RANDOM CENTER-SURROUND APPROACHES FOR MODELING VISUAL SALIENCY

TADMERI NARAYAN VIKRAM

Applied Informatics Group & Research Institute for Cognition and Robotics
Faculty of Technology
Bielefeld University

Thesis Adjudication Committee
Prof. Dr. Barbara Hammer, Bielefeld University
Prof. Dr. Britta Wrede, Bielefeld University
Prof. Dr. John K. Tsotsos, York University
Dr. Wolfram Schenck, Bielefeld University

Dedicated to the fond memories of my grandparents
Smt. Sowbhagyamma Keshavachar & Sri. Kalkunte Keshavachar

Sir, an equation has no meaning for me unless it expresses a thought of god

— Srinivasa Ramanujan Iyengar

ACKNOWLEDGMENTS

Many thanks to all of you who made this research possible. Over the last three years I have had the opportunity to interact and work the some of the finest minds in contemporary robotics research. This thesis work which I carried out, is essentially a product of these interactions and bears the influence of many wonderful ideas from cognitive sciences and computer vision.

First of all, I would like to thank Britta Wrede, Marko Tscherepanow and Agnes Swadzba as they were the principal investigators of this research project. It was Britta Wrede's decision to hire me and the subsequent generous funding that set the stage. The research was carried out under the able supervision of Marko Tscherepanow, who became my best friend and worst critic over the entire span. Subsequently, Agnes Swadzba directed the thesis writing and her efforts resulted in its timely completion.

Thanks to Franz Kummert who was always ready to help. Thanks also to Tom Ziemke and Robert Lowe who hosted me at their lab in University of Skovde, Sweden. Special thanks to Angelo Cangelosi, for fostering a vibrant interaction within the RobotDoC network which exposed me to many different aspects of research. Thanks to all of my co-authors, RobotDoC fellows and colleagues at the Applied Informatics Group for helping me maintain my spirits and motivation.

Many thanks to my parents and family for their efforts, patience and guidance all throughout. Lastly, I would like to thank my wife Subha, without whom I would not have been as productive as I am now.

ABSTRACT

Computational models of visual saliency have been used to detect salient regions and simulate human eye-gaze on images and videos. A majority of the existing approaches are highly parametric in nature. They are specialized to predict either eye-gaze or detect salient regions, but not both simultaneously. Like other computer vision approaches, the saliency models too impose pre-specified grids to process the image. In this context we explore ways of exploiting random/stochastic algorithmic approaches for saliency computation to address issues like pre-specified grids, computational efficiency, parameter set etc. We propose three different approaches for saliency computation on images and provide elaborate benchmarking results with respect to other saliency systems. Consequently, we have been successful in improving the state-of-the-art in terms of eye-gaze prediction and salient region detection performance of the saliency systems. In addition, we have extended one of our proposed saliency approaches to predict eye-gaze while viewing a tutoring or goal-directed action scenario. Along with the proposed algorithms, we also have created a video dataset for evaluating saliency systems in the context of goal-directed action. We hope that the proposed approaches for saliency computation, exprimental protocols, resulting video dataset and the ensuing discussions will help the community in developing more sophisticated systems of visual saliency.

CONTENTS

1 INTRODUCTION

In a clear night sky, the bright Venus stands out from the star clusters and captures our sight. A known or a familiar book attracts our interest for a moment, while scanning a stack in the library. A goal-keeper focuses on the striker during a penalty kick, and not at the spectators. This cognitive ability to discern the relevant stimuli from the rest is called attention.

1.1 VISUAL ATTENTION

From a philosophical point of view, there was much debate about attention beginning from the pre-historic times. Ancient Greeks postulated that the shift in attention from one object to another was due to a ray of light emanating from the eyes [128]. The Hindu and Confucian scholars wrote detailed texts which describe the link between attention and meditation. However, the first scientific attempt to define attention is attributed to the French mathematician Descartes [23] in 1649, where he describes attention as a process of thought suppression due to arousal.

Several descriptive definitions of attention have been proposed subsequently, but they were either incomplete or vague. In order to address this, the German psychologist Herbart [40] in 1824, proposed the first mathematical model of attention. This model attempted to capture emotion, arousal and other psychological attributes as differential variables. However, it was far ahead of its times, and the lack of supporting instrumentation coupled with sparse knowledge of neurobiology failed to advance it any further. The modern day definition of visual attention was put forward by James [49] in 1890, which states – " Everyone knows what attention is. It is the taking possession by the mind, in clear and vivid form, of one out of what seem several simultaneously possible objects or trains of thought. Focalization, concentration, of consciousness are of its essence. It implies withdrawal from some things in order to deal effectively with others, and is a condition which has a real opposite in the confused, dazed, scatterbrained state which in French is called *distraction*, and *Zerstreutheit* in German". This definition has remained as the bed rock of attention research, as all the theories proposed later conforms to it.

The most visible external manifestation of visual attention are the eye movements. Fovea, which is the center of the retina, has a higher resolution on the point of visual space where the human is attending to. However, Helmholtz [112] proved that humans could still attend to

various spatial locations on a scene without eye-gaze re-orientation. This was called as the *covert attention*, while the former is referred to as *overt attention*. Contemporary psychologists and neuroscientists view attention as the cognitive function which attenuates irrelevant features through a two stage selection process leading to a spotlight. The first stage is called as the *bottom-up attention*, which is purely driven by the input stimulus. It is also alternatively referred to as *exogenous, feature driven* or *context free attention*. The second stage is called as the *top-down attention*, which is goal driven. This is further referred to as *endogenous, goal driven* or *cued attention*. The distinction between these stages of attention is clearly demonstrated in the experiments conducted by Yarbus [123], an example of which is given in Fig. 1.

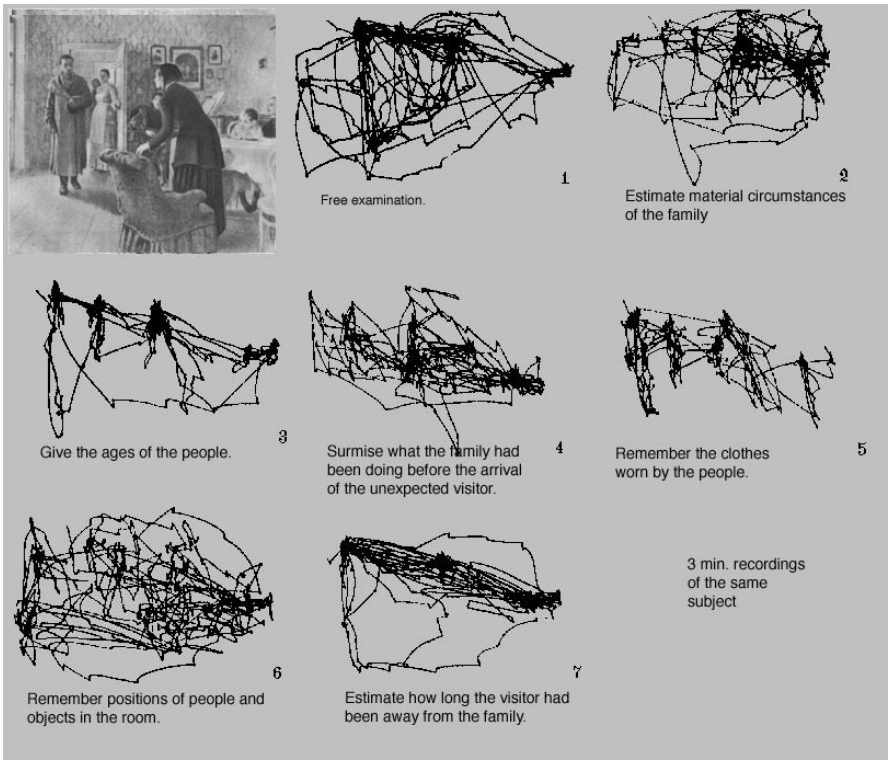


Figure 1: Study conducted by Yarbus [123]. (1) shows the eye-gaze in a free examination condition. (2) to (7) show that the eye-gaze patterns are specific for an objective that drives the image viewing. The result in (1) is a consequence of a bottom-up attention process, while the rest are due to top-down attention. Observe that the eye-gaze shifts vary with respect to the context.

The spotlight metaphor attributed to attention, describes the process where our consciousness is shifted from the current location to the next on the visual scene, despite the scene remaining static. The shifting of attentional spotlights from one location to another is seen as a result of the interaction between the top-down and bottom-up attentional processes [17]. The attention spotlight is seen as an effect,

while the reasons for the cause remained under investigation. Several associated issues like, location bias, size of the spotlight window, whether everything within the spotlight is processed equally, time required to shift from one location to another etc. continues to provoke new ideas to this date. In order to address some of these questions, particularly the bottom-up attentional process, the Feature integration theory [101] (FIT) was proposed. The theory assumed that visual scene is initially decomposed along a number of separable channels, such as color, orientation, spatial frequency, brightness and the direction of movement. Bottom-up attention is finally viewed as a product of the ensuing competition between these feature channels. The FIT [101] assumes that the feature competition is a black box process, and does not provide a mathematical framework to realize the theory.

