important objects and their relationships [102]. Visual saliency is thus crucial in determining human visual experience and also relevant to many applications, such as automatic image collection browsing and image cropping. Visual saliency has been extensively studied in signal processing, computer vision, machine learning, psychology and vision research literatures (e.g., [9, 18, 51, 42, 119]). However, most saliency models disregard the fact that human visual system operates in real 3D environments, while these models only investigate the cues from 2D images and the eye fixation data are captured in a 2D scene. However, stereoscopic contents provide additional depth

cues that are used by humans in the understanding of their surrounding and play an important role in visual attention [130]. Are the observer's fixations different when viewing 3D images compared to 2D images? How do the current state-of-the-art saliency models perform with additional depth cues? Not only are these questions interesting and important, answering them can also significantly benefit areas in computer vision research, such as autonomous mobile systems, 3D content surveillance and retrieval, advertising design, and adaptive image display on small devices.

In this chapter, we conduct a comparative and systematic study of visual saliency in 2D and 3D scenes. Whereas existing eye tracking datasets captured for 2D images contain hundreds of images, the largest available eye tracking dataset for 3D scenes contains only a limited size of 28 stereo images [54]. A comprehensive eye tracking dataset for 3D scenes is yet to be developed. Motivated by these limitations, we collect a larger eye fixation dataset for 2D-vs-3D scenes. A 3D camera with active infra-red illumination (Microsoft "Kinect"[3]) offers the capa-

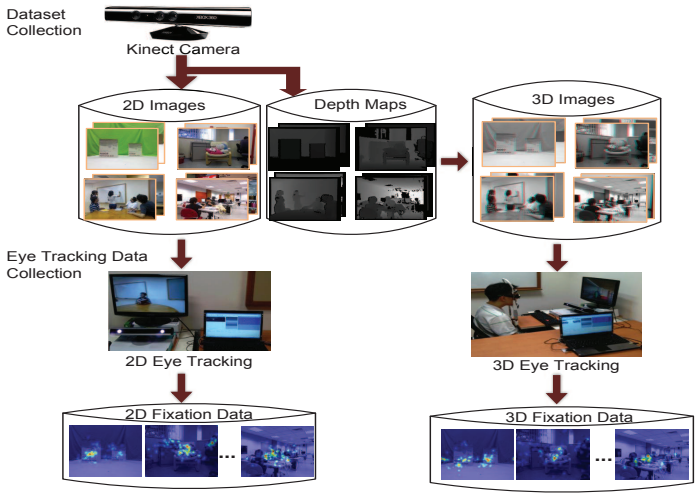


Figure 4.1: Flowchart on 2D-vs-3D fixation dataset construction. We collect eye-tracking data on both 2D and 3D viewing settings and each 2D or 3D image was viewed by at least 14 observers. Eye fixations are recorded for each observer. The final fixation maps are generated by averaging locations across all the observers' fixations.

bility to easily extract scene depth information in order to extend 2D stimulus to 3D versions. Using an eye tracker, we collect eye fixation data to create human fixation maps which represent where viewers actually look in 2D and 3D versions of each scene. Our work further aims at quantitatively assessing the contribution of depth cues in visual attention in 3D scenes and proposing depth priors to improve the performances of state-of-the-art saliency detection models. In summary, the contributions of our work mainly include:

1. An eye-fixation dataset is collected from a pool of 600 images and 80 participants in both 2D and 3D scenes.
2. We analyze and quantify the difference between 2D and 3D eye fixation data. Based on the observations, the novel depth priors are proposed and integrated into saliency detection models.
3. We comprehensively evaluate the performances of state-of-the-art saliency detection models augmented with proposed depth priors on 2D and 3D eye fixation data.

4.2 Literature Review

In order to understand what human attend to and qualitatively evaluate computational models, eye tracking data are used to create the human fixation maps, which will offer an excellent repository of ground truth for saliency model research. Most eye tracking datasets [18, 59, 21, 102] are constructed for 2D scenes and most saliency models only investigate the cues from 2D images or videos. In contrast, relatively few studies have investigated visual attention modeling on 3D contents.

Recently, several researchers have pioneered visual attention research on stereoscopic 3D contents. Jansen et al. [54] examined the influence of disparity on human behavior in visual inspection of 2D and 3D still images. They collected eye tracking data from 14 participants across 28 stereo images in a free-viewing task on 2D and

3D versions. A recent study [101] collected 21 video clips and the corresponding eye gaze data in both versions. However, compared with 2D eye tracking datasets, a comprehensive eye tracking dataset for 3D scenes is still absent. We believe that the studies on a rich and comprehensive 3D eye tracking dataset can offer interesting and important insights into how eye fixations are driven in real 3D environment. Motivated by this requirement, we collect a large 3D image database in this work, and then capture eye tracking data from an average of 14 participants per image across both 2D and 3D versions of 600 images.

Depth cues provide additional important information about contents in the visual field and can be regarded as relevant features for saliency detection [94]. Stereoscopic contents bring important additional binocular cues for enhancing human depth perception. Although there have been several efforts [52, 10, 34, 98] to include the depth channel into computational attention models, a major problem in extracting depth from stereo input is the computation time needed to process disparities. In this chapter, we study the discrepancies in human fixation data when viewing 2D and 3D scenes. The influence of depth on visual saliency is then studied and serves as the basis for learning depth priors to model 3D saliency.

4.3 Dataset Collection and Analysis

4.3.1 Dataset Collection

Our dataset aims to provide a comprehensive and diverse coverage of visual scenes for eye tracking analysis. We choose indoor and outdoor scenes that have natural co-occurrence of common objects. Furthermore, we systematically generate variants of scenes by varying parameters like depth ranges of different objects, number and size of objects and degree of interaction or activity depicted in the scene. We use the Kinect camera, which consists of an infra-red projector-camera pair as the depth camera that measures per pixel disparity, to capture a 640×480 pixel color image and its corresponding depth image at the same time. The dataset is named

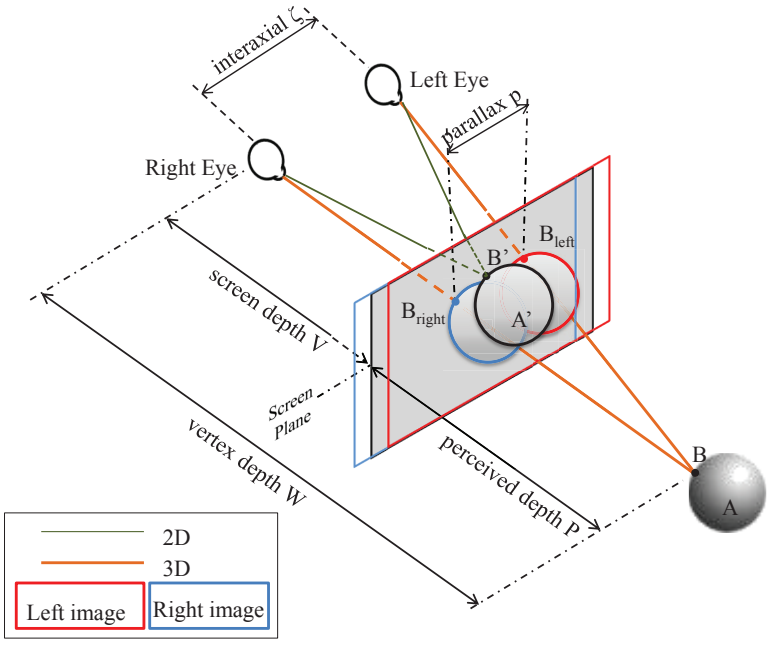


Figure 4.2: The relationship between 2D and 3D fixations. The fixation location captured from participant viewing at B is the same for both 2D and 3D experiment setups. Screen depth W is the distance from the participant to the screen, while perceive depth P is calculated based on the depth value.

NUS3D-Saliency dataset[1]. To the best of our knowledge, this is the largest 3D eye tracking dataset available to date for visual attention research, in terms of the total number of images and size of subject pool.

Stereoscopic image pair generation for 3D display Following the collection of the color and depth image pair, the next step is to create 3D stimulus which involves generating left and right images. Prior to generating left-right image pair, some pre-processing on the captured images are required.

Depth alignment and correction We first perform calibration on both depth and color cameras to find the transformation between their images in a similar way as [4]. Next, we over-segment the color image into superpixels [6]. Each

[1]The *NUS3D-Saliency* dataset is available at:
https://sites.google.com/site/vantam/nus3d-saliency-dataset

pixel, whose original depth value equal to 0, is assigned the average depth of the nearest neighbors in 8 directions in the same superpixel. Finally, we apply a default Laplacian filter with a 3×3 kernel for pixels whose depth values equal to 0 until all missing depth pixels are filled.

Stereoscopic image generation The stereoscopic image pair is produced by extracting parallax values from the smoothed depth map D and applying them to the left image I^l and right image I^r. For each pixel of the input color image I, the value of the parallax is obtained from its depth value. Figure 4.2 shows the relationship between 2D and 3D fixation. In both 2D and 3D viewing, for example, the fixation location for viewing B is recorded as the same position by the eye tracker. Considering the input image as a virtual central view, the left and right views are then obtained by shifting the input image pixels by a value ρ , $\rho = parallax/2$. In particular, the left image I^l and right image I^r can be obtained as $I^l(x_p^l) = I^r(x_p^r) = I(x_p)$, where x_p denotes the coordinate of the pixel in the color image I, the coordinate of the pixel in each view is calculated as

Figure 4.3: Exemplar data in our eye fixation dataset. From left to right columns: color image and raw depth map captured by Kinect camera, smoothed depth map, 2D fixation map, and 3D fixation map.

$x_p^l = x_p + \rho, x_p^r = x_p - \rho$. Following Figure 4.2, the *parallax* is calculated as follows:

$$parallax = \zeta \times (1 - \frac{V}{W}), \tag{4.1}$$

where ζ is the interaxial gap between two eyes, averaged as $60mm$, V is the screen depth or the distance from eyes to the screen and fixed as $80cm$ in our experiment setup, W is the vertex depth, equal to the summation of screen depth V and perceived depth P. For each pixel x in the image I, the perceived depth $P(x)$ can be calculated as $P(x) = D(x) \times \tau$, where τ ($\tau = 39.2$) is the ratio between the maximum depth distance captured by Kinect (10,000mm) and the maximum value in the depth image D (255). Since our dataset aims to provide the comprehensive and diverse coverage of visual scenes for eye tracking analysis, we reject images that are similar or have significantly overlapping content with other images in the dataset. Furthermore, images with significant artifacts after the smoothing process were rejected as well in an effort to minimize problematic images.

Participants The participants (students and staff members of a university) ranged from 20 to 33 years old (μ=24.3, σ=3.1), among them 26 females and 54 males with normal or corrected-to-normal vision. All participants are naive to the purpose of the study and sign consent forms for public distribution of recorded eye-fixation data.

Data collection procedure We use a block based design and free viewing paradigm. The subject views two blocks of 100 images that are unique and randomly chosen from the pool of 600 images, one of the blocks is entirely 2D and the other one entirely 3D. 3D images were viewed by using active shutter glasses on a 3D LCD display and 2D images were shown on the same screen in 2D display mode and the active shutter glasses switched off. In order to record subject eye gaze data, we used an infra-red based remote eye-tracker from SensoMotoric Instruments. The eye-tracker gives less than 1^o error on successful calibration. The eye tracker was calibrated for each participant using a 9-point calibration and validation method. Then images were presented in random order for 6 seconds followed by a gray mask for 3 seconds. We used a chin-and-forehead-rest to stabilize the participant's head position during each session.

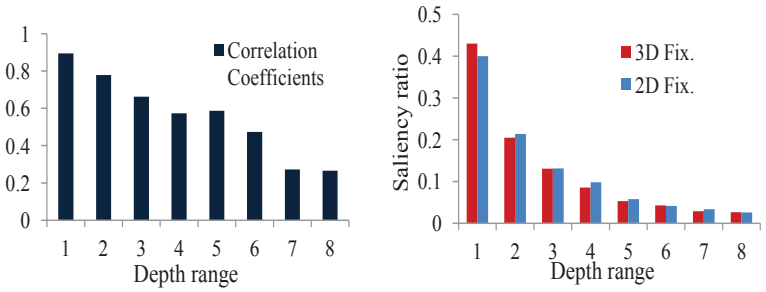
Human fixation maps Human fixation maps are constructed from the fixations of viewers for 2D and 3D images to globally represent the spatial distribution of human fixations. Similar to [123], in order to produce a continuous fixation map of an image, we convolve a Gaussian filter across all corresponding viewers's fixation locations. Six examples of 2D and 3D fixation maps are shown in Figure 4.3, the brighter pixels on the fixation maps denote the higher salience values.

4.3.2 Observations and Statistics

Using the recorded eye tracker data, we mainly investigate whether spatial distributions of fixations are different when human subjects view 3D images compared to 2D version. The interrelated observations are summarized as follows.

Observation 1: *Depth cues modulate visual saliency to a greater extent at farther depth ranges. Furthermore, humans fixate preferentially at closer depth ranges.*

In order to study the difference between 2D and 3D versions with respect to different depth range, fixation data for each 2D image I and its corresponding 3D



(a) (b)

Figure 4.4: (a) The correlation coefficients between 2D and 3D fixations in different depth ranges. We observe lower correlation coefficients for farther depth ranges.(b) Saliency ratio in different depth ranges for 2D and 3D scenes respectively. The participants fixate at closer depth ranges more often than farther depth ranges.

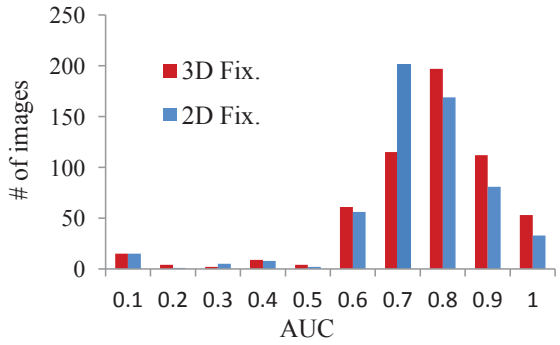


Figure 4.5: We examine the ability of 2D/3D fixation map to predict the labeled interesting objects and histogram of the AUC values for 2D and 3D fixation dataset are comparatively shown is blue and red colors, respectively.

image I', are divided into n ($n = 8$) depth ranges. Then for each depth range $r_b, b \in \{1, \cdots, n\}$, we compute the similarity between the 2D and 3D fixation distributions. We use the correlation coefficients (CC)[99] as similarity measure between two fixation maps. Figure 4.4(a) shows lower correlation coefficients for farther depth ranges.

Furthermore, in order to create a quantitative statistic of the relationship between the fixation distributions and depth ranges, we define saliency ratio as the description of fixation distribution. For the image I, we first compute saliency ratio $\gamma(r_b)$ as a function of the depth range, $\gamma(r_b) = \sum_x S(x)\delta(D(x) \in r_b) / \sum_x S(x)$, where $\delta(D(x) \in r_b)$ is set to 1 if x is in the depth range r_b. Figure 4.4(b) shows the saliency ratio vs. the depth range for 2D and 3D fixation data respectively. Looking at the data from the entire fixation dataset, the saliency ratio systematically decreases with the increase in depth range. From our analysis of fixation distribution and 2D-vs-3D correlation statistics over the entire dataset, we observe, (a) the larger discrepancy between 2D and 3D fixation data at further depth planes and, (b) the greater attenuation of visual attention at farther depth planes.

Observation 2: *A few interesting objects account for majority of the fixations and this behavior is consistent across both 2D and 3D.*

Such interesting objects such as human faces, body parts, text, cars and other

conspicuous objects are discussed in [59]. Other studies such as [89] have shown that the eye fixations are correlated to the locations of such objects. In order to analyze this relationship, we follow the method in [59] by manually labeling objects of interest. To form more object-like contours, annotation of such regions is done by over-segmentation using superpixels [6] for each color image in our dataset. Despite occupying only 7.6% of the image area on average, the area corresponding to interesting objects account for 54.2% and 51.2% of the fixation points for 3D and 2D respectively. To quantitatively measure how well a 2D/3D fixation map predicts interesting objects on a given 2D/3D image, we compute the AUC value [42], the area under receiver operating characteristic (ROC) curve for each image. We use the labeled objects of interest as ground truth along with the fixation map as the predicted map of the image, this method effectively marginalizes out the influence of depth planes and helps to understand the role of objects in isolation. Figure 4.5 shows the distribution of the AUC value for 600 labeled images. The average AUC for the entire 3D fixation dataset is 0.7399 and 0.7046 for 2D fixation dataset. Figure 4.6 gives examples of interesting objects along with corresponding fixation maps. These results suggest that the 2D and 3D fixation points show good correspondence with a few interesting objects.

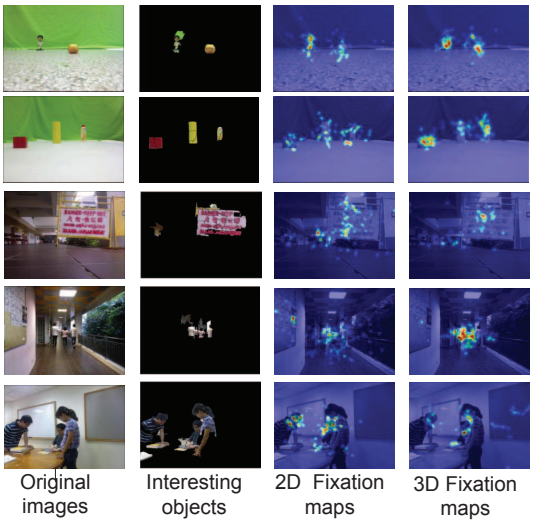


Figure 4.6: Exemplar interesting objects manually labeled and fixation maps for 2D and 3D images. It indicates that the participants frequently fixed on such areas.