1.2 SALIENCY MAPS

In the early 1970s, David Marr laid down the tri-level engineering perspective of biological vision [76]. This did not mention visual attention per se, but explained how a visual function would work. It elaborated a visual function in terms of computational, representational and physical perspectives. The computational aspect explains the objective of a visual function and its necessity. The representational aspect explains the algorithmic procedure which implements the function. Finally, the physical aspect explains the neuronal architecture which realizes the underlying algorithm. In this context, Koch and Ulmann [58] explained FIT [101] conforming to Marr's tri-level architecture [76]. They further proposed the existence of a saliency map, which encodes the degree of conspicuity at each spatial location in the scene. The first computational implementation of a saliency map which conformed to limited aspects of FIT was proposed by Sandon [91]. The saliency map proposed by Koch and Ulmann [58] was programmed a decade later by Itti et al. [48], when the image processing and computer vision programming routines started to mature. Since then the concept of saliency map has been central to the computer vision approach for modeling visual attention.

The ability of saliency maps to automatically predict interesting regions on images has been exploited in various computer vision and robotics applications. They are used to guide the robot's attention to visually interesting regions on a scene, thereby rendering a more human like behavior to its eye-gaze. It has also been used for applications like image thumbnailing, cropping, retargeting, collage creation, automatic target detection etc. The details of such applications are provided later in Section 2.10.

A wide variety of saliency maps have been proposed which operate on different computer vision paradigms. Most of them resemble the original center-surround architecture of Itti et al. [48], except that they

differ in the features used and the weighting given to each of them. Inspired by this prediction, several saliency maps were proposed which relies on the center-surround paradigm. It has been shown that the recent center-surround contrast based saliency maps, outperform a vast majority of existing saliency maps from other categories in terms of correlating with human eye-gaze [2, 3, 4, 106]. Despite this, the research on saliency maps is still relevant.

Most of the existing saliency models are constrained by the need of training data, large set of tunable parameters, ad-hoc fusion of features to generate the final saliency map, and not being scale invariant. The human visual system analyzes the input stimuli randomly, while the existing saliency systems process the image pixels sequentially. In addition, the saliency systems process all pixels equally where as in reality, the biological system may not equally process all parts of the visual stimuli. Overt attention is reflected in eye-gaze, while human vision also attends to salient regions without eye-gaze re-orientation. The current saliency systems are either efficient in predicting human eye-gaze on an image, or in detecting salient regions but not both.

Saliency systems that are tailored to handle static images vastly outnumber the ones which are capable of handling video streams. Computing saliency on video streams is relevant in the context of surveillance and human-robot interaction. A robust video based saliency system helps in guiding the robotic attention to relevant regions in an interaction scenario. One of the simple but inefficient solution is to decompose the video into image frames, and compute saliency on these images independently. The other sophisticated solutions propose to include motion history as an additional feature. The following strategies prove simplistic as most of the real life videos depict goal directed tasks which are driven by specific semantics (see Fig. 2). To the best of our knowledge, there does not exist a video saliency system which incorporates the task based semantics or contextual information to boost the performance of bottom-up saliency approaches on videos pertaining to actions and interaction.

1.3 THESIS CONTRIBUTIONS

The goal of this thesis is to build more reliable and complete bottom-up and top-down attention models. In the present time, saliency models are specialized to handle eye-gaze prediction on an image, salient region detection or video based saliency. All these three functionalities are relevant especially in the context of human-robot interaction and the saliency systems deployed on the robots are nothing but an aggregation of distinct components which caters to one of these tasks. The human visual system realizes a visual function through approximations in order to optimize computational performance. This issue is overlooked in most of the existing saliency approaches. In this

Figure 2: Goal directed events. Observe that the six video snap shots consists of a target or goal-directed task. These images do not consist of cues which hint towards any motion, and despite this we recognize the inherent motion pattern and deploy our attention accordingly. We also anticipate the trajectory in which the arrow, the horse or the balls move with great accuracy by conjuring mental images of the scene as if these were playing movies. Our attention system thus focuses only on the locations of the anticipated trajectory and suppresses those regions of the image which are not relevant to the action. [Images taken from Wikipedia]

context, we propose bottom-up saliency models which are stochastic in nature and could solve the problems of eye-gaze prediction and salient region detection concurrently. The eye-gaze correlation task involves predicting those regions in an image where the observer would fixate (along with the degree of fixation). On the other hand salient region detection task involves the automatic identification of those image regions which the observer thinks is most interesting. The saliency models are further extended to guide attention in an interaction scenario by fusing top-down information. The contributions of the thesis are:

We propose three different models of bottom-up saliency. The proposed models works on the paradigm of random center-surround contrast and have a single parameter which requires tuning. They also outperform most of the currently available saliency systems in terms of predicting human eye-gaze and detecting salient regions.

We provide a baseline to compare different saliency systems. It is necessary to quantitatively evaluate the performance of the saliency systems on multiple datasets and also for different tasks. The diversity in datasets and tasks enables us to draw a definitive conclusion about the robustness and reliability of the saliency systems.

We propose a saliency model which incorporates task based semantics to analyze goal directed actions. The proposed model incorporates a random center-surround contrast based saliency map which is coupled with task specific spatial priors. The incorporation of spatial priors reduce the search space and thereby redundant background is not processed at all. The proposed model outperforms other existing video based saliency systems in detecting saliency in a goal directed action video.

We provide a video dataset for saliency computation on goal directed actions. We make the entire video clips and the associated annotations retrieved from our experiments available online for the public.

2

REVIEW OF SALIENCY MODELS

Saliency is a perceptual quality which enables an object to stand-out of its immediate contexts. Models of visual attention generate a saliency map, which encodes the probability of a pixel location being salient. A computational model of visual attention is relevant in the context of processing an overload of multimedia and image data with limited computational resources.

A saliency map has applications with regard to automatically selecting visually salient regions and simulating human eye-gaze on a visual scene. The general architecture of a computational model of visual saliency was first presented in [48] as shown in Fig. 3.

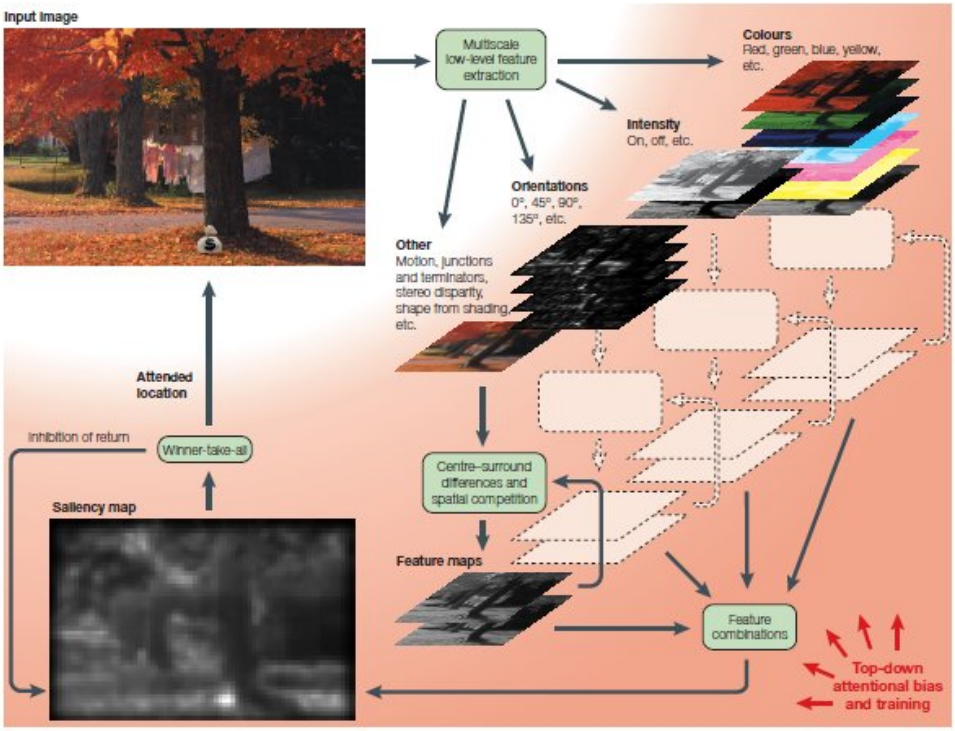


Figure 3: The architecture of the attention model proposed in [48]. The model decomposes the input image into several low level image features, and finally obtains the saliency map by the means of center-surround competition.

The aforementioned architecture has been the basis of the many existing models of visual saliency. In general the architecture of a saliency model has three distinct aspects namely, feature extraction, feature competition and feature fusion to obtain the saliency map. These models are essentially *bottom-up* in nature, as they predict saliency only in a free-viewing scenario which does not involve search or

recognition objectives. Hence, they are also alternatively referred to as bottom-up saliency models.

2.1 CATEGORIES OF SALIENCY MODELS

The theoretical framework for computation of saliency maps was first proposed by Koch et. al. [58] and later realized by Itti et. al. [48]. Subsequently it has led to the development of several other saliency approaches based on different mathematical and computational paradigms. The existing approaches can be classified into eight distinct categories based on the computational scheme they employ.

- Hierarchical approaches: They perform a multi-scale image processing and aggregate the inputs across different scales to compute the final saliency map.
- Spectral approaches: They operate by decomposing the input image into Fourier or Gabor spectrum channels and obtain the saliency maps by selecting the prominent spectral co-efficients.
- Power law based approaches: These approaches compute saliency maps by removing redundant patterns based on their frequency of occurrence. Rarely occurring patterns are considered salient while frequently occurring patterns are labeled redundant.
- Image contrast approaches: The mean pixel intensity value of the entire image or of a specified sub-window is utilized to compute the contrast of each pixel in the image. The contrast is analogously treated as the pixel saliency.
- Entropy-based approaches: The mutual information between patterns is employed to optimize the entropy value, where a larger entropy value indicates that a given pattern is salient.
- Center-surround approaches: These approaches compute the saliency of a pixel by contrasting the image features within a window centered on it.
- Hybrid approaches: Models of this paradigm employ a classifier in combination with one or more approaches to compute saliency.
- Top-down approaches: Such models couple components like face, object and line detection to re-weight the saliency map.

We briefly explore the existing saliency approaches based on the aforementioned categories in the sections to follow. Interested readers are pointed to [11, 30] for a more detailed and exhaustive review on saliency approaches.

2.2 HIERARCHICAL APPROACHES

The most popular approach in this category is the one proposed by Itti et al. [48]. This approach computes 41 different feature maps for a given input image based on color, texture, gradient and orientation information. Nine spatial scales are created using dyadic Gaussian pyramids. The resulting multi-scale image features are combined into a master saliency map. A normalization operator is employed, which globally promotes maps in which a small number of strong peaks of activity is present. Furthermore those maps which contain numerous comparable peak responses are globally suppressed. A dynamical neural network is employed which generates the eye-gaze locations in the order of decreasing saliency. The approach advocates a parallel implementation of feature extraction and attention-focusing system. In addition it down-scales the input image to an ultra low resolution ($\frac{1}{256}^{\text{th}}$ of the original), in order to achieve a fair computational run time. As mentioned earlier, this approach has served as a classical benchmark system and has been the basis of all the existing saliency systems.

The global selection and weighting of the multiscale information plays a crucial role in Itti et al. [48]. In order to partially address the issue of scale dependency, a saliency model based on generalized principal component analysis was proposed in Hu et al. [44]. The image is represented in a two dimensional space using polar transformation of its features so that each region in the image lies in a one dimensional linear subspace. The robustness of subspace estimation is improved by using a weighted least square approximation. The weights are calculated from the distribution of k nearest neighbors in the subspace to reduce the sensitivity of outliers. A region attention measure is further defined which calculates the saliency of each region by considering both feature contrast and geometric properties.

A saliency map is highly sensitive to the global properties of an image. As a result of this, they do not take into account the specific keypoints in an image which remain stable over different spatial scales. In order to factor both sensitivity and stability, a local image extrema based saliency map was proposed in Maruta et al. [77]. Salient regions in an image are extracted from multiresolutional two dimensional distribution of local extrema. Saliency is defined as the stability of a local extrema on the scale–space. The input image is convoluted with a Gaussian function to obtain the scale–space representation. The local extrema of an image is further extracted at each resolution level in the three channels i.e, luminance, red–green and blue–yellow. On these three channels, the saliency is computed on the stability of local extrema at multiple resolutions. Finally, the local extrema maps of three channels are summed-up to obtain the master saliency map. Local extremas are more biased towards high contrast and small sized

local patches than large image regions. Consequently, the approaches which rely on keypoints are less sensitive in detecting salient regions.

The saliency map proposed by Itti et al. [48] was tailored for static images. Several variants of this map were proposed, which could handle spatio-temporal data. Mean value theorem was employed in Shi and Yang [96] to compute motion saliency which was further integrated into the original architecture of Itti et al. [48]. A biologically plausible dynamic saliency model based on Gabor filters was proposed in Marat et al. [75]. The model extracts two signals from each frame that correspond to the two main outputs of the retina. Each signal is further decomposed into elementary features by a bank of cortical like filters. These filters are used to extract both static and dynamic information, according to their frequency selectivity, providing two saliency maps: a static and a dynamic one. Both saliency maps are combined to obtain a master spatio-temporal saliency map per video frame, and is found to have a good processing speed. Despite the biologically plausible nature of this approach, the design of cortex like filter bank involves tuning several parameters and appropriate selection of scale–space.

A computational model of dynamic visual attention on the sphere which combines the static saliency map of Itti et al. [48] and motion features is proposed by Bogdanova et al. [13]. This approach is employed to detect salient locations in omni-directional image sequences while working directly in spherical coordinates. The motion pyramid is built by applying block matching and varying the block size. The spherical motion saliency map is obtained by fusing together the spherical motion magnitude and phase conspicuities. Furthermore, the motion map is fused with the static spherical saliency map in order to obtain the master saliency map. Detection of the spots of attention based on the dynamic saliency map on the sphere is applied on a sequence of real spherical images. Such models have a potential of being useful in surveillance and crowd anomaly behavior detection scenarios.

Saliency maps which rely on simple edge and gradient information are also proposed in the literature. Such maps are based on the intuition that the presence of an edge leads to a higher probability of a salient object in that region. A saliency map based on gradient and isophote framework was proposed by Valentini et al. [103]. Isophotes are lines connecting points of equal intensity and the shape of each isophote is invariant to changes in the contrast and brightness of an image. This property of isophotes is exploited to prune false salient detections. The gradient slope information is further employed to detect salient regions in images. In addition to gradient, color boosting and pixel curvature information are used to generate a scale specific saliency map. An appropriate scale is selected by exhaustively searching for the scale value that obtains the best overall results for a salient

object detection task on a training dataset. Integral images are used to improve the computational performance of feature extraction. The color boosting aspect of this approach is further improved by Vazquez et al. [104]. This approach is highly sensitive to color space that is used to represent the original image.

Most of the hierarchical methods operate at the pixel level. Operating at the patch level instead of pixels is generally regarded to be more efficient in the context of object detection in images. In this regard a patch based saliency was proposed by Goferman et al. [33]. The method imposes a regular grid and extract patches at each scale. Each pixel is represented by the set of multi-scale image patches centered on it. A pixel is considered salient when its enclosing patch is highly dissimilar to all other image patches. Multiple scale processing is incorporated to further decrease the saliency of background patches, as they are more likely to repeat at multiple scales. A pixel is considered attended if its saliency value exceeds a certain threshold. Furthermore, each pixel outside the attended areas is weighted according to its Euclidean distance to the closest attended pixel.

In general, the performance of hierarchical methods are constrained by the fusion of multiple maps and the requirement to process multiple features. The fusion method employed to compute the master saliency map from the various feature maps plays a vital role in its accuracy. Arriving at a generalized rule of fusion for various maps is complicated and requires intensive cross-validation to fine tune the fusion parameters. In general, hierarchical methods tend to ignore visually significant patterns which are locally occurring as they are primarily driven by global statistics of an image [48].

2.3 SPECTRAL APPROACHES

The approaches in this category process the spectral parameters of an image signal (like phase and amplitude) to build a saliency map. A saliency map based on the log spectrum was proposed by Hou and Zhang [41]. It is based on the hypothesis that the spectral residual contains the novel or rare parts of an image. The spectral residual of an image is obtained by subtracting the log of Fourier spectrum from the general shape of log spectra. This serves like the compressed representation of a scene. Using an inverse Fourier transform, the compressed representation is further transformed into the spatial domain resulting in the saliency map. The saliency map thus contains the non-trivial part of the scene. This method has been very popular because its programming simplicity.

Inspired by the idea of spectral residual for image saliency detection, a temporal spectral residual on video slices was proposed by Cui et al. [22]. This can automatically separate foreground object motion from the background using threshold selection and voting

schemes. Different from conventional background modeling methods with complex mathematical models, this method is based on Fourier spectrum analysis. Saliency map is obtained by transforming the spectral residual back to spatial domain, where the high value pixels correspond to the salient regions.

A Gabor feature based saliency map was proposed by Gao et al. [31]. It is driven by the idea that in the absence of high-level goals, the most salient locations of the visual field are those that enable the discrimination between center and surround with smallest expected probability of error. The input image is decomposed into an intensity map and four broadly-tuned color channels. The four color channels are, in turn, combined into two color opponent channels. The opponent color maps and the intensity map are convolved with three Mexican hat wavelet filters, to generate nine feature channels plus a Gabor decomposition of the intensity map. The property of Gabor decomposition i.e. the bow-tie shaped conditional distributions, is exploited to estimate the posterior probability of a location being salient. Despite the efficiency of this approach, the generation of multiple feature channels causes a computational overload.

A saliency map based on phase spectrum of quaternion Fourier transform (PQFT) was proposed by Guo et al. [37]. A hierarchical selectivity framework based on the PQFT model was introduced to construct the tree structure representation of an image. The model resembles the one presented by Hou and Zhang [41], except that it operates only on phase spectrum and ignores the amplitude spectrum.

A framework based on the color and orientation distribution in images to compute saliency was proposed by Gopalakrishnan et al. [35]. The color saliency framework detects salient regions based on the spatial distribution of the component colors. A Gaussian mixture model is fit in the hue–saturation space to identify outliers. The orientation saliency framework detects salient regions in images based on the global and local responses of different orientations in the image. The master saliency map is selected as either color saliency map or orientation saliency map based on a pre-specified threshold.

A saliency map based on two dimensional log Gabor wavelets was proposed by Wang et al. [113]. The low-level image irregularities are initially isolated using log Gabor wavelets. These irregularities are subsequently integrated by considering a center bias matrix to construct a bottom-up saliency map.

Spectral approaches ignore local image information entirely. In order to address this issue, several models have been proposed which include local image information. A saliency detection model by combining global information from frequency domain and local information from spatial domain was proposed by Li et al. [66]. Redundant patterns in the image are eliminated by performing spectrum smooth-

ing, while the informative regions are enhanced using a center-surround mechanism in the spatial domain. The outputs from these two channels are further combined to produce the final saliency map.

As it can be seen from the illustrations by Guo and Zhang [38], Fourier-based methods are affected by the number of co-efficients selected for image reconstruction and the scale at which the input image is processed. Like the subspace analysis,the method results in loss of information during image reconstruction and is compromised by illumination, noise and other image artifacts.