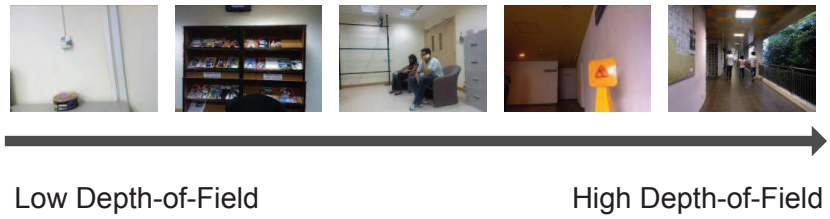


Figure 4.7: Examples with low and high depth-of-field values.

Observation 3: *The relationship between depth and saliency is non-linear and characteristic for low and high depth-of-field scenes.*

Importantly, we observe that strong correlation exists between the depth-of-field (DOF) of the image and the fixation distribution. The DOF value ℓ of the image I is inferred from the distance between farthest and nearest depth values. In this experiment, we assign depth values into n ($n = 21$) depth ranges. The ℓ is defined as $\ell = |\overline{h^s} - \overline{h^t}|$, where $\overline{h^s}$ and $\overline{h^t}$ denotes the mean of the depth value for the pixels in the nearest and farthest depth ranges. Figure 4.7 shows some examples corresponding to the different DOF values. To demonstrate the influence of DOF, we analyze the saliency ratio defined in Observation 1 on two subsets of 200 images each selected from our dataset, one low-DOF subset and one high-DOF subset.The low DOF and high DOF partitions have a significant overlap of object types and this effectively marginalizes out the influence of objects. Similar to the statistic described in Observation 2, we create the statistic of the relationship between saliency ratio γ and the depth range for these two image subsets respectively. As shown in Figure 4.8, the saliency ratio distribution on different depth ranges have noticeable discrepancies between low DOF and high DOF images, as well as the distribution is non-linear. We find that 2D(3D) low DOF and corresponding 2D(3D) high DOF saliency ratio distributions in Figure 4.8 are dissimilar at p = 0.05 using a non parametric Kolmogorov-Smirnov test, on the other hand, saliency ratio distribution for 2D low(high) DOF shows similarity to 3D low(high) DOF at $p = 0.05$. Motivated by this observation, we use the Gaussian Mixture Models (GMM) to model the distribution and the further implementation will be described in the next section.

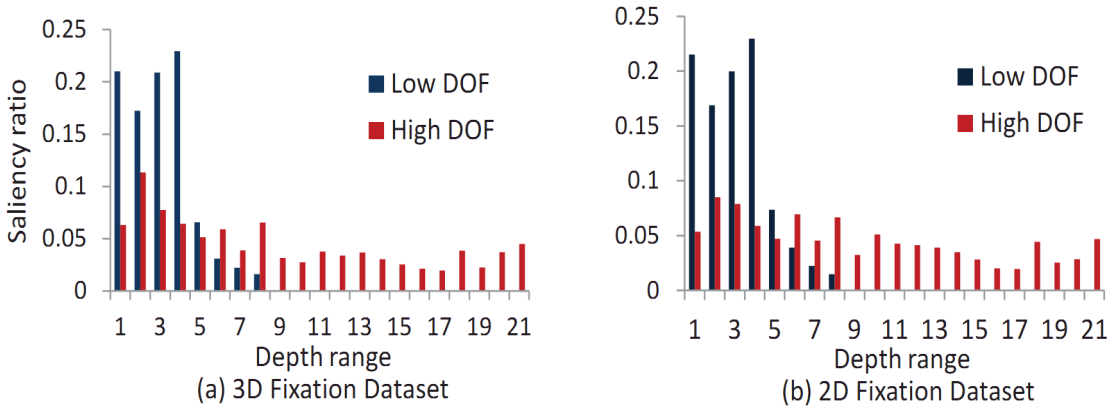


Figure 4.8: Saliency ratio as a function of depth range. The saliency ratio distribution for 200 lowest depth-of-field images and for 200 highest depth-of-field images calculated on (a) 3D and (b) 2D fixation dataset respectively. The plot indicates that depth-of-field has influence on the allocation of attention in both 2D and 3D images.

Table 4.1: The CC (correlation coefficient) comparison of fixation distribution on the 2D and 3D fixation data.

DOF	0-0.25	0.25-0.5	0.5-0.75	0.75-1	*Avg.*
CC	0.8066	0.5495	0.2721	0.3057	0.4835

Observation 4: *The additional depth information led to an increased difference of fixation distribution between 2D and 3D version, especially, when there are multiple salient stimuli located in different depth planes.*

In order to study the difference between 2D and 3D versions, we divide the image dataset into four groups according the DOF values and compute the correlation coefficients between two fixation maps for the four groups respectively. And Table 4.1 shows lower correlation coefficients for high DOF image groups. Figure 4.9 shows the fixation maps from the lower and higher depth-of-field images in 2D and 3D versions respectively.

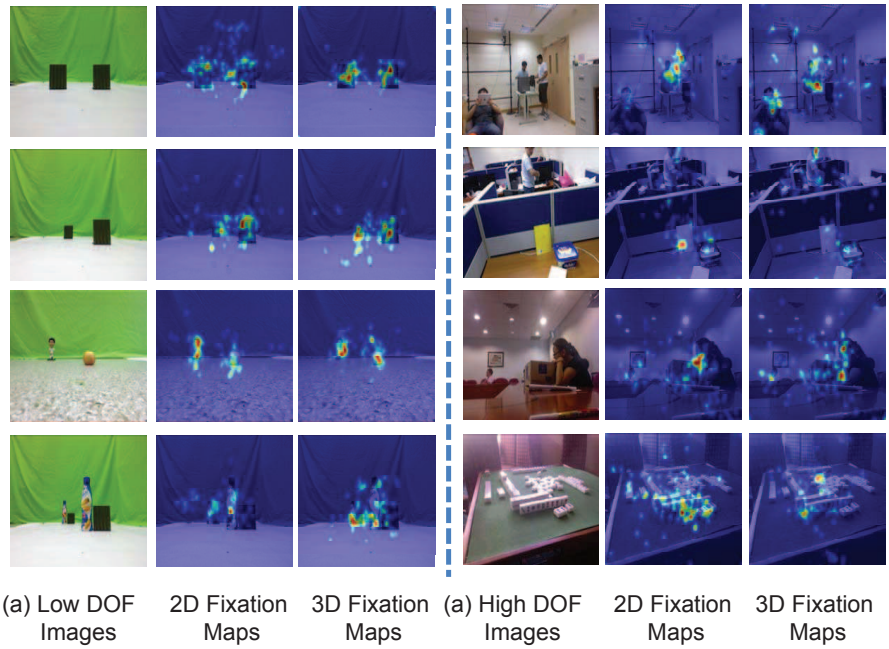


Figure 4.9: Fixation maps and fixation distributions for 2D and 3D images.The results indicate a clear difference between 2D and 3D fixation maps with the increased Depth-of-field of the images.

4.4 Saliency Detection with Depth Priors

Based upon the above observations, we seek the global-context depth priors in order to improve the performance of the state-of-the-art saliency detection models. In this section, we propose to model the relationship between depth and saliency by approximating the joint density with a Mixture of Gaussians. Note that we focus on bottom-up depth priors due to the difficulty in reliably detecting objects of interest in a scene.

4.4.1 Learning Depth Priors

Formally, let D and S represent the depth image and fixation map of the image I, respectively. And d and s denote N-dimensional vector formed by orderly concatenating the patches from a regular partition of the image D and S respectively, N

is the number of patches in the image. For the vector s, larger (smaller) magnitude implies that the patch is more salient (less salient). ℓ is the depth-of-field value introduced in the Section 4.3.2. The joint density between saliency response and depth distribution is written as

$$p(s, d|\ell) = \sum_{k=1}^{K} p(k)p(s|\ell, k)p(d|\ell, k), \tag{4.2}$$

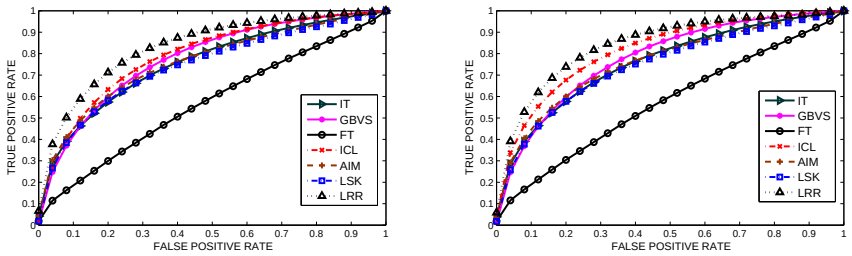
where k indicates the kth component of the GMM. From the joint distribution we calculate the conditional density required for the depth modulated saliency:

$$\begin{aligned} p(s|d, \ell) &= \frac{p(s, d|\ell)}{\sum_{k=1}^{K} p(k)p(d|\ell, k)} \\ &\propto \sum_{c=1}^{C} q_c(\ell) \sum_{k=1}^{K} \pi_k \mathcal{N}(s; \mu_k^c, \Lambda_k^c)\mathcal{N}(d; \nu_k^c, \Upsilon_k^c), \end{aligned} \tag{4.3}$$

where $q_c(\cdot)$ is a quantization function for the depth-of-field. We compute a C-bins histogram of depth-of-field on the whole dataset. If ℓ falls into the cth bin, $q_c(\ell) = 1$, otherwise, $q_c(\ell) = 0$. Finally, the conditional expected saliency of the test image I_t, with the depth vector d_t and depth-of-field ℓ_t, is the weighted sum of K linear regressors:

$$\begin{aligned} s_t &= \sum_{c=1}^{C} q_c(\ell_t) \frac{\sum_{k=1}^{K} \mu_k^c * w_k^c}{\sum_{k=1}^{K} w_k^c}, \\ w_k^c &= \pi_k \mathcal{N}(d_t; \nu_k^c, \Upsilon_k^c). \end{aligned} \tag{4.4}$$

The parameters of the model are obtained from the training dataset and the EM algorithm is applied for fitting Gaussian mixtures. For the image I_t, its corresponding depth saliency map can be defined as the predicted saliency s_t. Note that s_t has a non-linear dependency with respect to the image depth distribution d_t.



(a) 2D fixation data (b) 3D fixation data

Figure 4.10: ROC curves of different models. The results are from seven bottom-up saliency detection models to predict on the 2D and 3D fixation data individually.

4.4.2 Saliency Detection Augmented with Depth Priors

In order to investigate whether the depth priors are helpful for determining saliency, we extend seven existing methods to include the learned depth priors: Itti model (IT) [51], graph based visual saliency (GBVS) [42], frequency-tuned model (FT) [5], self-information (AIM) [18], incremental coding length (ICL) [46], local steering kernel (LSK) [109] and low-rank representation based model (LRR) [64]. The bottom-up saliency value predicted by the original models is denoted as $\psi(x)$. Note that we did not further report how to more elegantly integrate depth priors into the models themselves, and only do simple late fusion in this work. We will explore the methods along this direction in our further work. The final saliency can be achieved by simply using summation $\oplus$ or point-wise multiplication $\otimes$ as the fusion of two components. The final saliency is described by the equation:

$$S(x) = \psi(x)(\oplus/\otimes)p(s(x)|d(x), \ell). \tag{4.5}$$

4.5 Experiments and Results

In this section, we evaluate the saliency detection performance of the state-of-the-art models on our 2D and 3D fixation dataset. Furthermore, we quantitatively assess the effectiveness of the depth priors improving the performances of saliency prediction algorithms.

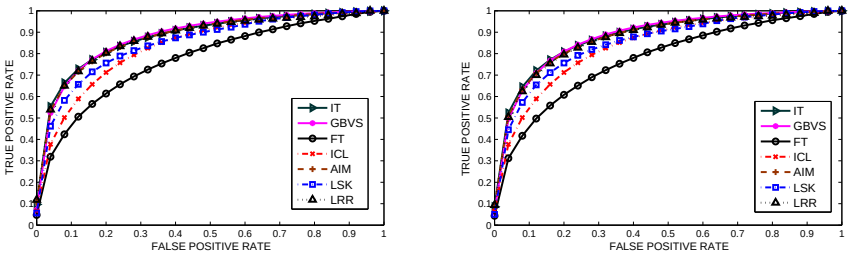
Table 4.2: The AUC and CC (correlation coefficient) comparison of different saliency models on the 2D and 3D eye fixation dataset.

Criteria	AUC		CC	
Method	2D Fix.	3D Fix.	2D Fix.	3D Fix.
IT	0.7270	0.7299	0.2706	0.2594
GBVS	0.7486	0.7506	0.2986	0.2844
FT	0.5707	0.5726	0.1402	0.1388
ICL	0.7676	0.7673	0.3095	0.2759
AIM	0.7293	0.7308	0.2868	0.2531
LSK	0.7142	0.7158	0.2658	0.2425
LRR	**0.8045**	**0.7971**	**0.3155**	**0.2975**
2D Fix.	--	0.7982	--	0.3797
3D Fix.	0.8156	--	0.3797	--

All saliency models use default parameter settings given by the corresponding authors. In order to learn depth priors, each image is resized to 200×200 pixels and regularly partitioned into 15×15 patches for training Gaussian models. The entire dataset is divided into four groups ($C = 4$) according to the depth-of-field value of each image. Depth saliency prediction is satisfactory with $K = 5$ as the number of the Gaussian components. For each image group, we randomly separate into 5 subsets, 4 subsets as the training set to learn the parameters of GMM and the remaining subset for testing. All the selected models are evaluated based on the following widely-used ROC and AUC. We also compute the correlation coefficients (CC) [99] between the fixation map and the predicted saliency map for evaluation.

4.5.1 Comparison of State-of-the-art Models

In this study, we examine the capability of seven bottom-up visual attention models to predict both the 2D and 3D fixation data in a free-viewing task. Figure 4.10 and Table 4.2 show the comparison results. First of all, most of models performed well for predicting human fixation when viewing 2D scenes. LRR exhibits stronger consistence with human eye fixations than the other models. We also evaluate the performance on 2D(3D) fixation maps to predict the 3D(2D) fixation maps. Interestingly, the AUC for the 2D fixation maps to predict 3D version is equal to



(a) 2D fixation dataset (b) 3D fixation dataset

Figure 4.11: ROC curves of different models. The results are from seven bottom-up saliency detection models by integrating depth priors to predict 2D and 3D fixation individually.

0.7982. On the contrary, the AUC of 0.8156 is obtained using 3D fixation maps to predict 2D fixation maps.

In contrast to 2D scenes, 3D scenes contain the additional information regarding the depth cue. This additional information could change the saliency of regions that are present in both 2D and 3D images. Here we show that the overall saliency is comparable in 2D and 3D scenes in terms of AUC. Thus, the bottom-up saliency models should also predict a fraction of the allocation of attention in 3D scenes. However, all models conducted on 2D scenes perform significantly better than 3D versions in terms of correlation coefficients. It further suggests that in a 3D attention model, depth could be considered as the important cue for saliency detection.

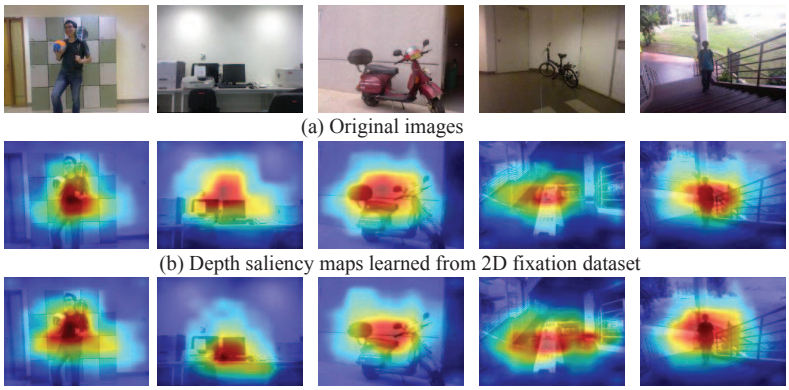
4.5.2 Depth Priors for Augmented Saliency Prediction

In this subsection, we assess the influence of the depth priors on saliency detection. To evaluate quantitatively the effectiveness of the proposed depth priors, the results of saliency models integrating depth priors are shown in Table 4.3. The ROC curves are illustrated in Figure 4.11. These results show that the models with predicted depth priors perform consistently better than those without such depth priors. Overall we observe a 6% to 7% increase in predictive power using depth based cues. Another important aspect brought out in Table 2 a multiplicative

Table 4.3: The AUC and CC (correlation coefficient) comparison of different saliency models with the depth priors on the 2D and 3D eye fixation dataset.