2.4 POWER LAW BASED APPROACHES

Power law models imply that rarely occuring features are salient. A saliency model based on Zipf's law and other aspects of linguistic analysis was proposed by Caron et al. [18]. Zipf's law is used to model the frequency of feature recurrence in an image as power law distributions. These models characterize the structural complexity of image textures. The input image is first partitioned into sub-images and Zipf's law is applied to these sub-images. They are subsequently classified according to the characteristics of the power law models. Saliency is thus inversely proportional to the occurrence of a texture pattern.

A saliency model based on Weibull's distribution was presented by Yanulevskaya et al. [122]. The contrast of an image is modeled using a two-parameter Weibull's distribution. This distribution captures the structure of the local contrast and edge frequency in a meaningful way. Using a set of images with associated eye movements, the joint distribution of the Weibull parameters at fixated and non-fixated regions are computed. Subsequently, a classifier based on the log-likelihood ratio between these two joint distributions is built to generate the final saliency map.

Despite their theoretic appeal, the power law based saliency models have a large parameter set and the heuristic to fix and optimize them constitutes a major drawback.

2.5 IMAGE CONTRAST BASED APPROACHES

Image contrast based approaches measure the variation in intensity or gray value in a specified region of an image to infer saliency. Local image contrast was first used by Ma and Zhang [72] to compute the image saliency. The image is partitioned into patches of equal size, and the saliency is measured as the inter-patch color contrast. A fuzzy region growing method is further incorporated to obtain a binary saliency map which highlights salient regions. The size of the imposed partition grid determines the accuracy of the saliency map. On similar lines, a saliency model based on local texture contrast was

proposed by Hu et al. [43]. The image is divided into local patches at several scales, and a Gabor wavelet transform is applied on each of the patches. Each patch is further represented by the mean and the standard deviation of the wavelet coefficients. Saliency is computed as the average mean difference and the average standard deviation difference over a neighborhood of patches. The multi-scale processing involved in this method thus addresses the problem of fixing a standard grid size to an extent.

A saliency map based on multi-scale local image contrast was proposed by Achanta et al. [2]. Saliency is determined as the local contrast of an image region with respect to its neighborhood at various scales. This is computed as the distance between the average feature vector of the pixels of an image sub-region to that of its neighborhood. This results in a combined feature map at a given scale by using feature vectors for each pixel. This approach was found to have good performance in detecting salient regions of an image.

An edge distance driven saliency map was proposed by Rosin [88]. Saliency of a pixel is modeled as the inverse of the multi-scale distance to its nearest edge. The model assumes that high intensity edges attracts eye fixation. A gradient based edge detector is employed to produce an edge transform of the input image. The resultant gradient image is thresholded at various gray-scale levels to produce a set of binary edge images. Subsequently, a distance transform is applied on all of the binary edge images to cascade the edge information. The resultant distance transformed images are further cumulated to obtain the master saliency map.

A method for salient region detection that generates full resolution saliency maps with well-defined boundaries of salient objects was proposed by Achanta et al. [3]. The object boundaries are preserved by retaining substantially more frequency content from the original image, as the method is based on global image contrast. It exploits features like color and luminance, and is computationally efficient. Experiments have revealed that the method fails to work where color and contrast are not the dominant features of an image. However, this drawback is offset by the advantage that it outperforms most of the existing methods to detect salient regions despite its simplicity.

A saliency map which is computed as a combination of two different local contrast measures was proposed by Huang et al. [45]. The method computes the saliency map by combining the results of two transforms namely the discrete moment transform (DMT) and the discrete symmetry transform (DST). The DMT is computed by evaluating local central moments around each pixel, while the DST computes the local annular symmetry. The DMT allows determination of large areas of interest, and the DST is able to locate finer details inside regions identified by DMT. The resulting coarse and fine-grained saliency maps are fused to obtain the final saliency map.

These methods are successfully applied for proto-object detection and out-performs many state-of-the-art methods without having many of their drawbacks. Poor global contrast of an image affects the performance of these methodologies, and local-statistics based approaches for saliency computation have been proposed in the literature to address this problem.

2.6 ENTROPY-BASED APPROACHES

A saliency map based on sparse representation of images was presented by Sun et al. [98]. A group of basis functions are learned using independent component analysis from short term statistics instead of large scale natural statistics. The original input data is represented by a linear combination of the learnt basis functions which minimizes the loss of information. Each basis function thus provides a unique feature channel. These feature channels are considered as a surrogate representative of neuronal cluster in the brain. The average activity of the feature, and the feature activation rate is computed to assess the energy consumption while viewing a visual pattern. Larger energy consumption thus indicates larger signal saliency.

In Sun et al. [99] the saliency is modeled as a sequential eye-fixation probability. Bottom-up saliency at a given location is defined as the conditional probability of being chosen as the next eye-fixation position given the previous fixations. Each location is characterized by discrete and integer cosine transform features of a patch centered on it. Saliency at a given location, is further computed as the weighted sum of the distance between its features and the mean value of the features extracted from all other locations. The conditional probability is further maximized using an entropy based representation of the prior probabilities.

A saliency map based on rank sparsity decomposition was proposed by Yan et al. [121]. It employs sparse bases to represent image patches and estimates saliency through sparsity matrix decomposition. In order to achieve a computationally tractable framework, the saliency computation is further modeled as a convex optimization procedure. Learning the sparse bases though a large image corpus is one of the drawbacks of this approach.

A saliency model based on context-mediated probability distributions was proposed by Xu et al. [119]. The model assumes that the visual saliency is based on efficient encoding of the probability distributions of visual variables in specific contexts of a natural scene. This model is based on the results from neuroscience of the early visual system. The computational units in the early visual system do not act as feature detectors but rather as estimators of the probability distributions of a full range of visual variables in natural scenes. This subsequently leads to a measure of visual saliency of the input

stimulus. The same philosophy is engineered to obtain this model. Independent component analysis (ICA) is further used to measure the visual saliency obtained on the basis of these distributions estimated from a set of natural scenes.

The popular approach of Bruce and Tsotsos [15] is based on local contrasts and maximizes the mutual information between features by employing ICA bases. A set of ICA bases is pre-computed using a patch size of 7×7 pixels. Subsequently it is used to compute the conditional and joint distribution of features for information maximization. Experiments conducted on the York University eye-gaze dataset [16] has proven its efficiency. But this method is constrained by its emphasis on edges while ignoring salient regions [106]. It also adds a spurious border effect to the resultant image, and requires re-scaling of the original image to a lower scale in order to make the computational process more tractable. Another ICA based approach was proposed by Zhang et al. [129] where image self-information is utilized to estimate the probability of a target at each pixel position. It is further fused with top-down features derived from ICA bases to build the final saliency map. The method proposed by Wang et al. [114], employs sparse bases to extract sub-band features from an image. The mutual information between the sub-band features is calculated by realizing a random-walk on them. An extension of this paradigm can been seen in the recent approach proposed by Lin et al. [67], where the entropy of a center versus a surround region is computed as the saliency value of a pixel. Other entropy-based approaches (like [42, 114]) employ incremental coding length to compute the final saliency map. These methods which rely on information theoretic approaches are in general constrained by the requirement of training bases, the patch size parameters and the size of the training bases.

2.7 CENTER-SURROUND APPROACHES

A saliency map based on the Kullback-Leibler Divergence (KLD) of center surround features was proposed by Klein and Frintrop [56]. Distributions of the visual feature occurrences for a center and a surround region are estimated. The KLD between these distributions statistically represents the feature divergence between the center and surround. An efficient scale-space computation of center-surround pairs of arbitrary sizes is further incorporated. This enables the method to be more robust as compared to the methods based on fixed grid sizes.

A saliency map based on the ratio of center-surround dissimilarity was proposed by Huang et al. [46]. The saliency of a pixel is defined as the ratio of total dissimilar pixels in its center and surround regions. The master saliency map is obtained by combining these ratios of dissimilarities over multiple scales.

A center-surround model based on human visual cognition was proposed by LeMeur and Chevet [60]. The model uses luminance and color features to compute the saliency map. Contrast sensitivity functions are used to compute the gradients, and multi-scale sub-band filters are employed to process the input image at different scales. The resulting saliency maps are fused using an exponential mapping to obtain the master saliency map. This model is an extension of their previous work [59], which introduced the usage of contrast sensitivity functions and perceptual decomposition for saliency computation.

A proto-object based saliency map was presented by Orabona et al. [84]. The input image is converted to log-polar form to handle in-plane rotation of the objects. The image is further decomposed into opponent color channels and subjected to center-surround filters to highlight proto-objects. A watershed transform is subsequently applied to enforce perceptual grouping of proto-objects. A center-surround modification of Achanta et al. [3] was originally introduced in Achanta and Süsstrunk [4]. Symmetric center-surround masks are imposed and the method proposed by Achanta et al. [3] is applied on each masks instead of the entire image. The saliency is cumulated to compute the final saliency map. The accuracy of the method depends on the probability that the masks enclose a salient region completely.

A color contrast based center-surround mechanism was introduced by Murray et al. [80]. The input image is initially processed using Gabor-filters at various scales. The size of the center-surround filters and normalizing weights for each scale is learned apriori using a prior fixation data. An inverse wavelet transform is further applied to compute the master saliency map. A gradient based center-surround contrast saliency map was proposed by Seo and Milanfar [93, 94]. Initially, the gradients are computed using the Sobel operator. A regular grid is imposed on the gradient image, and the saliency is computed as a function of the inter-patch contrast. This is obtained by a custom local steering kernel. The same method is also extended to handle spatio-temporal saliency by considering a video stream as a three dimensional image. This method has been tested on static and video data and applied for tasks like boundary detection, target detection, motion prediction etc. It is shown to be robust for noise and drastic illumination changes, but is computationally expensive as a set of compound features needs to be computed for each pixel in the image. In order to achieve a tractable run-time, the input image is down-scaled. Selecting an appropriate window size for center and surround patches plays an important role in obtaining a higher quality saliency map.