Criteria	Method	2D Fix. $\oplus$	2D Fix. $\otimes$	3D Fix. $\oplus$	3D Fix. $\otimes$
AUC	IT	0.8521	0.8536	0.8490	0.8539
	GBVS	0.8541	**0.8562**	0.8509	**0.8546**
	FT	0.7995	0.7458	0.7971	0.7449
	AIM	0.8502	0.8517	0.8495	0.8503
	ICL	0.8406	0.8088	0.8455	0.8077
	LSK	0.8496	0.8233	0.8453	0.8237
	LRR	0.8511	0.8495	0.8556	0.8463
CC	IT	0.4000	0.4202	0.3752	0.3977
	GBVS	0.4128	**0.4346**	0.3903	**0.4128**
	FT	0.3355	0.2804	0.3148	0.2680
	AIM	0.3651	0.4180	0.3419	0.3913
	ICL	0.4126	0.3704	0.3850	0.3248
	LSK	0.4064	0.3764	0.3793	0.3511
	LRR	0.4065	0.4085	0.3847	0.3953

modulating effect explains the influence of depth on saliency better than a linear weighted summation model, the latter has been used popularly in literature to combine results saliency maps derived from individual features [49].



(a) Original images

(b) Depth saliency maps learned from 2D fixation dataset

(c)Depth saliency maps learned from 3D fixation dataset

Figure 4.12: Representative examples in depth saliency prediction on 2D and 3D scenes respectively. The predicted depth saliency maps are similar between 2D and 3D versions due to the scenes with one conspicuous area/object clearly standing out from the others.

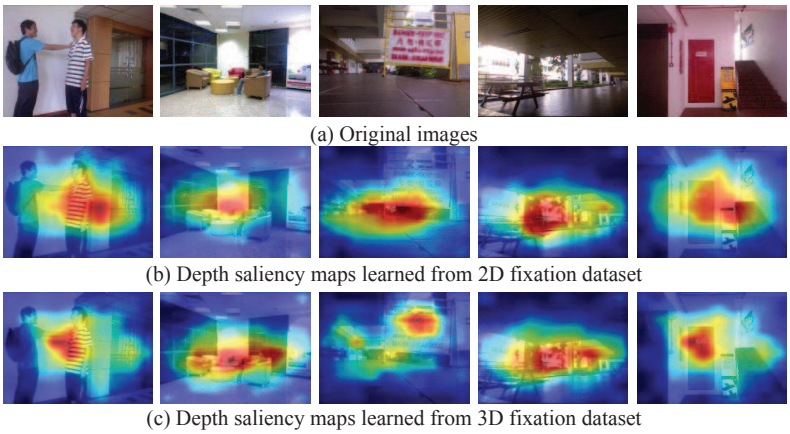


Figure 4.13: Representative examples in depth saliency prediction for 2D and 3D scenes respectively. The results show an obvious difference of the predicted depth saliency maps between 2D and 3D versions when multiply attractive objects or no conspicuous stimuli in the scenes.

Furthermore, Figure 4.12 and Figure 4.13 give some of the predicted saliency maps using depth priors alone (denoted as depth saliency map). As shown in Figure 4.12, the predicted depth saliency maps are similar in their spatial distribution between 2D and 3D versions when there is one conspicuous area or object clearly standing out from the others. On the other hand, when the scenes include multiple objects or no conspicuous objects, there is a noticeable difference between the predicted depth saliency maps of 2D and 3D cases.

Chapter 5

Static Saliency vs. Dynamic Saliency

In this chapter, we introduce the comparative study between the static saliency and dynamic saliency. Based on the observations on CMAO and Hollywood datasets, we propose a new computational model for video saliency prediction. We also utilize the video saliency into dynamic video captioning.

5.1 Introduction

The process of visual saliency has been the subject of numerous studies in psychology, neuroscience, computer vision and multimedia fields. Correspondingly, several computational models of saliency have been proposed in recent years [117, 42, 93, 139]. And many applications of automatic saliency detection have also been proposed such as image re-sizing [8], image automatic collage creation [126] and advertisement design [87]. Recently, other matters related to human attention such as depth information or attractiveness have also been explored [65, 95].

Although visual saliency has attracted the attention of researchers in the computer vision and multimedia fields for quite a long time, most of the visual saliency-related research works are conducted on still images. Video saliency receives much less research attention, though it is becoming more and more important along with

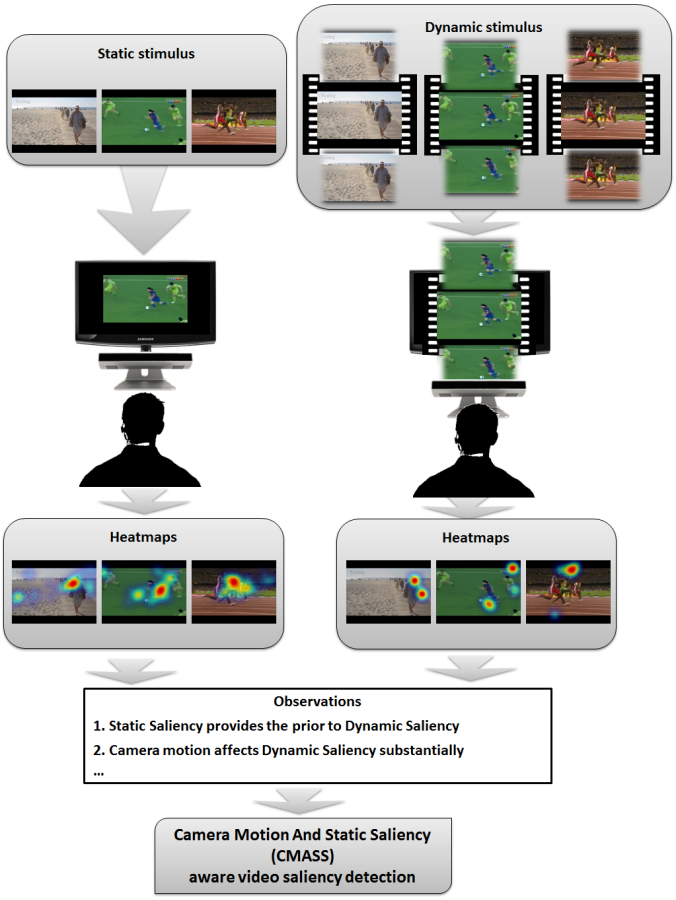


Figure 5.1: The comparative study of Static Saliency vs. Dynamic Saliency. We collect eye-tracking data on both static and dynamic viewing settings viewed by at least 10 observers. The CMASS framework is proposed to improve dynamic saliency detection.

the rapidly increasing demand of intelligent video processing. Moreover, in the existing works of video saliency [71, 11, 137], camera motions such as tilting, panning or zooming are disregarded during the saliency estimation. However, these camera motions ubiquitously exist in videos and may have great impacts on the saliency distribution, as experimentally validated in this work. Motivated by these two considerations, in this work, we conduct comprehensive comparison between the static saliency in still images and dynamic saliency in videos. Inspired by the observations in the comparison, we propose to utilize the static saliency as *a prior* information to improve the performance of dynamic saliency estimation in videos.

And we also investigate the role of camera motions in video saliency and integrate the estimated camera motion information into the saliency estimation procedure. Extensive experiments on two challenging benchmark datasets clearly show that our proposed saliency detection method outperforms the state-of-the-arts. Apart of proposing a novel method for saliency estimation, we introduce an interesting application of video saliency detection, *i.e.*, adaptive video subtitle insertion for assisting the patient with hearing impairment.

To facilitate the comparative study of static saliency vs. dynamic saliency, we first collect two video datasets for dynamic saliency estimation, namely the Hollywood and the Camera Motion (CAMO). Each of the two datasets contains the videos with camera motions. Then, volunteers are invited to participate the eye fixation map collection for these videos. Afterwards, the raw fixation data are converted to human fixation maps, which are considered as the groundtruth for saliency estimation. As aforementioned, in this work, we consider both the prior information from static saliency and camera motions in the video saliency. And we present a novel learning framework, called Camera Motion And Static Saliency (CMASS), to integrate the valuable information into the video saliency estimation. In particular, we train two neural networks which takes the camera motion parameters and position as inputs and outputs the optimal weights for the static saliency map and dynamic saliency map. In this way, the two available saliency maps can be adaptively fused to produce an improved dynamic saliency map estimation.

The proposed framework is shown in Figure 5.1, which includes the static and dynamic saliency detection for the same video and the fusion of the two detected saliency maps. The major contributions of this work can be summarized as follows,

1. To the best of our knowledge, we comprehensively conduct the first comparative study on the static saliency vs. dynamic saliency detection.

2. This is the first work to investigate the effects of camera motions in the dynamic saliency detection.

3. Inspired by the observed relationship between static and dynamic saliency, we propose a novel learning framework, *i.e.*, the CMASS method, for automatically fusing these two kinds of saliency maps to improve the performance of dynamic saliency detection.

4. We introduce a new and useful application for the dynamic saliency detection, namely adaptive video subtitle insertion for assisting people with hearing impairment.

5.2 Related Work

5.2.1 Learning to Predict Saliency

Through preliminary studies [19, 115], at early stages of free viewing, mainly bottom-up factors attract human attention (e.g., color, intensity, or orientation) and later on, top-down factors (e.g., humans, objects and interactions) guide eye movements. Some top-down factors in free-viewing are related to semantic factors. Elazary et al. suggested that interesting objects (annotations from LabelMe dataset [105]) direct human attention [32]. Einhauser et al. observed that objects are better predictors of fixations than bottom-up saliency [31]. Cerf et al. discovered that the meaningful objects such as faces and text attract human attention [21]. Judd et al., further showed that humans, faces, cars, text, and animals attract human gaze [59]. These interesting objects convey more information in a scene. During collecting NUSEF eye fixation dataset [102], Subramanian et al. found that fixations are focused on emotional and action stimuli.

Therefore, combining bottom-up and top-down factors may boost the existing models in order to better predict where human looks [60, 59, 100, 141]. The basic idea is that a weighted combination of features, where weights are learned from a large repository of eye movements over natural images, can enhance saliency detection compared with unadjusted combination of feature maps. [60], [59] and

[100] used image patches, a vector of several features at each pixel, and scene gist, respectively for learning saliency. Zhao et al. learned optimal weights for saliency channel combination separately for each eye-tracking dataset [141].

5.2.2 Saliency Prediction Models for Static and Dynamic Scenes

Visual attention analysis in static scenes has been long studied, while there is not much work on the dynamic scenes. In reality, we absorb the rich visual information that constantly changes due to dynamics of the world. Due to the large amount of information, visual selection is performed on both current scene saliency as well as the accumulated knowledge from chronological events. In the early works, few researchers have extended the spatial attention from static images to video sequences where motion plays an important role. Cheng et al. has incorporated the motion information in the attention model [25]. The motion attention model analyzes the magnitudes of image pixel motion in horizontal and vertical directions. Bioman et al. proposed a spatiotemporal irregularity detection in videos [11]. In this work, instead of reading motion features, textures of 2D and 3D video patches are compared with the training database to detect the abnormal actions present in the video. Le Meur et al. proposed a spatiotemporal model for visual attention detection [90]. Affine parameters were analyzed to produce the motion saliency map.

Recently, several researchers have studied modeling temporal effects on bottom-up saliency. Some methods fuse static and dynamic saliency maps to produce the final visual saliency maps (e.g., Li et al. [71] and Marat et al.[81]). A spatio-temporal attention modeling approach for videos is presented by combining motion contrast derived from the homography between two images and spatial contrast calculated from color histograms. Zhai et al. introduced a dynamic fusion technique is applied to combine the temporal and spatial models in order to achieve the spatiotemporal attention model [137]. The dynamic weights of the two individual models are controlled by the pseudo-variance of the temporal saliency values.

5.3 Fixation data collection

5.3.1 Data Collection

There are many datasets of still images (for studying static saliency) and videos (for studying dynamic saliency) [21, 59, 102]. However, none of the datasets can be used for studying the saliency for still images and videos simultaneously. Thus, in this work, we first construct two new datasets in order for studying these two kinds of saliency maps together.

Dataset Construction

To study the effects of camera motion in video saliency, we collect a new dataset named CAMO (Camera Motion) which consists of 120 videos of 6 different fundamental camera motions in cinematography: *dolly*, *zoom*, *trucking*, *tilt*, *pan*, and *pedestal motions*. Each video contains one single camera motion. Similar to the Hollywood dataset, we also randomly select one frame from each video for static saliency map collection. The information of each camera motions is listed as below.

- *Tilting*: the camera is stationary and rotates in a vertical plane.
- *Panning*: the camera is stationary and rotates in a horizontal plane.
- *Dolly*: the camera is mounted to the dolly and the camera operator and focus puller or camera assistant, usually ride on the dolly to operate the camera.
- *Trucking*: roughly synonymous with the dolly shot, but often defined more specifically as movement which stays a constant distance from the action, especially side-to-side movement.
- *Pedestal*: moving the camera position vertically with respect to the subject.

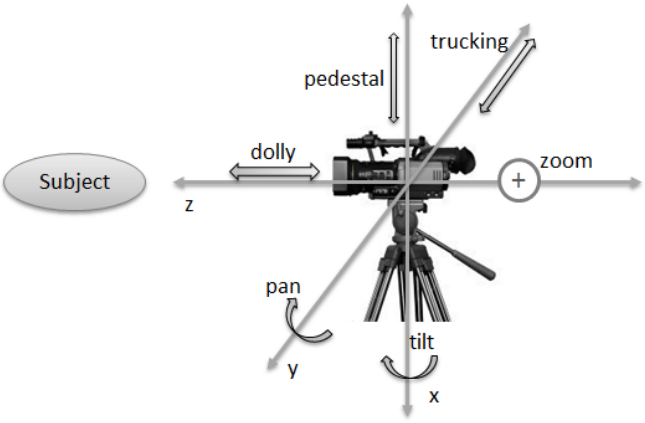


Figure 5.2: The fundamental camera motions in cinematography. Six basic types of motions are shown.

- *Zooming*: Technically this is not a camera move, but a change in the lens focal length with gives the illusion of moving the camera closer or further away.

Figure 5.2 illustrates six aforementioned camera motions. In the real world, many camera moves use a combination of these above mentioned techniques simultaneously. Therefore, we also collect another dataset named Hollywood. We select 500 random videos from Hollywood 2 dataset [82]. Hollywood 2 dataset consists of videos with natural human actions in diverse and realistic video settings. There exists one dataset collecting eye fixation on movies [84]. Therefore, we only collect fixation data on static images of that dataset. For each video, we extract one random frame which is not the shot boundary and stay close to the center frame of the video. The reason we select Hollywood 2 is that it contains realistic movies.

Human fixation data collection design

We invite 30 participants (students and staff members of a university), whose age ranged from 21 to 36 years old ($\mu = 26.9$, $\sigma = 3.1$), with normal or corrected-to-normal vision, to participate in the fixation map collection. All participants are

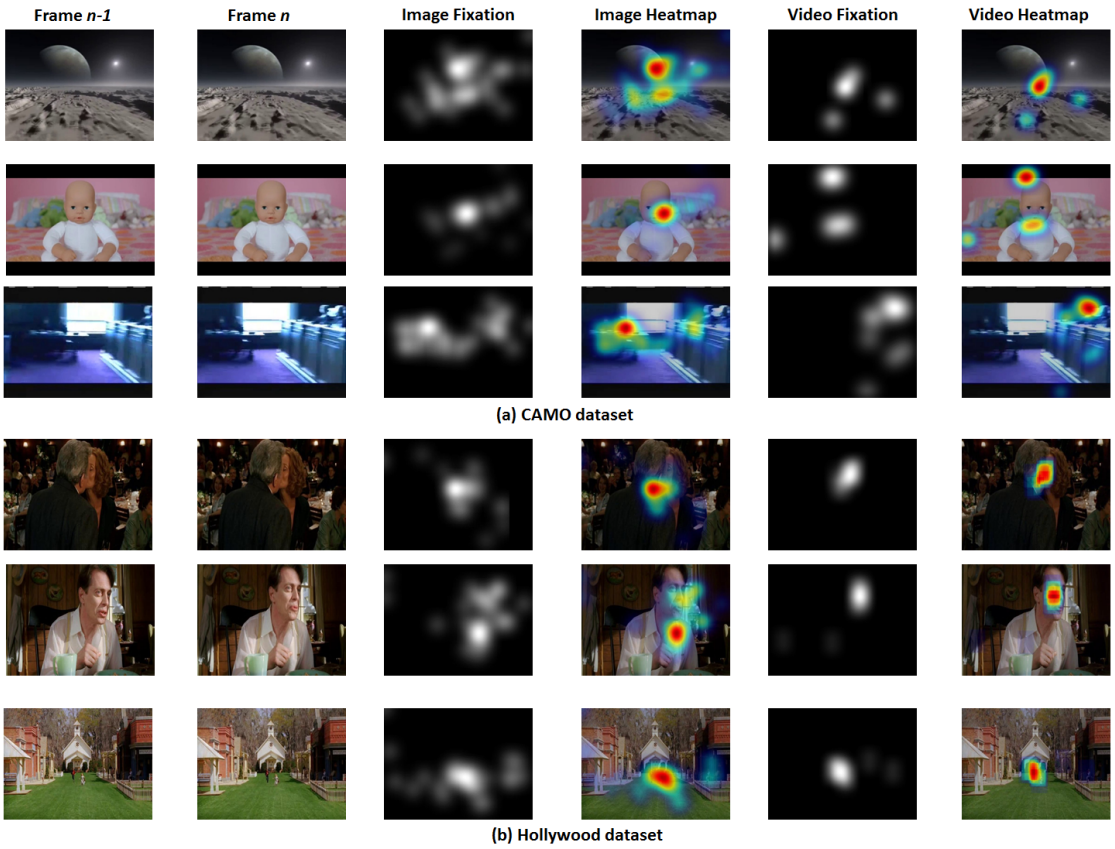


Figure 5.3: The exemplar images and their corresponding saliency maps and heat maps in CAMO and Hollywood datasets.

naive to the purpose of this study and have no prior exposure to experiments on vision. The participants have been split equally into three groups. Each group view only one of three following categories freely: *Hollywood static images*, *CAMO static images* and *CAMO videos*.