2.8 HYBRID APPROACHES

A camera motion based saliency map was proposed by Abdollahian et al. [1]. Camera motion is used as a feature for identifying regions of interest as it is an indicator of both camera person's and viewer's focus of attention in the scene. Feature maps such as color contrast, object motion, face detection are fused with estimated camera motion parameters to obtain the final saliency map.

An object or target specific saliency map was proposed by Wei et al. [117]. Multiple feature based saliency maps are computed for all positive and negative examples of an object during the training phase. A weight vector is subsequently computed as the ratio of the mean target class saliency and the mean negative class saliency for each feature. During the test phase, all feature maps of a scene are combined and is further modulated by the weight vector. Local region-based entropy is used to identify the salient regions, and thus the final saliency map is obtained.

A radically different approach which tries to detect salient regions by estimating the probability of detecting an object in a given sliding window was proposed by Alexe et al. [5]. They employ the concept of super pixel straddling, coupled with edge density histograms, color contrasts and the saliency map of Itti et al. [48]. A linear classifier is trained on an image dataset to build a bag-of-features to arrive at a prior for an object in an image. The method is theoretically very attractive, but is subjected to high variations in the performance as too many features, maps and parameters are involved which require fine tuning.

An object extraction based saliency map was proposed by Yu et al. [125]. The framework comprises of two different importance maps. The first one extracts the borders and second extracts salient regions. The two maps are later fused and modulated by object priors to obtain the master saliency map. The object priors enable a pixel-level classification, while most of the object detection techniques work on the basis of a sliding window. This reduces the computational burden of the method significantly.

An associative memory based saliency map was proposed by Wilder et al. [118]. The model assumes that the saliency map is used in a search scenario and hence specialized modules for a particular task is designed. Each module includes a task-specific associative memory that maps from the visual representation (color, gradients and orientation) to an activation map which indicates the presence of an object. A task specific training network is employed that binds the features and classes by associative learning. In order to reduce the size of the feature set, a dimensionality reduction is further applied. The model accounts for key results in visual search on synthetic images and real-world images.

A stochastic model which estimates the probability of an image patch being salient was proposed by Avraham and Lindenbaum [8]. In contrast to other methods the model does not emphasize preference for local contrast. The algorithm iterates from a random preattentive segmentation and then uses a graphical model approximation to efficiently reveal those image segments that are more likely to be salient.

A graph-based visual saliency approach was proposed by Harel et al. [39]. It implements a Markovian representation of feature maps and utilizes a pyschovisual contrast measure to compute the dissimilarities between features. On similar lines, saliency detection is modeled as a Markov random walk performed on pixels represented as nodes in Gopalakrishnan et al. [34, 36]. The global properties of the image are extracted from the random walk on a complete graph, while the local properties are extracted from a k-regular graph. The saliency of nodes are inversely proportional to the frequency of they being visited. The equilibrium reaching times of the ergodic Markov chain is used to further identify the most salient node. A seeded salient region identification mechanism is later incorporated to identify the salient parts of the image.

A support vector machine (SVM) based saliency model was first proposed by Kienzle et al. [53]. The model consists of a non-linear mapping from an image patch to a real value, trained to yield positive outputs on fixated, and negative outputs on randomly selected image patches. Instead of using a predefined set of feature maps, the classifying function (an SVM) is learned directly from human eye movement data. To represent fixations and background locations accordingly, a square image patch is positioned at each of these locations and the pixel values are extracted. Fixation patches are identified as positive examples, while background patches are labeled as negative and later a SVM is learned. Learning an ideal patch size forms one of the constraints of this approach.

A task specific saliency model which combines both low level image features and classifier input was proposed by Li et al. [64]. The model is based on multi-scale wavelet decomposition and unbiased feature competition. A learning algorithm is further used to learn the task-related scene specific saliency functions. Both the local visual attributes and global scene characteristics are considered simultaneously in the learning framework. Unlike other approaches which employ a generic fusion, the said approach also learns a scene specific fusion rule. A k–nearest neighbor classifier is used to select an appropriate fusion strategy for each new scene to re-configure the prediction weights.

A conditional random field based saliency model was proposed by Liu et al. [68]. A set of features like multi-scale contrast, center surround histogram and color spatial distribution are used to describe

an image patch. A conditional random field is subsequently learned to effectively combine these features to generate a saliency map. This work has resulted in a large dataset which is widely used in evaluation of saliency maps.

Inspired by [68], conditional random fields were further utilized for salient object detection in videos in Liu et al. [69]. The salient object sequence detection is modeled as an energy minimization problem within a conditional random field framework. Static, spatio-temporal saliency and a global topic model were defined and integrated to identify a salient object sequence. A dynamic programming procedure is further designed to compute a global optimization, which results in a rectangle to represent each salient object.

Object specific saliency maps based on random trees was proposed by Moosmann et al. [78]. Random sub-windows are sampled on the training images, and randomized decision trees are built from these sub-windows as a classifier. On the test images sub-windows are again sampled randomly and each window is classified by the decision trees. The importance of each leaf node in the decision tree is learnt by an SVM, while histogram of oriented gradients are used as the image features at sub-windows.

In the context of video archival and retrieval, a user search behavior specific saliency model was proposed by Li et al. [65]. The saliency model is represented as a ranking system which sorts the image segments in a scene with respect to their relevance to the searching intention. A multi-task rank learning approach is proposed where visual saliency is estimated as a pair-wise rank learning problem. Videos are decomposed into scene clusters and multiple visual saliency models are learned for each scene cluster. The models refines itself further by automatically learning and integrating those image features that best distinguish targets from distractors in that cluster. A center-surround filter is later used to generate features for each visual subset in a scene. Various pre-attentive visual features are integrated with linear weights for saliency estimation.

A saliency model based on kernel density estimation (KDE) for segmentation has been proposed by Liu et al. [70]. The input image is partitioned into a set of regions using the mean shift algorithm. The pixels in each segmented region are then used as the samples to construct a KDE based non-parametric model. The color likelihood of a pixel with respect to each KDE model is defined using a custom dissimilarity measure. The color saliency and spatial saliency of each KDE model are then evaluated based on its color distinctiveness and spatial distribution. Based on color saliencies and spatial saliencies of all KDE models, the pixel-wise color map and spatial map are generated and fused to produce the master saliency map.

A Gaussian mixture model based saliency map was proposed by Ren et al. [86]. An adaptive mean shift algorithm is used to extract

superpixels from the input image. The superpixels are clustered using a Gaussian mixture model which captures color similarity. The saliency value for each cluster is computed using compactness metric together with modified page rank propagation. As compared to other superpixel based approaches, this method is robust with respect to over-segmentation.

Cascade of linear SVMs are employed by Khuwuthyakorn et al. [52] for salient object detection. It exploits a divide-and-conquer strategy by partitioning the feature space into sub-regions of linearly separable data-points. This yields a structured learning approach where a linear SVM is learnt for each region, along with the mixture weights and the combination parameters. Thus, the method learns the combination of salient features such that a mixture of classifiers can be used to recover objects of interest in the image.

Most of the machine learning based techniques hitherto model the salient region. However, a method proposed by Zhang et al. [130] incorporates background model in addition to object model to make this formulation more robust. A scalable subtractive clustering algorithm is used to cluster image pixels in different feature channels. The clusters are modulated by prior eye-movement behaviors and a maximum saliency difference technique which assigns each cluster as either background or foreground. A Gaussian mixture model is learned for both background and foreground models separately. During the validation phase, a Bayesian framework is employed to classify each pixel into salient object or background using the learned mixture models. The above detection procedures are repeated until the detection results achieve a steady state.

2.9 TOP-DOWN APPROACHES

In addition to conspicuity, top-down approaches add line, object or face detection results to re-weight the saliency map. An object specific saliency map based on symbolic interval valued representation was presented by Sang et al. [92]. Color, orientation, intensity and texture features are extracted from object templates. The mean and standard deviation of every corresponding feature from the set of object templates is stored as an object class representative. A sliding window is moved across the image where the features are extracted and contrasted with the object class representative to compute the similarity. The similarities are cumulated to produce the final saliency map.

A saliency map which combines both task and object priors in an human-robot interaction scenario is presented in [125, 126]. The model has three stages. At the first stage, there is a pre-attentive segmentation which selects the region of interest where the task relevant object is present. The detected salient region is further pruned to isolate the learned set of objects. At the final step, the detected objects

are recognized and classified. In contrast to other saliency systems, this approach includes searching and recognition paradigms, which is in line with the recursive binding paradigm advocated by Tsotsos [102].

Bayesian learning is incorporated into phase Fourier spectrum based saliency maps by Pie et al.[85]. This is further used for object detection. The learning is based on low level image features, and hence is suitable as a pre-processing step to boost existing object detection techniques.

Learning where to attend in an interaction scenario is one of the important aspects which most of the existing saliency maps do not handle. In this direction, a robotic system capable of learning the gaze following behavior in a real-world environment is presented by Kim et al. [55]. The system learns to detect salient objects and to distinguish a caregiver's head poses in a semi-autonomous manner. Multiple scenes containing different combinations of objects and head poses to the robot head are fed as training sequences. The system learns to associate the detected head pose with the correct spatial location of an object using a biologically plausible reinforcement learning mechanism.