We use a block based design and free viewing paradigm. The subject views one of four designed blocks. In order to record subject eye gaze data, we used an infra-red based remote eye-tracker from SensoMotoric Instruments Gmbh. The eye-tracker gives less than 1^o error on successful calibration. The eye tracker was calibrated for each participant using a 9-point calibration and validation method. Then images were presented in random order for 6 seconds followed by a gray mask for 3 seconds.

Human fixation maps are constructed from the fixations of viewers to globally represent the spatial distribution of human fixations. Similar to [123], in order to produce a continuous fixation map of an image, we convolve a Gaussian filter across all corresponding viewers's fixation locations. Some examples of fixation maps of two new constructed datasets are shown in Figure 5.3, the brighter pixels on the fixation maps denote the higher salience values. These two datasets, CAMO and Hollywood are available at: https://sites.google.com/site/vantam/camo.

5.4 Observations

5.4.1 Camera Motion Effects

Using the recorded eye tracker data, we mainly investigate whether spatial distributions of fixations are different in static and dynamic settings. The key observations are summarized as follows.

1. The fixations of each video form a subset of the ones of the corresponding image if there is only single person or object. There are many explanations of this observation, such as, camera motion strongly narrows one's attention to certain parts of the scene, the viewer does not have enough time to examine all the details in the a moving scene, and also the accumulative knowledge of the previous scenes. Due to the temporal limitation in movie watching, the number of dynamic fixations is less than those of static fixations. This observation presents a close relationship between the static and dynamic saliency maps. We can use the static saliency map as a good prior to guide the dynamic saliency map estimation.

2. In some cases, for example, pedestal camera movement, the fixation lies on the anticipated direction, not on the objects. This observation shows the effect of camera motion on the dynamic saliency.

3. In the case that there are multiple persons or objects in the video, the fixations of the videos are not same as images. In other words, the fixations on videos and images focus on different people or objects.

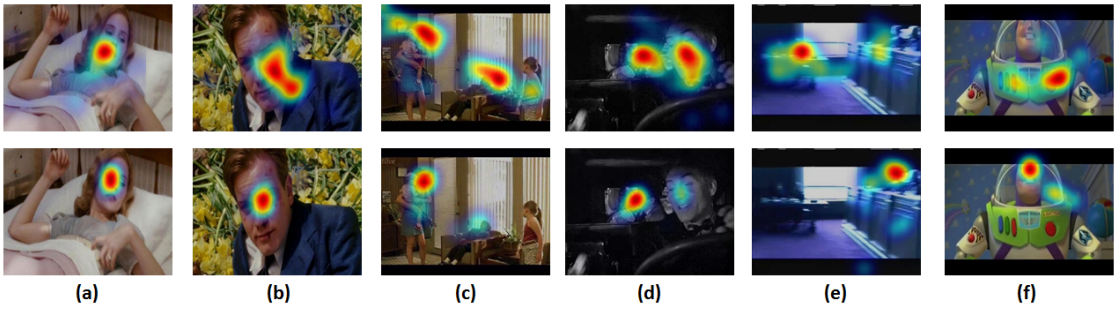


Figure 5.4: The observations of fixation data on the images (top row) and videos (bottom row). Note the difference of human fixations from column (c) to (f).

Some examples of the camera motion effects mentioned in Observation 2 are shown in Figure 5.4. The details of discrepancies of each camera motions are summarized as follows.

- *Pan*: The fixations may be either on the object of interest (*e.g.*, face of a walking person) or in the anticipated direction of the motion.
- *Pedestal*: The subject often tends to fixate on the anticipated direction of motion.
- *Tilt*: In case of a tilt shot, the subject also tends to fixate on the anticipated direction of motion.
- *Trucking*: Fixations in video are either a subset of the fixations in static images or they are in the anticipated direction of motion.
- *Dolly shot*: In the dolly shot, the cameraman is "moving closer" to the center or the object of focus. Therefore, the anticipated direction of motion can be considered to be the center or object of motion. While, the subject does fixate on the object of interest, it is not like the dolly shot which causes

the subject to fixate more/less on the object of interest. Thus, in case of dolly, the movement of the camera does not cause the subject to fixate on the anticipated direction of motion as "Pan", "Pedestal", or "Tilt".

- *Zoom*: We notice that the fixations are either on the object of interest or the peripheral motion of the camera.

5.4.2 Central Bias Investigation

We compute the average fixation maps to investigate the central bias of the fixation maps. Due to different sizes of testing images, the average fixation maps have cross-like shape. As can be seen in Figure 5.5, the center bias remains strong in the average fixation map of original images in the static part of both datasets. This agrees with the finding in [59]. The average map for video part of CAMO dataset is not center-biased due to the strong effects of the camera motions. Meanwhile, the average fixation map of Hollywood video is not so strong as the static version. In summary, the central bias is not significantly observed in video fixation.

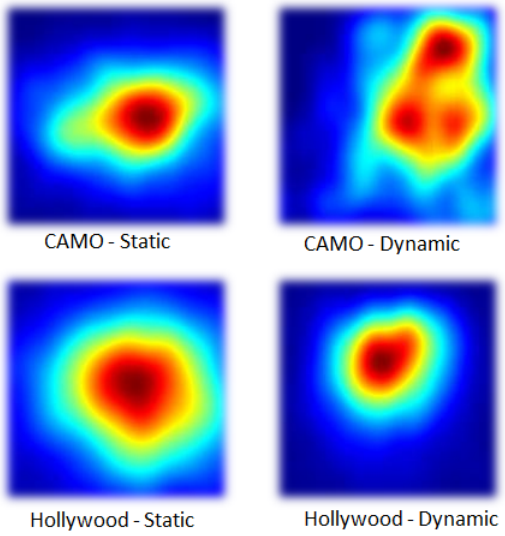


Figure 5.5: The average fixation static and dynamic maps from the two datasets. Warmer color indicates stronger fixation.

5.5 The proposed framework

In this section, we first explain the feature extraction applied on the given image or a certain frame in the video, followed by a novel framework which learns the mapping between image saliency and video saliency simultaneously.

5.5.1 Features

Static Features

To well describe the content of the images, we extract multiple static features and combine them together. The extracted features together describe both low-level appearance and high-level semantics. In particular, we use following low-level features: 13 local energy of the steerable pyramid filters in 4 orientations and 3 scales; 3 intensity, orientation, and color contrast channels (Red/Green and Blue/Yellow) as calculated by Itti and Koch's saliency method; 3 values of the red, green, and blue color channels as well as 3 features corresponding to probabilities of each of these color channels; 5 probabilities of above color channels as computed from 3D color histograms of the image filtered with a median filter at 6 different scales; 4 saliency maps of Torralba, SIM, SUN, and GBVS bottom-up saliency models. And we extract following high-level features: the horizontal line due to tendency of photographers to frame images and objects horizontally; person and car detectors implemented by Felzenszwalb's Deformable Part Model (DPM); face detector using the Viola and Jone's code.

Dynamic Features

In the temporal attention detection, saliency maps are often constructed by computing the motion contrast between image pixels. In this work, we generate dense saliency maps based on pixel-wise computations, mostly dense optical flow fields.

Here, we first resize each image/video frame to 200×200 pixels and then extract a set of features as aforementioned for every pixel.

5.5.2 CMASS for Dynamic Saliency Detection

Learning to Predict Image/ Video Saliency

In this subsection, we provide a simple linear regression based saliency estimation method. In the training phase, each training sample contains features at one pixel along with a $+1$ (salient) or -1 (non-salient) label. Positive samples are taken from the top p percent salient pixels of the human fixation map (smoothed by convolving with a Gaussian filter with window size $\sigma = 0.1$) and negative samples are taken from the bottom q percent. We chose samples from the top 20% and bottom 40% in order to have samples that were strongly positive and strongly negative. Training feature vectors are normalized to have zero mean and unit standard deviation. Assuming a linear relationship between feature vector f and saliency map s, we solve the following optimization problem to obtain the linear model W:

$$\min \|FW - S\|^2 + \lambda \|W\|^2,$$

where F and S are matrices by column-wisely stacking the vectors f and s of the training data. W is obtained in a closed-form manner, $W = (F^T F + \lambda I)^{-1} F^T S$. For a testing image, features are first extracted and then the learned mapping was applied to generate a vector which is later resized to a 200×200 saliency map.

CMASS Video Saliency Prediction

Inspired by the observations given in Section 5.4, we propose a novel learning-based method, *i.e.*, Camera Motion And Static Saliency (CMASS), to improve the performance of dynamic saliency prediction by utilizing the information from camera motion and static saliency results. Each frame in the videos is divided

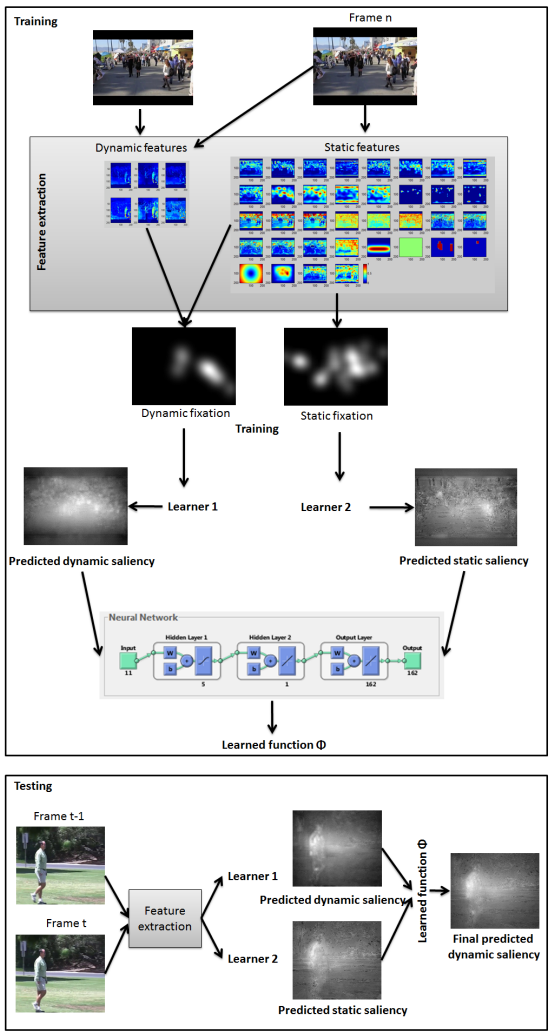


Figure 5.6: The learning framework. The upper panel shows the learning process, including the neural network parameters learning. The bottom panel shows the testing phase.

regularly into patches with the size of 9×9 pixels. For the jth patch in the training samples, let p_i^j denote the saliency map vector obtained from the image and p_v^j denote the saliency map vector from the video. The groundtruth saliency map for the jth patch is denoted as p^j. The camera motion parameter is denoted as CM^j. The position of the patch in the image is denoted as (x^j, y^j). According to the Observation 1, the generated saliency map is a weighted combination of the static and dynamic saliency maps. According to the Observation 2, camera

motion has great impact on the saliency map. For different patches, the camera motion and spatial position of the patches are different. Thus, the weights for their two kinds of saliency maps should be different. Based on these two considerations, in CMASS, we construct two neural networks to weight the static saliency map and dynamic saliency map in the final saliency estimation, respectively. The input to the neural network is the camera motion parameters and the position of the patches, and the output is the weight for the saliency map. The function of the neural networks are denoted as $\phi_i(CM^j, x^j, y^j)$ and $\phi_v(CM^j, x^j, y^j)$. And they are learned by minimizing the following loss function,

$$\mathcal{L}(\phi_i, \phi_v) = \sum_j \|\phi_i(CM^j, x^j, y^j)p_i^j + \phi_v(CM^j, x^j, y^j)p_v^j - p^j\|_2^2.$$

After learning the functions of the neural network ϕ_i and ϕ_v, we can directly obtain the saliency map for each new sample through

$$\tilde{p} = \phi_i(CM, x, y)p_i + \phi_v(CM, x, y)p_v,$$

where CM, x, y are the motion and position parameters for the input patch, and p_i and p_v are its two saliency maps.

However, directly training the neural network involves quite complicated optimization procedure, which damages the efficiency of the proposed method. In this work, we introduce two auxiliary variables w_i^j and w_v^j for ϕ_i^j and ϕ_v^j to simplify the optimization procedure. Then the loss function is formulated as:

$$\mathcal{L} = \sum_j \|w_i^j p_i^j + w_v^j p_v^j - p^j\|_2^2 + \lambda\{(\phi_i^j - w_i^j)^2 + (\phi_v^j - w_v^j)^2\}. \tag{5.1}$$

The above optimization problem can be solved by various methods. And the algorithm iteratively learns two phases within the same objective function. The solver is used for efficiency and outlined in Algorithm 1. Step 1 of the algorithm has closed form solution. Step 2 is solved via the optimization of the neural network. To ensure that the auxiliary variables approximate the original variables, the trade-off parameter λ will be increased in each iteration.

For the camera motions, we extract homography matrix of 15 frames with the selected frame is the middle one. To avoid the motion of the human or object, we use the work of [39]. Then, 8 values of the homography matrix (except the last element on the diagonal line) represent the camera motions.

For the implementation, we utilize 2 hidden layers with Transfer functions are 'tansig', and 'purelin', respectively. Backpropagation network training function is Levenberg-Marquardt. Note that we initialize NN of this current step by using the weights of its previous step. λ is set as 0.1 which takes the role of controlling the convergence speed. 40 to 50 iterations are required for convergence.

5.6 Evaluation

In this section, we describe the extensive experiments conducted on the new collected datasets for the better understanding about the performance of the proposed learning framework.

5.6.1 Learning to Predict Saliency

To quantitatively measure how well an individual saliency map predictors on a given frame, we compute the area under the receiver operating characteristic (ROC) curve (AUC) and linear correlation coefficient (CC) values. As the most popular measure in the community, ROC is used for the evaluation of a binary classifier system with a variable threshold (usually used to classify between two methods like saliency vs. random). Using this measure, the model is treated as a binary classifier on every pixel in the image; pixels with larger saliency values than a threshold are classified as fixated while the rest of the pixels are classified as non-fixated. Human fixations are then used as ground truth. By varying the threshold, the ROC curve is drawn as the false positive rate vs. true positive rate, and the area under this curve indicates how well the saliency map predicts actual human eye fixations. Meanwhile, CC

Algorithm 1 Solving Problem (5.1)

Input: Saliency map vectors p_i^j, p_v^j, p, parameters λ_0, $\rho = 1.5$.
Initialize: $t = 0$, $w_i^{j(t)} = 0.5$, $w_v^{j(t)} = 0.5$, $\lambda^{(t)} = \lambda_0$.
while not converged **do**

1. $t \leftarrow t + 1$

2. $f_i^{(t+1)} \leftarrow \phi_i(CM^j, x^j, y^j)$, $f_v^{(t)} \leftarrow \phi_v(CM^j, x^j, y^j)$

3. Update the auxiliary variables:

$$w_i^{j(t+1)} = (p_i^{jT} p_i^j + \lambda I)^{-1} \times \left(p_i^j p^{jT} - p_i^j p_v^{jT} w_v^j + \lambda(f_i^{(t)} + f_v^{(t)} - w_v^j)\right)$$

.

$$w_v^{j(t+1)} = (p_v^{jT} p_v^j + \lambda I)^{-1} \times \left(p_v^j p^{jT} - p_v^j p_i^{jT} w_v^j + \lambda(f_i^{(t)} + f_v^{(t)} - w_i^j)\right)$$

.

4. Train ϕ_i, ϕ_v.

5. Update parameters w_{ϕ_i} of ϕ_i, w_{ϕ_v} of ϕ_v.

6. $\lambda^{(t+1)} \leftarrow \rho\lambda^{(t)}$

end while
Output: The learned neural network ϕ_i and ϕ_v.

measures the strength of a linear relationship between human fixation map and predicted saliency map.

Table 5.1 shows the predicted results on our collected datasets, Hollywood and CAMO. We used static features to predict static saliency. Similarly, we used static features and dynamic features to predict dynamic saliency. The performance of dynamic saliency prediction is worse than the static saliency prediction based on static features only. That shows the need to improve the performance of dynamic saliency prediction.

Table 5.1: AUC and CC of saliency detection on the two datasets.

Dataset	AUC	CC
CAMO - images	0.74	0.52
CAMO - videos	0.64	0.20
Hollywood - images	0.71	0.45
Hollywood - videos	0.75	0.30

5.6.2 Dynamic Saliency Evaluation

We compare the performance of CMASS framework with the following four baseline methods:

1. Video saliency prediction from visual features [59].
2. Video saliency prediction from visual and motion features.
3. Fixed mapping weight to fuse static saliency and video saliency.
4. Adaptive mapping weight to fuse static saliency and video saliency [137].