On similar lines, a developmental robotics based multi-modal attention learning system was presented Aryananda [7]. An integrated framework is presented, which combines an object-based perceptual system, an adaptive multimodal attention system and spatio-temporal perceptual learning. This allows a robot to interact while collecting relevant data in an unsupervised way. The multi-modal attention system for the robot is coupled with a spatio-temporal perceptual learning mechanism. This incrementally adapts the saliency parameters for different types and locations of stimuli based on the agent's past sensory experiences. A genetic algorithm addon is presented in Verma and McOwan [105] which modifies the bottom-up attention map to detect changes in the scene in an optimal way.

The discriminant saliency mechanism for videos is further improved by computing tracking priors for salient region movement by Mahadevan and Vasconcelos [73]. A learning stage, combines a focus of attention mechanism and bottom-up saliency to identify a maximally discriminant set of features for target detection. The detection stage uses a feature based attention mechanism and a target-tuned top down discriminant saliency to detect the target. The tracker iterates between learning discriminant features from the target location in a video frame and detecting the location of the target in the next frame. Well known properties of natural image statistics are exploited to implement the tracker and achieve computational efficiency.

Most of salient objects are man made and have straight line structures. In order to exploit this feature, a Gabor wavelet based model which generates orientation specific saliency maps was proposed by

Fang et al. [27]. The dominant line orientations are computed, and the low-level bottom-up saliency maps are re-weighted using this orientation information. This method has been applied for the detection of moving vehicles in a scene.

Most of the object based top-down attention models assume that the target object also happens to be visually salient. However, a target object being salient in all scenarios need not be the case in reality. In order to address this problem, a saliency map was proposed by Moren et al. [79] which fuses Feature Gate framework with the saliency map of Itti et al. [48]. The Feature Gate assumption enhances the target object image features thus boosting its visual saliency. It further modulates the winner-takes-all (WTA) mechanism to detect the target object despite not being visually salient.

Top-down attention models which are used for a specific task like object detection or recognition have never been throughly investigated for their biological plausibility. In order to address this issue, the work presented by Han and Vasconcelos [90] introduces a top-down modulation to the famous HMAX model [57] for object recognition. The model precisely establishes the connection between saliency and object recognition, which was hitherto hypothesized in selective tuning model of attention [102]. The model also relies only on those statistics, which can be realized by a biological circuit. The framework is tested on standard object detection datasets and is found to be effective.

Saliency computation is viewed as an optimization problem in Borji et al. [14]. The saliency map proposed by Itti et al. [48] is further refined, by replacing the feature competition process by a convex optimization function. The goal of the optimization is to maximize the saliency with minimum processing cost. This is further integrated and tested for an object detection task.

Face and skin have been important cues to guide visual attention in natural scenes. This was first integrated as top-down cues in the saliency system proposed in Lee et al. [61]. Low level image feature maps are extracted in a series of segmentation processes. In the bottom-up module, all the features are combined into a bottom-up map where a target candidate has a vector form of input. The top-down input is determined by the geometrical relationship and Gaussian distance between the location of a target and the location of a cue. The face and skin color cues are also integrated into the top-down component. The bottom-up and top-down maps are fused with a neural network that has a dynamic and modulatory property which guides attention shifts sequentially. Similarly the graph-based visual saliency model [39] is also further extended to incorporate face maps in [19]. This modification has shown considerable performance improvement of the original model with respect to saliency computation in natural images.

A machine learning approach to address the relationship between object recognition and saliency is presented by Chang et al. [20]. A graphical model which views objectness and saliency as a factor graph formulation is presented. The framework conceptually integrates these two concepts via constructing a graphical model to account for their relationships. This concurrently improves their estimation by iteratively optimizing a novel energy function. This further helps in realizing the saliency model.

An adaptive saliency map which re-weights top-down and bottom-up features with regard to the context was presented by Xu et al. [120]. The top-down attention selection in the task space and the bottom-up attention selection in the image space is evaluated and combined using information theory. An information based scene context classification considering scene dynamics is formulated to bias attention selection. The method is computationally inexpensive to implement as the fusion rule between the maps depends on Bayesian statistics.

An extension to the work of Kienzle et al. [53] is presented in Lee et al. [62]. The method learns a regression model from fixated and non-fixated image patches from a training video sequence. The fixation strength is added as an ordinal label while computing the regression model. During a test sequence, patches are extracted and classified which obtains a class specific saliency map. This model is trained and tested on ten different categories of video sequences, and has a higher performance as compared to the performance of pure bottom-up based approaches on the test video sequence.

An object form detection based saliency map is proposed by Ban et al. [10]. The form detection is enabled by a Harris corner detector. The model generates top-down bias signals of form and color features for a specific object. The desired object is localized by an incremental learning mechanism together with object feature representation scheme. A fuzzy topology adaptive resonance theory network is used for object color and form biased attention. It incrementally learns and memorizes color and form features of arbitrary objects, and also generates top-down bias signal for selectively attending to a target object.

Shape cues has been incorporated for the first time to compute object specific saliency maps by Khan et al. [95]. The model is employed for recognizing object categories when using multiple cues by separating the color and shape. The color is used to guide attention by means of a top-down category-specific attention map. The color attention map is further deployed to modulate the shape features to boost those regions that are likely to contain an object instance.

2.10 APPLICATIONS

Artificial visual systems which are deployed in real-time are data intensive. They are expected to process a huge quantum of image and multimedia data with high speed without compromising on effectiveness. Such systems thus benefit by models of visual attention, as they automatically provide a prior knowledge about the relative importance of each pixel. This property can be utilized to help robots focus on an interesting location of a scene. It can also improve segmentation algorithms by automatically labeling the background. The image search can also be optimized by re-weighting its associated keywords based on saliency. We now explain a few such representative applications which have made a significant impact.

2.10.1 *Developmental Robotics*

Developmental robotics is a field which attempts to integrate human traits like life long learning through social interaction into humanoid robots. In order to sustain a coherent dialogue with an interaction partner, a robot needs to be equipped with a suitable attention mechanism. In this regard, the work presented by Nagai [81] is highly relevant. The saliency model proposed by Itti et al. [48] was modified to handle dynamic scenes in [81]. This system could predict human eye-gaze in an interaction scenario with greater accuracy. It further showed that a bottom-up saliency map could moderately predict semantically relevant regions in a dynamic scene without additional top-down inputs. The said architecture is presented in Fig. 4

This application of the saliency system has the potential to achieve intelligent robotic systems which can sustain a dialogue with a human interaction partner.

2.10.2 *Digital Photography*

Converting a color image into a gray scale without compromising on the aesthetics is one of the important aspects of digital photography. One such application can be seen in the work presented by Ancuti et al. [6]. The color conversion framework employs the saliency map of Itti et al. [48] to compute weighting factors for each location. Such applications have the potential to be employed for colorizing black-and-white motion pictures by combining it with a suitable machine learning algorithm. An associated illustration is given in Fig. 5.

2.10.3 *Image Segmentation*

Automatic detection and segmentation of salient regions in an image is an important computer vision task. An efficient image seg-

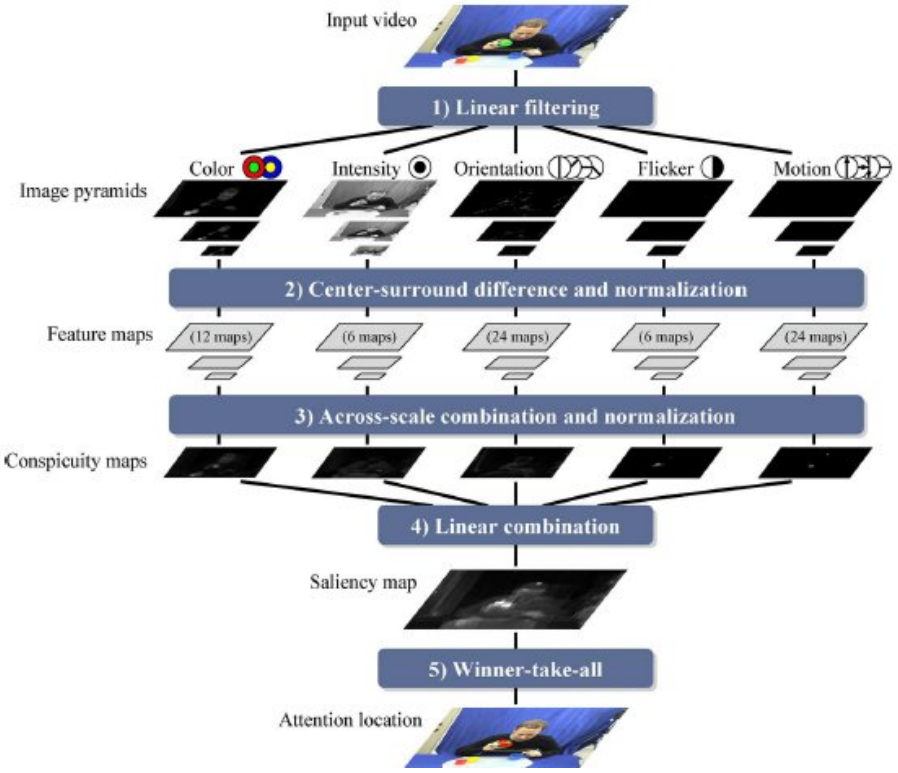


Figure 4: The architecture of the social attention model proposed in [81]. The model resembles the architecture of Itti et al. [48], but integrates retinal and stochastic filtering to predict eye-gaze fixations.