Their performance comparison in terms of AUC and CC are shown in Table 5.2. As can be seen in Table 5.2, the results of the dynamic saliency prediction method from static information only are the worst in all cases. Combining the visual and motion features improves the performance generally across the two dataset. Fixed mapping weight is learned from a simple linear regressor. And the performance is further improved incrementally. The adaptive weight method of Zhai *et al.* achieves better performance than the rest of baselines, improving the performance over fixed mapping weight by around 2 to 3 percentage. Our proposed CMASS achieves the best performance in terms of AUC and CC for both datasets, Hollywood and CAMO. Generally it outperforms the results from Zhai *et al.* by 4 to 6 percentages. This improvement is rather significant. It shows the advantages of our combined static saliency and camera motion in boosting the performance of dynamic saliency prediction.

Table 5.2: Performance of CMASS on video saliency prediction on CAMO and Hollywood datasets.

Method	Hollywood		CAMO	
	AUC	CC	AUC	CC
Judd *et al.* [59]	0.72	0.25	0.61	0.18
Visual and motion feat.	0.75	0.30	0.64	0.20
Fixed mapping weight	0.74	0.28	0.63	0.19
Zhai *et al.* [137]	0.76	0.31	0.64	0.22
CMASS	**0.80**	**0.37**	**0.69**	**0.28**

5.7 Application to Video Captioning

Assisting the disabled persons by applying computer vision/multimedia techniques consistently attracts the attention from many researchers. Recently, a technique for assisting hearing impairment patients in watching videos [45] is developed, which automatically inserts the dialogue nearing the talking persons to help the patients understand who is talking and the content of the dialogue. However, there is often a need to insert the subtitle into the video without human appearance (*i.e.*, only narration appears in the video), such as documentary and introductory films. In this section, we introduce the new application which automatically insert the subtitle into such videos based on the video saliency map intelligently, in order to help the patients understand the content of the narration.

The basic criteria of the subtitle insertion are two-folds. Firstly, the selected position of the frame to insert the subtitle should have low saliency score. Otherwise, the inserted subtitle will overlap with the salient objects and worsen the watching experience of the audience. Second, the selected position should be near to the high saliency position. Thus the inserted subtitle will not distract the audience's attention.

The technique for the suitable position detection based on saliency map is introduced as follows. The predicted saliency map is first split into multiple blocks,

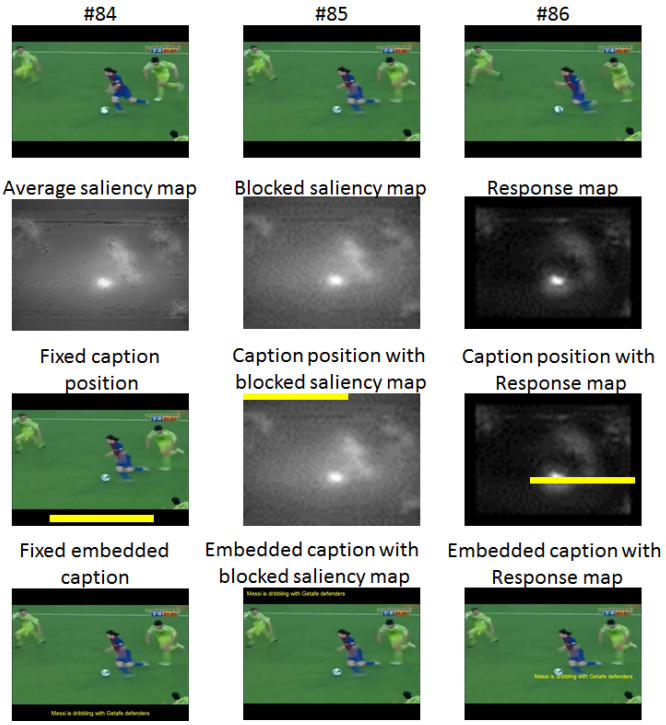


Figure 5.7: The usage of response map for inserting subtitles. The first row shows the frames of the video. The second row shows the saliency map from different saliency detection methods. The third row shows the found position for inserting the subtitles. And the last row shows the final results.

each of which having the size of 10×10 pixels. Each small block i has the mean saliency value s_i. We transform the saliency map to the response map for the use of determining the position of inserted subtitles. The response value of a certain pixel k in block i is computed as below.

$$r_k = \alpha_1 \sum_{j \in \mathcal{N}(i)} |s_i - s_j| - \alpha_2 s_i, \tag{5.2}$$

where $\mathcal{N}(i)$ represent the neighboring blocks of block i and s_i, s_j are the saliency values of the block i and block j respectively. r_k is the calculated response value for the kth pixel. The weights α_1, α_2 are empirically set as 0.5 and 0.5 throughout our implementation. The first term in the response calculation characterize the saliency contrast while the second term encourages to find the position with low saliency.

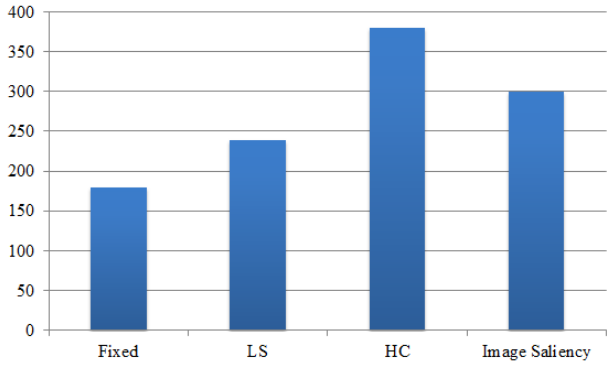


Figure 5.8: Results of evaluation of four methods in terms of the content comprehension. The compared methods include Fixed, Low Saliency Driven (LS) and High Contrast Drive (HC) and Static saliency detection based. The vertical axis represents the sum of the scores obtained by each group of participants. Higher score indicates better performance.

The size of inserted text will be calculated based on its length. Then we perform the exhaustive search on the response map in order to find the most suitable position with the largest response value. Figure 5.7 and 5.10 illustrate the examples of inserting subtitle into the documentary video without human appearance.

To evaluate the quality of the inserted subtitle and whether the watching experience is improved, we conduct the user study on both the content comprehensive and user impression. There are 24 users participating in the study. Their ages vary from 22 to 30 years old. We prepare 5 video clips with embedded caption for the evaluation.

For the content comprehension study, we randomly divide all the participants into four groups (each group has 6 participants) to avoid the repeated playing of a video which will cause knowledge accumulation. Therefore, each group merely evaluates one of the four paradigms for each video clip. We have designed five questions related to caption content comprehension. These questions are carefully designed to broadly cover the content in the video clips. The participant watches the clips under the task-free setting. Their results are the converted percentage of the correct answers. We compare the proposed high-contrast (HC) driven method with the following three methods. The first one is that the position of the subtitle

is fixed at the bottom of the frame. The second one is that the subtitles are inserted into the position with low saliency value, which is called low-saliency (LS) driven method. And the third one is also based on high-contrast but the saliency map is estimated from static saliency detection (Static). As shown in Figure 5.8, our method outperforms all of the other saliency detection methods for subtitle insertion. It demonstrates that the video saliency estimation method proposed in this work can find the most suitable position to insert the subtitle, where the subtitle is informative to the audience. In contrast, the image saliency based one performs worse since the inserted subtitle is not close enough to the salient regions. And the fixed caption performs worst as the audience cannot focus on the subtitle and video content at the same time.

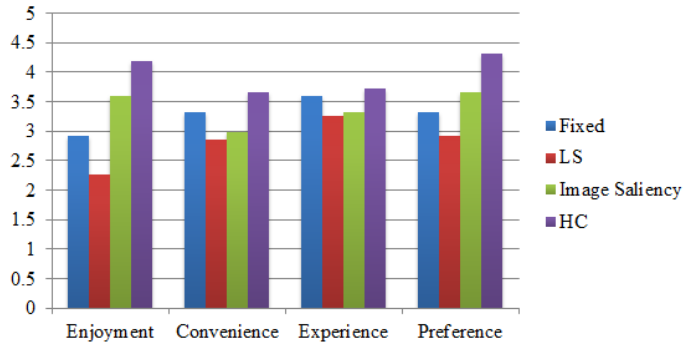


Figure 5.9: Results of evaluation on the user impression. Three methods are compared, namely Fixed, Low Saliency Driven (LS) and High Contrast Drive (HC). The methods are compared in terms of four criteria, namely Enjoyment, Convenience, Experience and Preference. Each user has been asked to assign a score between 1 (most unsatisfactory) and 5 (most satisfactory) for each criterion.

We further compare the four subtitle insertion schemes, *i*.e., fixed, LS, Image Saliency and HC, in terms of the user impression. We invite another 15 evaluators who are requested to indicate their satisfaction with respect to the following perspectives: 1) Enjoyment: How do you feel that the video is enjoyable? 2) Convenience: How do you feel the visual appearance of subtitle is convenient? 3) Preference: How do you prefer that captioning method? 4) Experience: How does the caption help you experience the video? For each sample, the participant rates each method on a 5-point scale from the best (5) to the worst (1)[74]. The video order is randomized. Figure 5.9 depicts the results of user impression evaluation.

Generally, our method outperforms others in all aspects since it optimizes allocated position. Low saliency driven captioning yields relatively low score due to uncommon appearance in the video frames.

5.8 Discussions

In this work, we have conducted comparative studying between the static saliency and dynamic saliency. To the best of our knowledge, this is the first research attempt to investigate this problem in depth. We first build the datasets of human fixation on both images and videos for the comparison purpose. Then we report several important observations of the relationship of static and dynamic saliency. Inspired by these observations, we propose the noval CMASS learning framework to fuse static saliency into dynamic saliency estimation to improve the video saliency prediction. Extensive experimental evaluations on the constructed datasets well demonstrate the effectiveness of the proposed method for video saliency prediction. We also apply the video saliency prediction method to the application of helping patients with hearing impairment in watching videos with narration. Suggested future work includes extensive user studies as a means to explore the potential of our approach under different conditions and for different application domains.

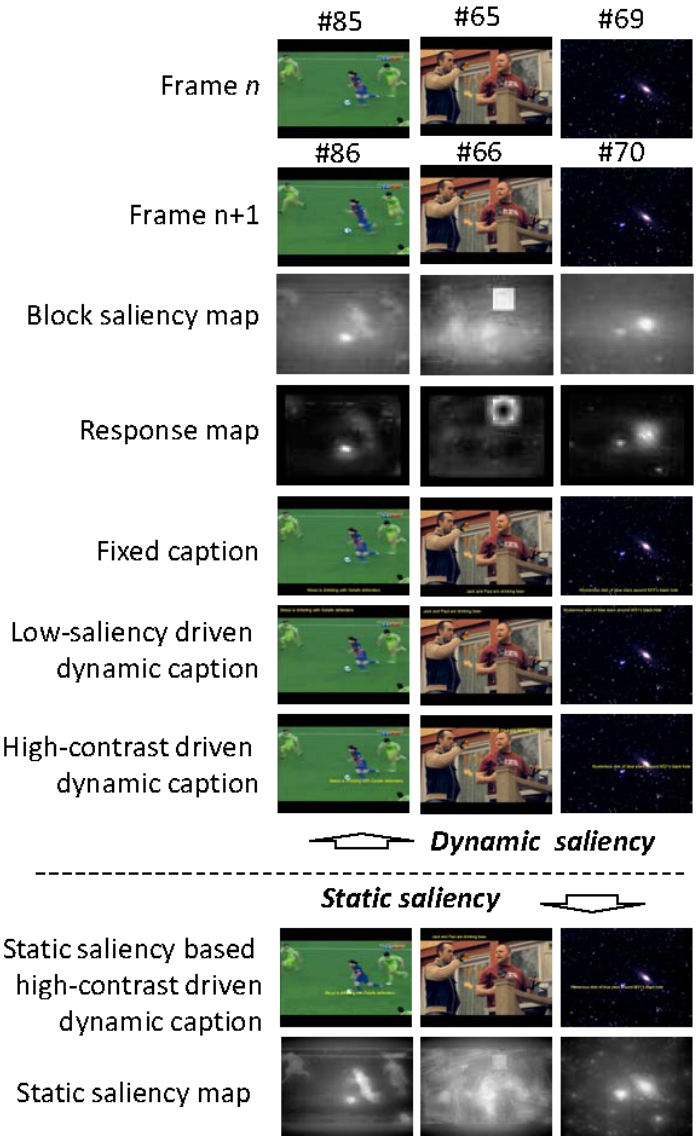


Figure 5.10: The examples of inserting subtitle into the documentary video. The original frames, the detected saliency maps, calculated response maps are shown from top to down. The top panel shows the result from the dynamic saliency detection. And the bottom panel shows the results from the static saliency detection.

Chapter 6

STAP: Spatial-Temporal Attention-aware Pooling for Action Recognition

---

In this chapter, we further investigate the impact of video saliency in human action recognition. We introduce the new saliency-guided pooling method which outperforms the state-of-the-arts.

6.1 Introduction

Recognizing human action in realistic videos has attracted much attention in computer vision community. Large-scale datasets, modern feature extraction methods and machine learning techniques are innovating this task. Improvements have been made using classifiers trained based on bag-of-words representations, which are computed from feature descriptors extracted at informative image locations [67, 83, 66, 113].

To some extent, action recognition in videos shares similar issues as object recognition in static images. Both tasks have to deal with significant intra-class variations, background clutter and occlusions. The conventional bag-of-words image classification framework was first adapted for action recognition in [66]. It pools

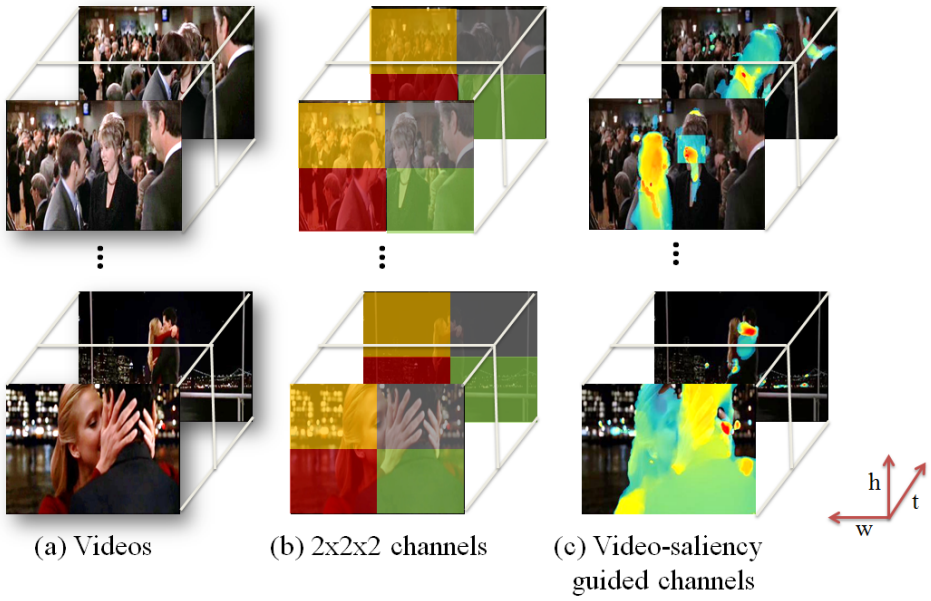


Figure 6.1: The illustration of the spatial-temporal attention-aware feature pooling for action recognition. The figure shows our work is superior over spatial pyramid matching due to the implicit background/foreground matchings. The local features are pooled according to (b) traditional SPM pooling with $2 \times 2 \times 2$ channels in spatial-temporal domain and (c) the proposed saliency-aware feature pooling with video saliency guided channels. For better viewing of all of the rest of figures in this thesis, please see original color pdf file.

all local features averagely to obtain a video representation. Extensional work also used pooling on spatial-temporal channels of video frames [67], which is similar to the Spatial Pyramid Matching (SPM) [68] used in the image classification. These approaches try to model global geometric correspondence by pooling video frame features to increasingly fine spatial channels. The success of SPM-based methods originates from the valid assumption that the videos with similar scene and geometry layout possibly belong to the same category.

However, we argue that the SPM-based representation may not be optimum for action recognition in human-centric videos. Most visual cues contributing to action recognition are not regularly located in certain spatial channel of the videos. As Figure 6.1 indicates, the spatial channels based on SPM may cause the misalignment problem due to different object locations and scene layouts. On the other

hand, better matching in the video representation can be achieved by respectively describing the video action/foreground area and the video scene/background area. For example, to recognize the human kissing action as shown in the last row of Figure 6.1, the information from the human face and the environment provides different cues and should be modeled separately.

To construct video representation that separates the video foreground and background information implicitly, we propose the Spatial-Temporal Attention-aware Pooling (STAP) framework. Inspired from the fact that actions/foregrounds attract human visual attention, we propose to utilize the video saliency to guide the construction of the video feature representation. In particular, we propose a new method to fuse the saliency maps from different saliency prediction models. This new saliency model borrows the prior knowledge from existing saliency models which often reveal some visual semantics, e.g., face, moving objects. By using such prior knowledge, we can construct a representation which matches the key objects implicitly. We then apply spatial-temporal feature pooling driven by the predicted video saliency maps to pool the video features. Besides the implicit foreground/object correspondence, the video backgrounds can also be better matched owing to the guidance of visual attention, since the visual saliency model can also predict the non-salient areas in the videos. The background context such as scene information is especially useful for certain action classification. Thus better recognition performance can be achieved by pooling based on the foreground and background separation in the classification process.

Our proposed spatial-temporal attention-aware feature pooling scheme is evaluated on three popular video action datasets and considerable performance improvements are achieved, specifically 62.5% on Hollywood2 (better by 4.2 %), 95.3% on UCF Sports (better by 0.3 %) and 87.9% on YouTube dataset (better by 3.7 %).

6.2 Related Work

In this section, we summarize the state-of-the-art works on video action recognition regarding the commonly used features, SPM-based pooling methods and the

integration of visual attention.

6.2.1 Feature Representations

Local descriptors computed around video interest points or on densely sampled patches are two popular methods for video representation as shown in Table 6.1. Interest point based local descriptors have been extended from images to videos. Laptev [66] introduced spatial-temporal interest points by extending the Harris detector. Other popular interest point detectors include detectors based on Gabor filters [16] and the determinant of the spatio-temporal Hessian matrix [128]. Meanwhile, Wang et al. [125] introduced an approach to model videos using densely sampled features. Since dense sampling has shown better performance [125], we also choose the dense sampling method in this work.

Among the existing descriptors for action recognition, the combination of HOG (Histograms of Oriented Gradients) and HOF (Histograms of Optical Flow) [66] has achieved excellent results on a variety of datasets [67]. HOG [28] focuses on the static appearance information, whereas HOF captures the local motion information. Dalal et al. [29] introduced MBH (Motion Boundary Histogram) to the problem of action recognition. The recent work [124] has demonstrated the effectiveness of this new feature. Trajectories are also used as a description to the interest point locations. Messing et al. [88] extracted feature trajectories by tracking Harris3D interest points [66] with the KLT tracker [79]. Sun et al. [113] extracted trajectories by matching SIFT descriptors between two consecutive frames. There is also a middle layer attribute description introduced by Liu et al. for the action recognition [76]. Recently, inspired by the Object Bank method [72], Sadanand et al. proposed to use action bank to explore how a large set of action detectors, which ultimately act like the bases of a high-dimensional "action-space", combined with a simple linear classifier can form the basis of a semantically-rich representation for action recognition and other video understanding challenges [107]. In this work, the prevalent HOG, HOF and MBH features are used.

6.2.2 Spatial Pyramid Matching based Pooling

Laptev et al. [67] presented a framework that classifies more than ten visual action classes. Their approach to video classification was inspired from the image recognition methods [13, 138] and extended the SPM [68] to the spatial-temporal pyramid. They used bag-of-words representation which computes the histogram of visual word occurrences over each particular spatial-temporal volume. This framework was also further followed in [83].

6.2.3 Visual Attention and Action Recognition

The problem of visual attention and the prediction of visual saliency have long been of interest in the human vision community [51, 18, 42, 47, 46]. Recently there has been a growing trend of training visual saliency models based on human fixations mostly in static images [59]. Jhuang et al. [55] proposed the model accounting only for part of the visual system, the dorsal stream of the visual cortex, where motion-sensitive feature detectors analyze visual inputs. Ullah et al. improved bag-of-words action recognition with non-local cues [120]. Recently Mathe et al. [84] explored the relationship between human visual attention and computer vision, with emphasis on action recognition in videos. However, they only introduced saliency as a criterion to select features for action recognition. In this chapter, we are interested in proposing new pooling-based computational model for recognizing actions. We are also inspired by recent works in image recognition [56, 23] which

Sampling method	**Without attention-aware pooling**	**With attention-aware pooling**
Interest points	Laptev et al. (HOG/HOF) [67] Marszalek et al. (HOG/HOF/SIFT) [83] Dense Trajectories (HOG/HOF/MBH/Traj.) [124] Klaser et al. (HOG3D) [61]	Ullah et al. (HOG/HOF) [120] Mathe et al. (HOG/HOF/MBH/Traj.) [84]
Dense sampling	Wang et al. (HOG3D/HOG/HOF) [125]	Our proposed method

Table 6.1: Where are we? The summary of related works of action recogntion in videos.

similarly utilize saliency as a cue. Table 6.1 locates our work along with the previous works in literature.

6.3 Spatial-Temporal Attention-aware Pooling for Action Recognition

In this section, we introduce the framework using the video saliency information for a saliency-aware feature pooling. Figure 6.2 depicts the flowchart of our proposed framework.

The proposed Spatial-Temporal Attention-aware Pooling (STAP) procedure aims to pool video local descriptors into channels using the predicted video saliency maps. As we have utilized the video saliency predictor described in Chapter 5, the pooling procedure is introduced as below.

Given a video $\vartheta = \{v_i, i = 1, \cdots, m\}$ with m frames and $S = \{S_i\}$ as their saliency maps, the local descriptors $X = \{x_j\}$ can be extracted densely from the frame patches. Note that one video frame is divided into overlapping patches. Each

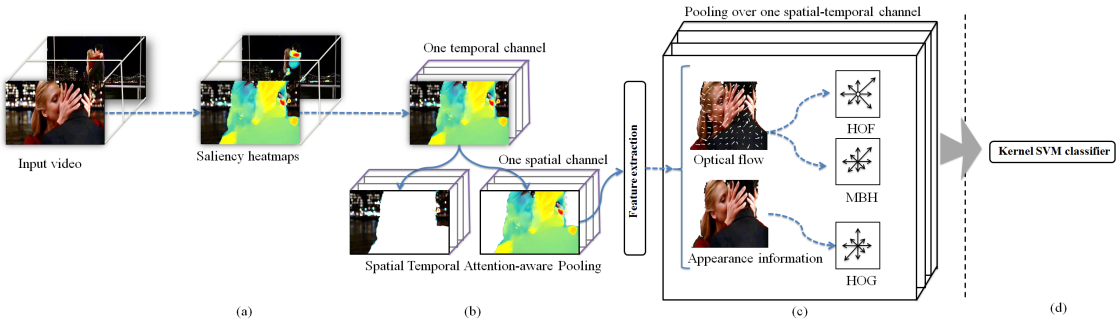


Figure 6.2: The flowchart of the proposed framework for action recognition in videos. (a) The saliency maps are predicted from the input video frames. (b) The local features are clustered to different channels according to the video saliency information. (c) The feature pooling is then operated on each channel to form a representation of the video. (d) Finally, Kernel SVM is used for action classification.

patch's parameters include spatial patch size $W_s = 15$ and overlapping size $O = 5$. We denote each patch's descriptor as $x = \{\mathbf{d}_x, s_x\}$ where $\mathbf{d}_x$ is a sparse histogram vector which is the result of the projection of the raw descriptors onto the codebook elements (the value of the closest entry's index is one, and the rest is zero), and s_x is the patch's saliency value. We compute s_x by averaging the saliency values within the patch area.

Unlike traditional SPM [68] where the descriptor is assigned to the corresponding spatial channel based on its location, we utilize the saliency-guided descriptor grouping for spatial domain. Denote L as the number of spatial layers, the total number of spatial channels is $2^L - 1$. For l-th layer, video descriptors are grouped to 2^{l-1} channels according to threshold values $\theta_l = \{\frac{1}{2^{l-1}}, \frac{2}{2^{l-1}}, \cdots, \frac{2^{l-1}}{2^{l-1}}\}$. Based on their s_x values, the local descriptors are assigned to the corresponding spatial channel. Thus the attention-aware spatial channels of descriptor x of all L layers can be defined as:

$$G_a(x) \subset \{1, 2, \cdots, 2^L - 1\}, \tag{6.1}$$

where $G_a(x)$ denotes the set of attention-aware channels that x belongs to. Note that each descriptor may belong to multiple spatial channels.

Similar to spatial domain, the video frames are divided into T temporal layers and the temporal channel of each descriptor x is denoted as:

$$G_t(x) \subset \{1, 2, \cdots, 2^T - 1\}. \tag{6.2}$$

Then the visual descriptors belonging to the a-th attention-aware channel and t-th temporal channel are pooled to produce the descriptor $\mathbf{f}_{at}$:

$$\mathbf{f}_{at} = \frac{\sum_{x|l \in G_a(x), t \in G_t(x)} \mathbf{d}_x}{\sum_{x|l \in G_a(x), t \in G_t(x)} 1}. \tag{6.3}$$

For classification, we use a non-linear kernel support vector machine (SVM)

[121] with the kernels defined as

$$K(\vartheta_1, \vartheta_2) = \exp\big(-\sum_{a=1}^{2^L-1}\sum_{t=1}^{2^T-1}\frac{1}{A_{at}}D(\mathbf{f}_{at}^1, \mathbf{f}_{at}^2)\big). \tag{6.4}$$

Obviously, STAP pooling will fall back to a standard bag-of-words model when $L = 1$, $T = 1$. Similar to [124, 120, 125], we choose D as a χ^2 distance function and A_{at} is the mean value of χ^2 distances among the training samples for the at-th channel.

6.4 Implementation Details

In this section, we introduce the implementation details used in our STAP framework including the feature extraction and model learning for action recognition.

6.4.1 Video Representation

In this work, we compute the dense combination of HOG, HOF and MBH [66, 29] as the local feature description. For both HOG and HOF, orientations are quantized into 8 bins using full orientations, with an additional zero-motion bin for HOF, namely, 9 bins in total. Both descriptors are normalized with their ℓ_2 norm. For MBH, we obtain an 8-bin histogram for the horizontal and vertical components of the optical flow and normalize them separately with the ℓ_2 norm. For both HOF and MBH descriptors, we reuse the dense optical flow that is already computed to extract motion magnitude images for saliency prediction [17].

Dense local feature sampling is used in this work. Two parameters related to the dense local feature sampling, the temporal and spatial sampling size W_t and W_s, are investigated. These two parameters denote the sampling duration and

patch size while computing each local feature. Larger sampling size generates less but smoother local features in videos.

For bag-of-words (BoW) representation, we construct a codebook for each descriptor type (HOG, HOF, and MBH) separately and use Vector Quantization to build the BoW representation. To limit the complexity, we train the codebooks by clustering on $500,000$ randomly selected training features using k-means implemented in [122]. The size of the codebooks C is further investigated in our work.

Finally for the pooling parameters defined in Section 6.3, the spatial layer number L is set to 2 for all experiments in this work in order to limit the computational complexity. But the temporal layer number T is further evaluated as shown in the experiments.

6.4.2 Learning with Kernel SVM

We use Kernel SVM to learn action classifiers. To build a multi-class classifier, we combine binary classifiers using one-against-all strategy. Note, however, that in our setup all problems are binary, i.e., we recognize each class independently and concurrent presence of multiple class labels (namely multiple actions) is allowed.

6.5 Experiments

6.5.1 Datasets and Evaluation Metrics

We systematically evaluate the effectiveness of the proposed STAP method on three realistic human action datasets: Hollywood2, UCF Sports and YouTube. These three databases are chosen for evaluation because they exhibit the difficulties in recognizing human actions, in contrast to the controlled settings in other related

databases. Figure 6.3 depicts some exemplar frames of the datasets utilized for the evaluation.

Hollywood2 dataset has been collected from 69 different Hollywood movies. There are 12 action classes: answering the phone, driving car, eating, fighting, getting out of the car, hand shaking, hugging, kissing, running, sitting down, sitting up, and standing up. In total, there are 1,707 action samples divided into a training set (823 sequences) and a test set (884 sequences).

UCF Sports Action dataset contains ten different types of human actions: golf swinging-bench, diving, kicking a ball, weight-lifting, horse-riding, running, skateboarding, swinging-side, golf swinging and walking. The dataset consists of 150 video samples which show large intra-class variabilities. We use a leave-one-out cross-validation strategy similar as in [125, 124].

YouTube dataset [77] contains 11 action categories: basketball shooting, biking/cycling, diving, golf swinging, horse back riding, soccer juggling, swinging, tennis swinging, trampoline jumping, volleyball spiking, and walking with a dog. This dataset is challenging due to large variations in camera motion, object appearance and pose, object scale, viewpoint, cluttered background and illumination

Figure 6.3: Exemplary frames from video sequences of UCF Sports (top row), Hollywood2 (middle row), and YouTube (bottom row) human action datasets.

conditions. The dataset contains a total of 1,168 sequences. We follow the original setup [77] using leave-one-out cross validation for a pre-defined set of 25 folds.

For Hollywood2 dataset, to compare the overall system performance, we compute a mean average precision (mAP) over a set of binary classification problems as in [83, 125]. For UCF Sports and YouTube datasets, average accuracy over all classes is reported as performance measurement as in [125, 77]. Since different videos are in different resolutions (e.g., UCF Sports contains some videos with high resolution, 720×576 pixels), we resize all videos to the same size, namely, 640×480 for UCF Sports dataset (smaller scale), 320×240 for Hollywood2 and YouTube datasets (larger scale) to control the computational complexity.

6.5.2 Performance of Saliency Prediction

We compare our predicted saliency maps with other predicted saliency baselines. We take $1,000$ random frames for the test set from Hollywood2 fixation dataset [84]. We compare our approach with SIM, SUN, LSK, GBVS, Cerf et al., Motion map, and Central bias map. We first evaluate our saliency predictors under the AUC metric, which interprets saliency maps as predictors for separating fixated pixels from the rest. We also report the Correlation Coefficient (CC) for comparing predicted saliency maps to the human ground truth. As shown in Table 6.2, combining predictors improves performance under those metrics, whereas bottom up

Saliency model	AUC	CC	STAP on UCF Sports
SIM [93]	0.71	0.10	93.3
LSK [109]	0.68	0.11	91.7
GBVS [42]	0.76	0.26	92.3
SUN [139]	0.69	0.12	92.7
Cerf et al. [21]	0.66	0.18	93.0
Motion map [17]	0.65	0.14	91.7
Central bias map [59]	0.81	0.21	92.7
Ours	**0.87**	**0.29**	**95.3**

Table 6.2: Evaluation of saliency prediction models.

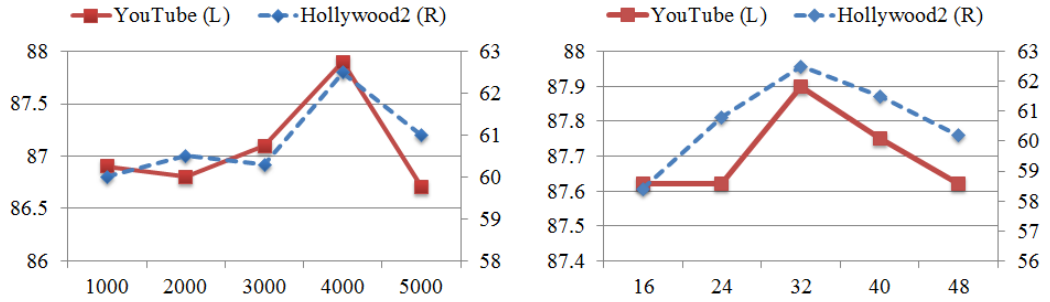
saliency maps are better predictors than top down ones. GBVS achieves the highest AUC among static image saliency models. The central bias has a high value of AUC which shows the bias in cinematography. Meanwhile, motion shows the lower performance compared with ones of other saliency models. Our top performance illustrates the significant advantage in combining various kinds of saliency information.

6.5.3 Evaluation of Parameter Settings

To evaluate the different parameter settings for STAP, we report results on two larger and more challenging datasets, YouTube and Hollywood2. We study the impact of the codebook size, sampling spatial size, temporal size and temporal layer number.

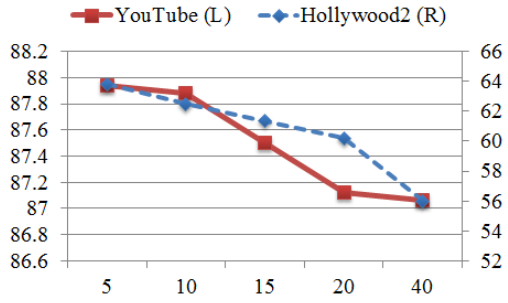
As shown in Figure 6.4, the performance degrades when the codebook size is too small (i.e., 1000, 2000) or too large (i.e, 5000). The best performance is achieved with codebook size $C = 4000$. It agrees with the finding in previous works [67, 124]. Similarly, too small or too large sampling sizes also cause the performance decrease. The best performance is obtained when the spatial sampling size W_s is 32. Regarding the temporal sampling size W_t, the performance decreases when W_t increases. However, the smaller sampling size costs more memory for feature extraction. Therefore, 10 is acceptable in terms of performance and memory storage since there is only a minor gain compared to $W_t = 5$. For the temporal layer number T, the best performance is when $T = 2$ and the performance degrades when we increase T to 3 or 4. We observe the similar patterns on both datasets throughout the parameters setting experiments.

Finally, the parameters for our STAP are fixed to the following values, i.e., $C = 4000$, $W_s = 32$, $W_t = 10$, $T = 2$.

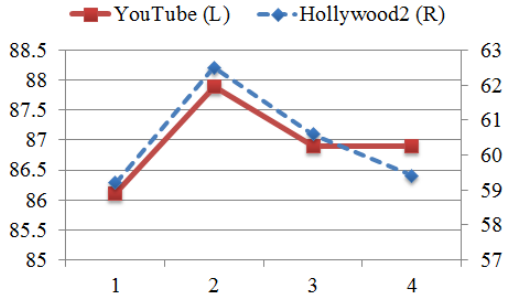


(a) Codebook size C

(b) Spatial sampling size W_s



(c) Temporal sampling size W_t



(d) Temporal layer number T

Figure 6.4: Results for different parameter settings on Hollywood2 and YouTube datasets (left Y axis is for YouTube, whereas right Y axis is for Hollywood2).