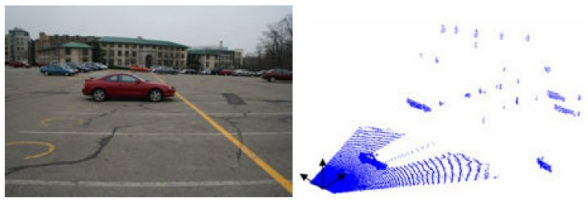
mentation has implications for digital photography, image resizing, thumbnailing and other computer vision tasks. Image segmentation has application to robotics, where it helps the robot to focus on specific a region of a scene. In terms of driver assistance systems for cars, segmentation helps in identifying obstacles, lanes, traffic lights etc. An application of saliency models for scene segmentation is given in Fig. 6

2.11 SUMMARY

In this chapter, we briefly reviewed many of the existing models for visual saliency. Majority of these models emphasize on a large feature set and computationally complex strategies. The attention modules developed so far are specialized in nature. They are either used to predict eye-gaze in a free viewing condition or are designed to handle searching and recognition task, but not both concurrently. The absence of a unified saliency system which can handle both of these aspects leads to additional software engineering effort to make a robot work in real time.

Since the saliency systems are computationally complex, they downsize the input image into a lower scale. This process may sometime

Figure 5: A saliency driven color conversion mechanism presented in [6]. The left shows the original stimulus, and the middle one shows the gray scale color conversion due to interpolation. The right image shows the re-adjusted color conversion by considering the saliency of each pixel. Observe that the saliency driven color conversion is visually more appealing than the normal grayscale based conversion.



(a) Input: an image and its corresponding 3-D scan data

(b) Output: saliency clustering and segmentation

Figure 6: A saliency driven 3D segmentation mechanism presented in [54]. The results shown in this work corroborates our assumption that integrating visual saliency mechanism improves the image segmentation algorithms.

eliminate smaller objects and finer details that are visible at the original scale. It can also be observed in the literature that centre-surround contrast plays a pivotal role in directing attention in a free viewing condition. Not surprisingly, many of the successful models of visual saliency are driven by centre-surround contrast algorithms. One of the advantages of this mechanism is that it is computationally less expensive and hence the input image need not be down-scaled to a lower resolution.

Another important factor that influences the saliency system is the number of tunable parameters present in it. Cross validation is necessary to fine tune the parameters. This introduces a dataset bias wherein the generic nature of the saliency system is compromised. A few high performing saliency systems (like [39, 16, 15]) require training bases. This introduces a classifier bias where the saliency system works more appropriately on the images which resemble the ones in

the training corpus. Any improvements on one or more of these issues will positively improve the performance of saliency systems. As we have already mentioned that the saliency models have a plethora of computer vision applications. The enhancements would thus cascade into the other applications, thereby enhancing the performance and user experience.

3 BOTTOM-UP ATTENTION MODELS

In the literature review we explained several important bottom-up saliency models with their strengths and short-comings. Based on these discussions one could sum-up the important properties of an ideal bottom-up saliency system as:

1. Generate full resolution saliency map
2. Minimal set of tunable parameters
3. Simple to program
4. Absence of a training corpus
5. Minimal set of feature maps
6. Highlight the objects uniformly and not just the boundaries
7. Perform well on both eye-gaze correlation and salient region detection tasks
8. Computationally efficient

Most of the saliency models process the pixels in a sequential and grid-like fashion. But in reality, the human vision processes purposeful spatial positions on a visual scene and does not follow a serial or sequential method. The existing models use a plethora of simple and compound image features to compute saliency maps. But the saliency systems which employ center-surround contrast features are the ones which are more biologically plausible. Based on these observations we thus propose three different approaches to compute bottom-up saliency maps.

The proposed saliency maps are described from Section. 3.1 to Section. 3.3. The associated experiments and results are explained in Section. 3.4. The chapters ends with a brief conclusing note in Section. 3.5.

3.1 RANDOM PIXELS BASED SALIENCY (PR1)

Nosofsky [83] proposed that each stimulus is influenced by every other stimulus present in the attention space. A stimulus is thus attenuated or boosted as a result of this interaction. The ensuing interactions contributes to the final cumulated salience. A majority of the stimulus interaction models have two components, the first one being the similarity function and second one being the biasing function.

Based on this paradigm we propose an interaction formulation as in Eq. 1.

Let $\mathbf{I}$ be an image of dimension $r \times c$. Let (x_i, y_i) and (x_j, y_j) be two distinct co-ordinate positions in $\mathbf{I}$. The corresponding intensity values are given by I_{x_i,y_i} and I_{x_j,y_j} respectively. The attention value $(V(\mathbf{I}, x_i, y_i, x_j, y_j))$ resulting out of the interaction between $I(x_i, y_i)$ and $I(x_j, y_j)$ is given by their gradient normalized by the Euclidean distance between them. This can be formulated mathematically as in Eq. 1.

$$V(\mathbf{I}, x_i, y_i, x_j, y_j) = \frac{\left|I_{x_i,y_i} - I_{x_j,y_j}\right|}{\sqrt{(x_i - x_j)^2 + (y_i - y_j)^2 + 1}} \tag{1}$$

Our input is an image and we do not consider patches or image segments as stimuli but rather consider each pixel. This helps in solving the issue of patch or grid size for feature extraction as in the case of other saliency models like [94, 16, 129, 53, 56]. This is also in line with the center-surround contrast paradigm, except that we do not restrict to a pre-specified radius to the choose pixels from. Thus we propose to randomly generate a set of n_p random co-ordinates which act as stimulus keypoints. We study the interaction between the stimulus (pixel intensity value) that is present in each of these keypoints and a stimulus that is present in another set of n_p random co-ordinates in the image. The algorithm **Random_Pixel_Saliency** describes the proposed formulation.

However, a saliency model is expected to operate on color images and not just gray scale versions. We therefore recommend to convert the input color image into $L^*a^*b^*$ color space, and further operate on the decomposed L^*, a^*, and b^* channels separately. We recommend to use $L^*a^*b^*$[1] color space as it preserves the perceptual difference between the colors in the Euclidean space. The obtained channel specific saliency maps are fused by employing Euclidean norm to generate the master saliency map. We are motivated to use this fusion rule as it is non-parametric and also does not violate the metric properties. Usage of Euclidean norm as a fusion rule has also been recommended by Achanta et al. [3, 4]. A Gaussian filter is used as a pre-processor to remove spurious spikes and noise present in the input image. Furthermore, a median filter is used as a post-processor to propagate the saliency values across the neighboring pixels. We specifically chose median filter for this purpose, as it has the property of blurring the image without suppressing the edges and the boundaries. This property helps the final saliency map to highlight both regions and boundaries simultaneously. The complete framework of the proposed scheme which works on a color image is thus given in the algorithm **PR1**.

1 The original CIE document - http://www.electropedia.org/iev/iev.nsf/display?openform&ievref=845-03-56

Algorithm I.(a) : **Random_Pixel_Saliency**

Input : (1) $\mathbf{I}$ (Grayscale Image) of size $r \times c$

: (2) n_p - Number of random pixels

Output : $\mathbf{S}^T$ - Component specific saliency map of size $r \times c$

Method

Step 1 : Set all elements of $\mathbf{S}^T$ to 0

Step 2 : Update $\mathbf{S}^T$

for i= 1 to n_p

x_i = Random number in$[1, r]$

y_i = Random number in$[1, c]$

for j= 1 to n_p

x_j = Random number in$[1, r]$

y_j = Random number in$[1, c]$

$S^T_{x_j,y_j} = S^T_{x_j,y_j} + V(\mathbf{I}, x_i, y_i, x_j, y_j)$

end-j

end-i

Algorithm I : **PR1**

Input : (1) $\mathbf{I}_{RGB}$ (RGB Image) of size $r \times c \times 3$

: (2) n_p - Number of random pixels

Output : $\mathbf{S}$ - Saliency map of size $r \times c$

Method

Step 1 : Apply Gaussian filter on $\mathbf{I}_{RGB}$

Step 2 : Convert input $\mathbf{I}_{RGB}$ to $L^*a^*b^*$ space

Step 3 : Generate saliencies for each component

$\mathbf{S}^L$ = Random_Pixel_Saliency$(\mathbf{L}^*, n_p)$

$\mathbf{S}^a$ = Random_Pixel_Saliency$(\mathbf{a}^*, n_p)$

$\mathbf{S}^b$ = Random_Pixel_Saliency$(\mathbf{b}^*, n_p)$

Step 4 : Compute $\mathbf{S}$ by pixel-wise Euclidean norm

$\mathbf{S}$ = Fusion$(\mathbf{S}^L, \mathbf{S}^a, \mathbf{S}^b)$

Step 5 : Apply median filter on $\mathbf{S}$

Step 6 : Normalize $\mathbf{S}$ in [0, 255]

Algorithm I.(b) : Fusion

Input : (1) Matrices **A**, **B** and **C** of size $r \times c$

Output : (1) Fused matrix **F** of size $r \times c$

Method

Step 1 : Set all elements of **F** to 0

for i= 1 to r

for j= 1 to c

$$F_{i,j} = \sqrt{A_{i,j}^2 + B_{i,j}^2 + C_{i,j}^2}$$

end-j

end-i

3.2 RANDOM RECTANGULAR SUB-WINDOW BASED SALIENCY (PR2)

The algorithm **PR1** models saliency as the cumulated contrast between random pair of pixels. It assumes that the input image is noise filtered and smooth, otherwise the spikes are detected as salient. In addition, the method is not scale invariant as it uses the Euclidean distance to normalize the gradients between two random pixels. Euclidean distance is an absolute measure, and is not a relative entity like aspect ratio which remains constant despite a uniform scale change of an image. In order to address these two issues, we reformulate saliency computation as sampling random sub-windows from an image and cumulating the local saliencies to obtain the master saliency map. The random scale and location of a sub-window overcomes the need for a pre-determined grid size as in the case of [94, 129].