6.5.4 Comparison with the State-of-the-arts

Table 6.3 compares our results with the state-of-the-arts. In all three datasets, STAP outperforms all known methods in the literature, and in some cases by a significant margin. On UCF Sports dataset, STAP outperforms the state-of-the-art performance [134] by 4%. In Figure 6.5, we show the confusion matrix of human action recognition on UCF Sports. We achieve 100% accuracy on 7 out of 10 classes. We are aware of the small gap between our work and Action Bank [107] on UCF Sports (0.3%). These two works, however, have two different approaches. The target of our work is to improve the pooling of local features and hence can even further boost the performance of [107] with better recognition to each elemental action. For YouTube dataset, our framework outperforms the current state-of-the-art method [124] by 3.7%. The performance of the proposed framework on

Hollywood2 is 62.5% which is an improvement of 4.3% over [124].

Hollywood2		UCF Sports		YouTube	
Wang et al. [125]	47.7%	Klaser et al. [61]	86.7%	Liu et al. [77]	71.2%
Gilbert et al. [40]	50.9%	Kovashka et al. [63]	87.3%	Zhang et al. [140]	72.9%
Ullah et al. [120]	53.2%	Dense Trajectories [124]	88.2%	Ikizler-cinbis et al. [48]	75.2%
Mathe et al. [84]	57.6%	Wu et al. [134]	91.3%	Le et al. [69]	75.8%
Dense Trajectories [124]	58.3%	Action Bank [107]	95.0%	Dense Trajectories [124]	84.2%
STAP with fixation	59.6%	STAP with fixation	94.3%	STAP with fixation	–
Our method	**62.5%**	Our method	**95.3%**	Our method	**87.9%**

Table 6.3: Comparison of our proposed method with state-of-the-art methods in the literature.

We also compare the Average Precision per action class for Hollywood2 and YouTube datasets. On Hollywood2, we compare against the approach of [124] and [84]. As seen in Table 6.4, our STAP yields best results for 9 out of 12 action classes. On YouTube, STAP gives best results for 7 out of 11 action classes when compared with [124] and [69] as shown in Table 6.4.

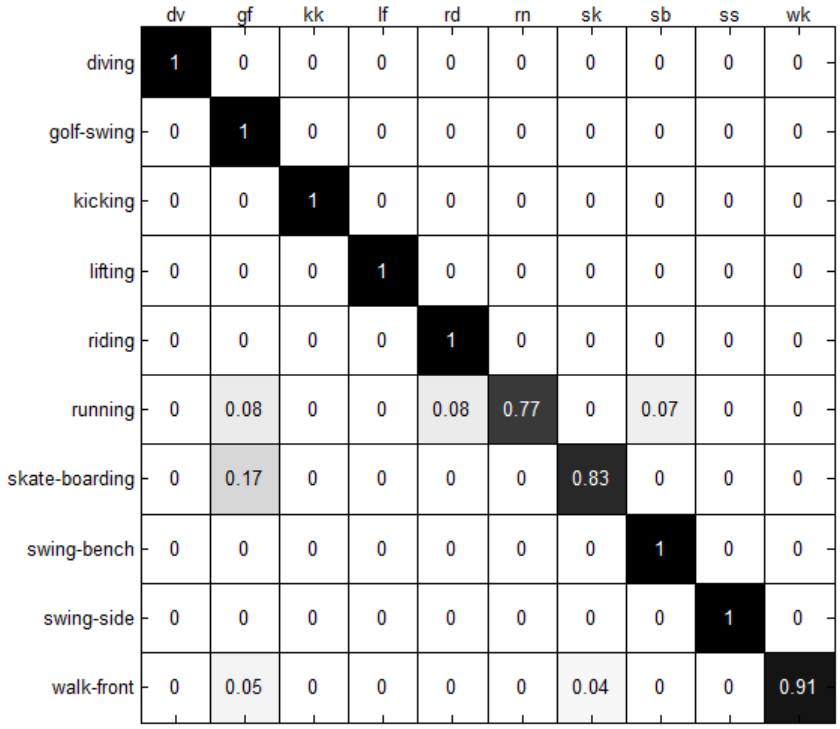


Figure 6.5: The confusion matrix of STAP on UCF Sports dataset.

Hollywood2			
	STAP	Dense Trajectories [124]	Mathe et al. [84]
AnswerPhone	**44.0**	32.6	23.7
DriveCar	**94.6**	88.0	92.8
Eat	**70.5**	65.2	70.0
FightPerson	77.6	**81.4**	76.1
GetOutCar	**55.9**	52.7	54.9
HandShake	**34.5**	29.6	27.9
HugPerson	45.0	**54.2**	39.5
Kiss	**68.2**	65.8	61.3
Run	**84.9**	82.1	82.2
SitDown	**73.9**	62.5	69.0
SitUp	25.1	20.0	**34.1**
StandUp	**75.7**	65.2	63.9
mAP	**62.5**	58.3	57.6

YouTube			
	STAP	Dense Trajectories [124]	Le et al. [69]
Shooting	**64.6**	43.0	46.5
Biking	90.3	**91.7**	86.9
Diving	98.7	**99.0**	93.0
Golf	92.3	**97.0**	85.0
Riding	**89.9**	85.0	76.0
Juggle	**80.8**	76.0	64.0
Swing	**91.2**	88.0	88.0
Tennis	**89.2**	71.0	56.0
Jumping	**95.0**	94.0	87.0
Spiking	**96.6**	95.0	81.0
Walking	78.1	**87.0**	78.1
Accuracy	**87.9**	84.2	75.2

Table 6.4: Average Precision and Accuracy (%) per action class for the Hollywood2 (upper) and YouTube (lower) dataset.

In addition, since we get inspired from human attention, we also conduct another comparison between STAP using our saliency prediction method and STAP with ground truth fixation. Our method achieves better results over the one with ground truth fixation. This is because the predicted saliency is generally more consistent in training set and testing set. We observe the failure case of action recognition based on STAP with ground truth fixation. For example, in the horse-riding action, the human fixation fixates on the human face, while the predicted saliency focuses on both human and the horse which are more informative corresponding to the action. We also perform STAP across various saliency prediction

models on UCF Sports dataset. As shown in Table 6.2, the performance of all saliency models surpasses all of baselines except [107]. STAP from our saliency prediction model achieves the best performance which shows the advantage of our proposed model for video saliency prediction.

6.6 Discussion

In this chapter, we have presented STAP, a simple yet effectively powerful method for action recognition on a wide variety of realistic videos. The proposed method combines dense sampling with spatial-temporal feature pooling driven by video saliency information. Extensive experimental results have clearly demonstrated the proposed STAP can achieve the state-of-the-art performances on diverse and popular action recognition datasets.

One may argue about the costly processing time of computing predicted saliency maps for each individual method. For a given 320×240 pixel image, the average processing time in second unit is as follows: Itti-Koch (0.23), AIM (2.01), ICL (0.89), SIM (1.1), FT (0.07), LSK (0.52), SR (0.81), GBVS (0.91), Signature-LAB (0.12), the motion map (1.21). Note that all of the implementation is currently not optimized in MATLAB. Our experimental computer is equipped with quad-core 2.67 GHz CPU and 16 GB RAM.

There exists a concern about the video size, which is currently fixed at 240×320 pixels for two large-scale datasets. [125] reported the performance decreasing when the video's resolution shrinks. It means the results are encouraging since the results on the original videos can be even better though more time consuming.

Last but not least, it is worth noting that dense sampling produces a very large number of features. Therefore, the experiment is more data storage and memory consuming than the one with the relatively sparse number of interest points as in [67, 83].

Chapter 7

Conclusion and Future Work

7.1 Conclusion

This thesis makes contributions on four aspects of visual saliency analysis, namely image re-attentionizing, 3d saliency, video saliency prediction, and action recognition. Each contribution is novel and interesting. Each work provides the proposed models with the extensive experiments which show superior performance than other state-of-the-art methods.

Image Re-Attentionizing. We propose a novel computational framework for the image re-attentionizing task. Our work is based on a premise that human eyes tend to look at the unique area in the image in both global and local sense. The experiments demonstrate that the recolored images successfully attract human attention to the target region(s) and in the meantime both spatial coherence and color coherence are well preserved. Although the proposed method yields a better experience, it still has limitations. The first is the boundary artifact when selecting target regions from superpixels. To overcome this issue, interactive methods can be applied to provide better region selection [73], [91]. Another solution is to increase the number of superpixesl in the image to provide finer over-segmentation. The second issue is the *unnatural* color for the objects which do not exist in the patch dataset. The remedy for this is to increase the dataset size.

3D Saliency. As aforementioned, the obtained depth from stereo images is unreliable. In addition, the existing datasets are small-scale. Therefore, the thesis

focuses on analyzing the depth matters on saliency based on a large-scale fixation dataset. We introduce an eye fixation dataset compiled from 600 images for both 2D and 3D scenes viewed by 80 participants. Using the state-of-the-art models for saliency detection, we have shown new performance bounds for this task. We expect that the newly built 3D eye fixation dataset will help the community enhance the study of visual attention in a real 3D environment. Furthermore, based on the analysis of the relationship between depth and saliency, extending the saliency models to include the proposed depth priors can consistently improve performance of current saliency models.

Video saliency prediction. We conduct the comparative studies between static saliency and dynamic saliency. To the best of our knowledge, this is the first research attempt to investigate this problem in depth. We first build the datasets of human fixation on both images and videos for the comparison purpose. Then we report several important observations of the relationship of static and dynamic saliency. Inspired by these observations, we propose the novel CMASS learning framework to fuse static saliency into dynamic saliency estimation to improve the video saliency prediction. Extensive experimental evaluations on the constructed datasets well demonstrate the effectiveness of the proposed method for video saliency prediction. We also apply the video saliency prediction method to help patients with hearing impairment to watch videos with narration. Extensive user studies clearly demonstrate the superiority of the proposed method in automatically determining the most suitable position to insert the subtitle.

STAP. We further investigate the application of saliency in a basic problem of computer vision, i.e., action recognition task. To tackle this task, we present an approach to recognize actions in videos by combining dense sampling with hierarchical spatial temporal pooling. An important contribution of this part is the proposed Spatial-Temporal Attention-aware Pooling scheme together with designed varieties of saliency information which can achieve state-of-art performance on diverse and popular action recognition datasets. We believe that our work will shed light on further researches in learning descriptors, saliency models and search algorithms for action recognition and for validating biological models of human visual attention.

7.2 Future Work

In the future, four applications of visual saliency, namely image re-attentionizing, 3D saliency, video saliency prediction and action recognition, will be the foci of our research work. We plan to learn from human fixation data to de-emphasize the areas which attract human fixation in the original image, yet are not the target region(s). The more challenging cases in dynamic scenes, namely video re-attentionizing, may also invite further research. Superpixels show their benefits to the image re-attentionizing.

Regarding 3D saliency, we are interested in how to integrate depth priors into various models instead of the late fusion methods. We want to include depth information directly into the computational models. This future approach is very promising since it moves another step to understand the human vision. We also would like to analyze the relationship between eye fixation and object attributes such as size, location, depth plane.

For video saliency prediction, given that time is an important parameter in the dynamic video scenes, we will consider the function of viewing time as in [20]. Currently the proposed framework only considers the camera motion. The target motion will be considered in the future work.

For action recognition, more semantically meaningful saliency information and new approaches for combining different types of saliency information will be further explored. Our current framework will be augmented with the depth information provided in some public datasets [96, 53]. The temporal channels are fixed in this work, and more work will be done for automatic partition and alignment in temporal domain.

We also would like to explore more interesting topics. For instance, though the models do well qualitatively, they have limited applications because they frequently do not match actual human saccades from eye-tracking data, and finding a closer match depends on tuning many design parameters. Thus, we want to investigate

the possibility of learning the color and depth information to predict where humans look as discussed in [59, 141, 78]. We also intend to learn the benefits of this saliency in more interesting applications, such as image classification, thumbnailing and video retargeting.

Bibliography

[1] http://www.oed.com/view/Entry/125348?redirectedFrom=naturalness.

[2] http://www.forbes.com/lists/2011/forbes-top-brands/list.html.

[3] Kinect: http://www.xbox.com/kinect.

[4] http://nicolas.burrus.name/index.php/Research/Kinect/Calibration.

[5] R. Achanta, S. Hemami, F. Estrada, and S. Süsstrunk. Frequency-tuned salient region detection. In *CVPR*, pages 1597–1604, 2009.

[6] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk. Slic superpixels compared to state-of-the-art superpixel methods. *IEEE Trans. Pattern Anal. Mach. Intell.*, 34(11):2274–2282, 2012.

[7] F. Attneave. Some informational aspects of visual perception. In *Psychological Review*, 1954.

[8] S. Avidan and A. Shamir. Seam carving for content-aware image resizing. *ACM Trans. Graph.*, 2007.

[9] T. Avraham and M. Lindenbaum. Esaliency (extended saliency): Meaningful attention using stochastic image modeling. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2010.

[10] M. Aziz and B. Mertsching. Pre-attentive detection of depth saliency using stereo vision. In *AIPR*, 2010.

[11] O. Boiman and M. Irani. Detecting irregularities in images and in video. *IJCV*, 2007.

[12] A. Borji and L. Itti. State-of-the-art in visual attention modeling. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(1):185–207, 2013.

[13] A. Bosch, A. Zisserman, and X. Muñoz. Representing shape with a spatial pyramid kernel. In *CIVR*, pages 401–408, 2007.

[14] Y. Boykov and V. Kolmogorov. An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. In *IEEE Trans. Pattern Anal. Mach. Intell.*, 2004.

[15] Y. Boykov, O. Veksler, and R. Zabih. Efficient approximate energy minimization via graph cuts. In *IEEE Trans. Pattern Anal. Mach. Intell.*, 2001.

[16] M. Bregonzio, S. Gong, and T. Xiang. Recognising action as clouds of space-time interest points. In *CVPR*, pages 1948–1955, 2009.

[17] T. Brox, A. Bruhn, N. Papenberg, and J. Weickert. High accuracy optical flow estimation based on a theory for warping. In *ECCV*, pages 25–36, 2004.

[18] N. Bruce and J. Tsotsos. Saliency based on information maximization. In *NIPS*, 2006.

[19] R. Carmi and L. Itti. The role of memory in guiding attention during natural vision. *Journal of Vision*, 2006.

[20] R. Carmi and L. Itti. Visual causes versus correlates of attentional selection in dynamic scenes. In *Vision Research*, 2006.

[21] M. Cerf, E. Frady, and C. Koch. Faces and text attract gaze independent of the task: Experimental data and computer model. *Journal of Vision*, 2010.

[22] K.-Y. Chang, T.-L. Liu, and S.-H. Lai. From co-saliency to co-segmentation: An efficient and fully unsupervised energy minimization model. *CVPR*, 2011.

[23] Q. Chen, Z. Song, Y. Hua, Z. Huang, and S. Yan. Hierarchical matching with side information for image classification. In *CVPR*, pages 3426–3433, 2012.

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

[25] W.-H. Cheng, W.-T. Chu, J.-H. Kuo, and J.-L. Wu. Automatic video region-of-interest determination based on user attention model. In *ISCAS*, 2005.

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

[27] N. Courty and É. Marchand. Visual perception based on salient features. In *IROS*, pages 1024–1029, 2003.

[28] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, pages 886–893, 2005.

[29] N. Dalal, B. Triggs, and C. Schmid. Human detection using oriented histograms of flow and appearance. In *ECCV*, pages 428–441, 2006.

[30] A. Dankers and A. Zelinsky. Cedar: A real-world vision system. *Mach. Vis. Appl.*, 16(1):47–58, 2004.

[31] W. Einhauser, M. Spain, and P. Perona. Objects predict fixations better than early saliency. *Journal of Vision*, 2008.

[32] L. Elazary and L. Itti. A Bayesian model for efficient visual search and recognition. *Vision Research*, 2010.

[33] W. Freeman, E. Pasztor, and O. Carmichael. Learning low-level vision. *IJCV*, 2000.

[34] S. Frintrop, E. Rome, A. Núchter, and H. Surmann. A bimodal laser-based attention system. *CVIU*, 2005.

[35] R. Gadde and K. Karlapalem. Aesthetic guideline driven photography by robots. In *IJCAI*, pages 2060–2065, 2011.

[36] R. Galitz, 2011. http://www.galitz.co.il/en/articles/composition-.shtml.

[37] D. Gao, V. Mahadevan, and N. Vasconcelos. On the plausibility of the discriminant center-surround hypothesis for visual saliency. *Journal of Vision*, 2008.

[38] V. Garcia, E. Debreuve, and M. Barlaud. Fast k nearest neighbor search using gpu. In *CVPR Workshop on Computer Vision on GPU*, 2008.

[39] B. Ghanem, T. Zhang, and N. Ahuja. Robust video registration applied to field-sports video analysis. ICASSP, 2012.

[40] A. Gilbert, J. Illingworth, and R. Bowden. Action recognition using mined hierarchical compound features. *IEEE Trans. Pattern Anal. Mach. Intell.*, 33(5):883–897, 2011.

[41] R. Gonzalez and R. Woods. Digital image processing. In *Prentice Hall*, 2002.

[42] J. Harel, C. Koch, and P. Perona. Graph-based visual saliency. In *NIPS*, pages 545–552, 2006.

[43] J. Ho, A. Peter, A. Ranganranjan, and M. Yang. An algebraic approach to affine registration of point sets. In *ICCV*, 2009.

[44] B. Hong and M. Brady. A topographic representation for mammogram segmentation. In *MICCAI (2)*, pages 730–737, 2003.

[45] R. Hong, M. Wang, M. Xu, S. Yan, and T.-S. Chua. Dynamic captioning: video accessibility enhancement for hearing impairment. In *ACM Multimedia*, 2010.

[46] X. Hou and L. Yan. Dynamic visual attention: Searching for coding length increments. In *NIPS*, 2008.

[47] X. Hou and L. Zhang. Saliency detection: A spectral residual approach. In *CVPR*, 2007.

[48] N. Ikizler-Cinbis and S. Sclaroff. Object, scene and actions: Combining multiple features for human action recognition. In *ECCV*, pages 494–507, 2010.

[49] L. Itti and C. Koch. A comparison of feature combination strategies for saliency-based visual attention systems. In *SPIE HVEIC*, 1999.

[50] L. Itti and C. Koch. Computational modelling of visual attention. *Neuroscience*, 2001.

[51] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Trans. Pattern Anal. Mach. Intell.*, 20(11):1254–1259, 1998.

[52] Y. Jang, S. Ban, and M. Lee. Stereo saliency map considering affective factors in a dynamic environment. In *ICONIP*, 2008.

[53] A. Janoch, S. Karayev, Y. Jia, J. T. Barron, M. Fritz, K. Saenko, and T. Darrell. A category-level 3-d object dataset: Putting the kinect to work. In *ICCV Workshops*, pages 1168–1174, 2011.

[54] L. Jansen, S. Onat, and P. Konig. Influence of disparity on fixation and saccades in free viewing of natural scenes. *Journal of Vision*, 2009.

[55] H. Jhuang, T. Serre, L. Wolf, and T. Poggio. A biologically inspired system for action recognition. In *ICCV*, pages 1–8, 2007.

[56] Y. Jia, C. Huang, and T. Darrell. Beyond spatial pyramids: Receptive field learning for pooled image features. In *CVPR*, pages 3370–3377, 2012.

[57] M. Jordan. Learning in graphical models. In *MIT Press*, 1998.

[58] A. Joulin, F. Bach, and J. Ponce. Discriminative clustering for image co-segmentation. In *CVPR*, pages 1943–1950, 2010.

[59] T. Judd, K. Ehinger, F. Durand, and A. Torralba. Learning to predict where humans look. In *ICCV*, pages 2106–2113, 2009.

[60] W. Kienzle, M. Franz, B. Schölkopf, and F. Wichmann. Center-surround patterns emerge as optimal predictors for human saccade targets. *Journal of Vision*, 2009.

[61] A. Kläser, M. Marszalek, and C. Schmid. A spatio-temporal descriptor based on 3d-gradients. In *BMVC*, 2008.

[62] V. Kolmogorov and R. Zabih. What energy functions can be minimized via graphcuts. In *IEEE Trans. Pattern Anal. Mach. Intell.*, 2004.

[63] A. Kovashka and K. Grauman. Learning a hierarchy of discriminative space-time neighborhood features for human action recognition. In *CVPR*, pages 2046–2053, 2010.

[64] C. Lang, G. Liu, J. Yu, and S. Yan. Saliency detection by multi-task sparsity pursuit. *TIP*, 2011.

[65] C. Lang, T. Nguyen, H. Katti, K. Yadati, M. Kankanhalli, and S. Yan. Depth matters: Influence of depth cues on visual saliency. In *ECCV*, pages 101–115, 2012.

[66] I. Laptev. On space-time interest points. *International Journal of Computer Vision*, 64(2-3):107–123, 2005.

[67] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *CVPR*, 2008.

[68] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, pages 2169–2178, 2006.

[69] Q. Le, W. Zou, S. Yeung, and A. Ng. Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis. In *CVPR*, pages 3361–3368, 2011.

[70] H. Li and K. N. Ngan. A co-saliency model of image pairs. *IEEE Transactions on Image Processing*, 20(12):3365–3375, 2011.

[71] J. Li, Y. Tian, T. Huang, and W. Gao. Probabilistic multi-task learning for visual saliency estimation in video. *IJCV*, 2010.

[72] L. Li, H. Su, E. Xing, and F. Li. Object bank: A high-level image representation for scene classification & semantic feature sparsification. In *NIPS*, pages 1378–1386, 2010.

[73] Y. Li, J. Sun, C. Tang, and H. Shum. Lazy snapping. *ACM Trans. Graph.*, 2004.

[74] R. Likert. A technique for the measurement of attitudes. *Archives of Psychology*, 1932.

[75] H. Liu, S. Jiang, Q. Huang, and C. Xu. A generic virtual content insertion system based on visual attention analysis. In *ACM Multimedia*, pages 379–388, 2008.

[76] J. Liu, B. Kuipers, and S. Savarese. Recognizing human actions by attributes. In *CVPR*, pages 3337–3344, 2011.

[77] J. Liu, J. Luo, and M. Shah. Recognizing realistic actions from videos. In *CVPR*, pages 1996–2003, 2009.

[78] T. Liu, Z. Yuan, J. Sun, J. Wang, N. Zheng, X. Tang, and H. Shum. Learning to detect a salient object. *IEEE Trans. Pattern Anal. Mach. Intell.*, 33(2):353–367, 2011.

[79] B. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *IJCAI*, pages 674–679, 1981.

[80] V. Mahadevan and N. Vasconcelos. Spatiotemporal saliency in highly dynamic scenes. In *IEEE Trans. Pattern Anal. Mach. Intell.*, 2010.

[81] S. Marat, M. Guironnet, and D. Pellerin. Video summarization using a visual attention model. *Proceedings of the 15th European Signal Processing Conference*, 2007.

[82] M. Marszalek, I. Laptev, and C. Schmid. Actions in context. In *CVPR*, 2009.

[83] M. Marszalek, I. Laptev, and C. Schmid. Actions in context. In *CVPR*, pages 2929–2936, 2009.

[84] S. Mathe and C. Sminchisescu. Dynamic eye movement datasets and learnt saliency models for visual action recognition. In *ECCV*, pages 842–856, 2012.

[85] T. Mei, X. Hua, and S. Li. Videosense: A contextual in-video advertising system. *IEEE Trans. Circuits Syst. Video Techn.*, 19(12):1866–1879, 2009.

[86] T. Mei, L. Li, X. Hua, and S. Li. Imagesense: Towards contextual image advertising. *TOMCCAP*, 8(1):6, 2012.

[87] T. Mei, L. Li, X.-S. Hua, and S. Li. Imagesense: Towards contextual image advertising. *TOMCCAP*, 2012.

[88] R. Messing, C. Pal, and H. Kautz. Activity recognition using the velocity histories of tracked keypoints. In *ICCV*, pages 104–111, 2009.

[89] O. Meur and J. Chevet. Relevance of a feed-forward model of visual attention for goal-oriented and free-viewing tasks. *TIP*, 2010.

[90] O. L. Meur, P. L. Callet, and D. Barba. Predicting visual fixations on video based on low-level visual features. *Vision Research*, 2007.

[91] E. Mortensen and W. Barrett. Intelligent scissors for image composition. In *SIGGRAPH*, 1995.

[92] C. Muhl, Y. Nagai, and G. Sagerer. On constructing a communicative space in hri. In *KI*, pages 264–278, 2007.

[93] N. Murray, M. Vanrell, X. Otazu, and A. Párraga. Saliency estimation using a non-parametric low-level vision model. In *CVPR*, pages 433–440, 2011.

[94] K. Nakayama and G.Silverman. Serial and parallel processing of visual feature conjunctions. *Nature*, 1986.

[95] T. Nguyen, S. Liu, B. Ni, J. Tan, Y. Rui, and S. Yan. Sense beauty via face, dressing, and/or voice. In *ACM Multimedia*, pages 239–248, 2012.

[96] B. Ni, G. Wang, and P. Moulin. Rgbd-hudaact: A color-depth video database for human daily activity recognition. In *ICCV Workshops*, pages 1147–1153, 2011.

[97] L. OHare and P. Hibbard. Visual discomfort and blur. In *i-Perception*, 2011.

[98] N. Ouerhani and H. Hugli. Computing visual attention from scene depth. In *ICPR*, 2000.

[99] N. Ouerhani, R. Wartburg, and H. Hugli. Empirical validation of the saliency-based model of visual attention. *Electronic Letters on CVIA*, 2004.

[100] R. Peters and L. Itti. Beyond bottom-up: Incorporating task-dependent influences into a computational model of spatial attention. In *CVPR*, 2007.

[101] H. Quan and L. Schiaatti. Examination of 3d visual attention in stereoscopic video content. *SPIE-IST Electronic Imaging*, 2011.

[102] S. Ramanathan, H. Katti, N. Sebe, M. Kankanhalli, and T.-S. Chua. An eye fixation database for saliency detection in images. In *ECCV*, 2010.

[103] R. Rao and D. Ballard. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-eld effects. *Nature neuroscience*, 1999.

[104] E. Reinhard, M. Ashikhmin, B. Gooch, and P. Shirley. Color transfer between images. *IEEE Computer Graphics and Applications*, 2001.

[105] B. Russell, A. Torralba, K. Murphy, and W. Freeman. Labelme: A database and web-based tool for image annotation. *IJCV*, 2008.

[106] N. G. Sadaka and L. J. Karam. Efficient perceptual attentive super-resolution. In *ICIP*, pages 3113–3116, 2009.

[107] S. Sadanand and J. Corso. Action bank: A high-level representation of activity in video. In *CVPR*, pages 1234–1241, 2012.

[108] F. Schumann, W. Einhauser, J. Vockeroth, K. Bartl, E. Schneider, and P. Konig. Salient features in gaze- aligned recordings of human visual input during free exploratoin of natural environments. In *Journal of Vision*, 2008.

[109] H. Seo and P. Milanfar. Static and space-time visual saliency detection by self-resemblance. *Journal of Vision*, 2009.

[110] J. Shotton, J. Winn, C. Rother, and A. Criminisi. Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context. *IJCV*, 2009.

[111] C. Siagian and L. Itti. Biologically inspired mobile robot vision localization. *IEEE Transactions on Robotics*, 25(4):861–873, 2009.

[112] S. Su, F. Durand, and M. Agrawala. An inverted saliency model for display enhancement. In *MIT Student Oxygen Workshop*, 2004.

[113] J. Sun, X. Wu, S. Yan, L. Cheong, T. Chua, and J. Li. Hierarchical spatio-temporal context modeling for action recognition. In *CVPR*, pages 2004–2011, 2009.

[114] X. Sun, H. Yao, R. Ji, and S. Liu. Photo assessment based on computational visual attention model. In *ACM MM*, 2009.

[115] B. Tatler, R. Baddeley, and I. Gilchrist. Visual correlates of fixation selection: effects of scale and time. *Vision Research*, 2005.

[116] A. Toet. Computational versus psychophysical bottom-up image saliency: A comparative evaluation study. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2011.

[117] A. Torralba, A. Oliva, M. Castelhano, and J. Henderson. Contextual guidance of eye movements and attention in real-world scenes: The role of global features in object search. *Psychological Review*, 2006.

[118] A. Treisman and G. Gelade. A feature-integration theory of attention. *Cognitive Psychology*, 1980.

[119] J. Tsotsos, S. Culhane, W. Wai, Y. Lai, and N. Davis. Modelling visual attention via selective tuning. *Artificial Intelligence*, 1995.

[120] M. Ullah, S. Parizi, and I. Laptev. Improving bag-of-features action recognition with non-local cues. In *BMVC*, pages 1–11, 2010.

[121] V. Vapnik, S. Golowich, and A. Smola. Support vector method for function approximation, regression estimation and signal processing. In *NIPS*, pages 281–287, 1996.

[122] A. Vedaldi and B. Fulkerson. Vlfeat: an open and portable library of computer vision algorithms. In *ACM Multimedia*, pages 1469–1472, 2010.

[123] B. Velichkovsky, M. Pomplun, J. Rieser, and H. Ritter. Eye-movement-based research paradigms. In *Visual Attention and Cognition*, 2009.

[124] H. Wang, A. Kläser, C. Schmid, and C. Liu. Action recognition by dense trajectories. In *CVPR*, pages 3169–3176, 2011.

[125] H. Wang, M. Ullah, A. Kläser, I. Laptev, and C. Schmid. Evaluation of local spatio-temporal features for action recognition. In *BMVC*, 2009.

[126] J. Wang, L. Quan, J. Sun, X. Tang, and H. Shum. Picture collage. In *CVPR*, pages 347–354, 2006.

[127] W. Wang, Y. Wang, Q. Huang, and W. Gao. Measuring visual saliency by site entropy rate. *CVPR*, 2010.

[128] G. Willems, T. Tuytelaars, and L. Gool. An efficient dense and scale-invariant spatio-temporal interest point detector. In *ECCV*, pages 650–663, 2008.

[129] J. Wolfe. Guided search 2.0: A revised model of visual search. *Psychonomic Bulletin and Review*, 1994.

[130] J. Wolfe and T. Horowitz. What attributes guide the deployment of visual attention and how do they do it? *Neuroscience*, 2004.

[131] J. Wolfe, M. V, K. Evans, and M. Greene. Visual search in scenes involves selective and nonselective pathways. In *Trends in Cognitive Sciences*, 2011.

[132] L. Wong and K. Low. Saliency enhanced image aesthetics class prediction. In *ICIP*, 2009.

[133] L. Wong and K. Low. Saliency retargetting: An approach to enhance image aesthetics. In *WACV*, 2011.

[134] X. Wu, D. Xu, L. Duan, and J. Luo. Action recognition using context and appearance distribution features. In *CVPR*, pages 489–496, 2011.

[135] J. Yang, J. Wright, T. Huang, and Y. Ma. Image super-resolution via sparse representation. *IEEE Transactions on Image Processing*, 19:2861–2873, November 2010.

[136] A. Yarbus. Eye movements and vision. In *Plenum Press*, 1967.

[137] Y. Zhai and M. Shah. Visual attention detection in video sequences using spatiotemporal cues. In *ACM Multimedia*, 2006.

[138] J. Zhang, M. Marszalek, S. Lazebnik, and C. Schmid. Local features and kernels for classification of texture and object categories: A comprehensive study. *International Journal of Computer Vision*, 73(2):213–238, 2007.

[139] L. Zhang, M. Tong, T. Marks, H. Shan, and G. Cottrell. Sun: A bayesian framework for saliency using natural statistics. *Journal of Vision*, 2008.

[140] Y. Zhang, X. Liu, M. Chang, W. Ge, and T. Chen. Spatio-temporal phrases for activity recognition. In *ECCV (3)*, pages 707–721, 2012.

[141] Q. Zhao and C. Koch. Learning visual saliency by combining feature maps in a nonlinear manner using adaboost. *Journal of Vision*, 12(6):1–15, 2012.

Publication List

1. Tam V. Nguyen, Mengdi Xu, Guangyu Cao, Qi Tian, Mohan Kankanhalli, Shuicheng Yan. Static Saliency vs. Dynamic Saliency: A Comparative Study. In ACM Multimedia 2013 (Oral presentation)

2. Tam V. Nguyen, Bingbing Ni, Hairong Liu, Wei Xia, Jiebo Luo, Mohan Kankanhalli, Shuicheng Yan. "Image Re-Attentionizing". In IEEE Transactions on Multimedia (T-MM), 2013

3. Tam V. Nguyen, Si Liu, Jun Tan, Bingbing Ni, Yong Rui, Shuicheng Yan. "Towards Decrypting Attractiveness via Multi-Modality Cues". In ACM Transactions on Multimedia Computing, Communications and Applications (TOMCCAP), 2013.

4. Tam V. Nguyen, Si Liu, Bingbing Ni, Jun Tan, Yong Rui, Shuicheng Yan. "Sense Beauty via Face, Dressing, and/or Voice". In ACM Multimedia 2012 (Oral presentation)

5. Tam V. Nguyen, Lusong Li, Jun Tan, Shuicheng Yan. "3DME: 3D Media Express from RGB-D Images". In ACM Multimedia 2012

6. Si Liu* , Tam V. Nguyen* , Jiashi Feng, Meng Wang, Shuicheng Yan. "Hi, Magic Closet, Tell Me What to Wear!". In ACM Multimedia 2012 (* indicates equal contribution)

7. Congyan Lang*, Tam V. Nguyen*, Harish Katti*, Karthik Yadati, Mohan Kankanhalli, Shuicheng Yan. "Depth Matters: Influence of Depth Cues on Visual Saliency". In European Conference on Computer Vision (ECCV) 2012. (* indicates equal contribution)

8. Tam V. Nguyen, Shuicheng Yan. "Seeing Your Weight–An application in targeted advertisement". In International Conference on Computer Vision (ICCV) 2011 (demo)

Awards

1. Best Technical Demonstration in ACM Multimedia 2012
2. 2nd Prize, ICPR 2012 contest of "Kitchen Scene Context based Gesture Recognition"
3. Student Travel Grant to ACM Multimedia 2012
4. Best paper presentation of Signal Processing and New Media Track, ECE Graduate Student Symposium 2013