We sample n_r random sub-windows over **I**. The upper left and the lower right co-ordinates of the i^{th} random sub-window is denoted by (x_{1i}, y_{1i}) and (x_{2i}, y_{2i}) respectively. The saliency value at a particular co-ordinate position is defined as the sum of the absolute differences of the pixel intensity value to the mean intensity values of the random sub-windows in which it is contained. The algorithm **Random_Rectangular_Sub_Window_Saliency** describes the proposed formulation.

Algorithm II.(a) : **Random_Rectangular_Sub_Window_Saliency**

Input : (1) $\mathbf{I}$ (Grayscale Image) of size $r \times c$

: (2) n_r - Number of random sub-windows

: (3) Co-ordinate vectors $\mathbf{x_1}$, $\mathbf{y_1}$, $\mathbf{x_2}$, $\mathbf{y_2}$ each of size n_r

Output : $\mathbf{S}^T$ - Component specific saliency map of size $r \times c$

Method

Step 1 : Set all elements of $\mathbf{S}^T$ to 0

Step 2 : Update $\mathbf{S}^T$

for i= 1 to n_r

$Area_i = (x_{2i} - x_{1i} + 1) \cdot (y_{2i} - y_{1i} + 1)$

$Sum_i = 0$

for j= x_{1i} to x_{2i}

for k= y_{1i} to y_{2i}

$Sum_i = Sum_i + I_{j,k}$

end-k

end-j

$\mu_i = \frac{Sum_i}{Area_i}$

for j= x_{1i} to x_{2i}

for k= y_{1i} to y_{2i}

$S^T_{j,k} = S^T_{j,k} + \left| I_{j,k} - \mu_i \right|$

end-k

end-j

end-i

The issue of noise affects **PR1** because the region of support is a pixel and not a patch centered on it. By sampling patches instead of pixels we solve the issue of spikes being highlighted as salient. We do not fix a scale or position for the sampled patches or regions of interest. This is because, that we would like to maximally enclose a salient object or a region in an image. Further, salient regions or objects can occur at arbitrary positions, shapes and scales in an image. To compute the local saliency of a patch we recommend to compute pixel divergence, where all the pixels in the patch are replaced by their absolute differences between the mean pixel intensity value of the patch. The saliency value indicated at a pixel position in the master saliency map is nothing but the cumulated sum of the computed local saliency values. The aforementioned idea holds good for a two dimensional gray scale image. As recommended in **PR1** we convert

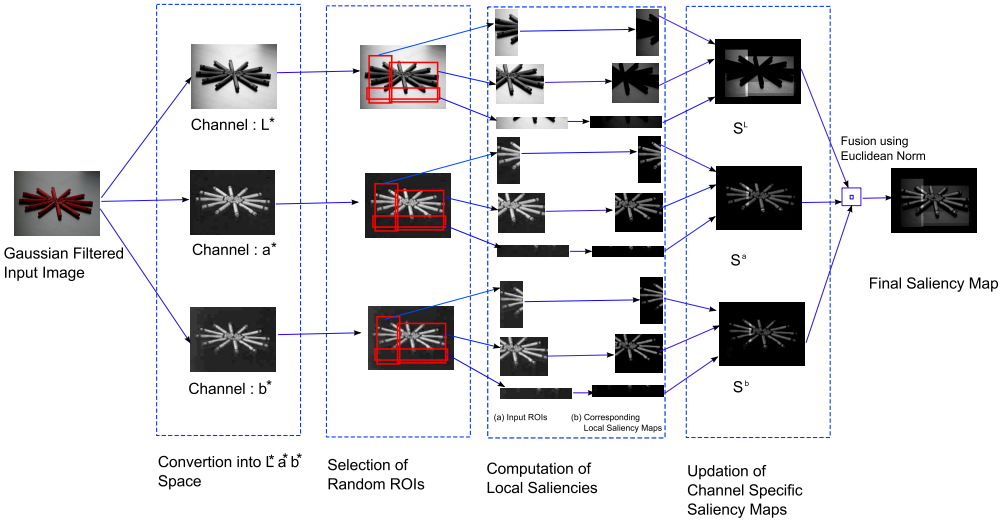


Figure 7: An illustration of the **PR2** approach. The input image is subjected to Gaussian filter in the first stage. Subsequently it is converted into the $L^*a^*b^*$ space and the individual L^*, a^* and b^* channels are obtained. For the sake of simplicity we have considered three random regions of interest (ROI) on the respective L^*, a^* and b^* channels. Local saliencies are computed over each of these ROIs and the channel specific saliency maps ($\mathbf{S}^L$, $\mathbf{S}^a$ and $\mathbf{S}^b$) are updated. The final saliency map is then computed by fusing the channel specific saliency maps by a pixel-wise Euclidean norm.

an input color image into $L^*a^*b^*$ space and decompose them into L^*, a^*, and b^* and apply **Random_Sub_Window_Saliency** algorithm on each of these channels separately. We obtain the master saliency map by fusing the channel specific saliency map by means of Euclidean norm. The complete framework of the proposed scheme which works on a color image is thus given in the algorithm **PR2**. A graphical illustration of the **PR2** approach is given in Fig. 7.

Algorithm II : **PR2**

Input : (1) $\mathbf{I}_{RGB}$ (RGB Image) of size $r \times c \times 3$

: (2) n_r - Number of random sub-windows

Output : S - Saliency map of size $r \times c$

Method

Step 1 : Apply Gaussian filter on $\mathbf{I}_{RGB}$

Step 2 : Convert input $\mathbf{I}_{RGB}$ to $L^*a^*b^*$ space

Step 3 : Generate random window co-ordinates

$$[\mathbf{x}_1, \mathbf{y}_1, \mathbf{x}_2, \mathbf{y}_2] = \text{Generate\_Random\_Sub\_Windows}(n_r, 1, 1, r, c)$$

Step 4 : Generate saliencies for each component

$$\mathbf{S}^L = \text{Random\_Rectangular\_Sub\_Window\_Saliency}(\mathbf{L}^*, n_r, \mathbf{x}_1, \mathbf{y}_1, \mathbf{x}_2, \mathbf{y}_2)$$

$$\mathbf{S}^a = \text{Random\_Rectangular\_Sub\_Window\_Saliency}(\mathbf{a}^*, n_r, \mathbf{x}_1, \mathbf{y}_1, \mathbf{x}_2, \mathbf{y}_2)$$

$$\mathbf{S}^b = \text{Random\_Rectangular\_Sub\_Window\_Saliency}(\mathbf{b}^*, n_r, \mathbf{x}_1, \mathbf{y}_1, \mathbf{x}_2, \mathbf{y}_2)$$

Step 5 : Compute $\mathbf{S}$ by pixel-wise Euclidean norm

$$\mathbf{S} = \text{Fusion}(\mathbf{S}^L, \mathbf{S}^a, \mathbf{S}^b)$$

Step 6 : Apply median filter on $\mathbf{S}$

Step 7 : Normalize $\mathbf{S}$ in [0,255]

Algorithm II.(b) : Generate_Random_Sub_Windows

Input : (1) n_r - Number of random sub-windows

: (2) l_x - Lower x co-ordinate

: (3) l_y - Lower y co-ordinate

: (2) u_x - Upper x co-ordinate

: (3) u_y - Upper y co-ordinate

Output : $\mathbf{x}_1, \mathbf{y}_1, \mathbf{x}_2, \mathbf{y}_2$ each of length n_r

Method

Step 1 : Generate random sub-window co-ordinates

Set all elements of $\mathbf{x}_1, \mathbf{y}_1, \mathbf{x}_2, \mathbf{y}_2$ to 0

for i= 1 to n_r

x_{1i} = Random number in $[l_x, u_x - 1]$

y_{1i} = Random number in $[l_y, u_y - 1]$

x_{2i} = Random number in $[x_{1i} + 1, u_x]$

y_{2i} = Random number in $[y_{1i} + 1, u_y]$

end-i

3.3 RANDOM FIXATION BASED SALIENCY (PR3)

In the algorithm **PR2**, we proposed to compute saliency over random rectangular patches. We implicitly assumed that the centroid of each patch was a region where the eye-gaze of a human fixates, and the saliency is a result of the divergence between mean patch intensity to a given pixel intensity. This helps us in overcoming the problem of fixing the grid size apriori, as this is taken care by the random sampling of sub-windows. The saliency of given pixel in **PR2** remains invariant to the spatial arrangement of other pixels in the patch. However, the perceptive mechanism in our visual system is sensitive to the spatial arrangement.

In order address these issues, we propose to sample square shaped random sub-windows instead of rectangles as in **PR2**. Computing circular patches is computationally expensive, and hence we decided to sample square shaped patches instead. Secondly, we use patch centroid intensity instead of mean pixel intensity value as a contrasting factor. The mean pixel intensity value remains invariant to the spatial arrangement of the pixels, while the patch centroid is generally sensitive to the spatial arrangement. We sample n_f random square sub-windows are over $\mathbf{I}$. The upper left and the lower right co-ordinates of the i^{th} random square sub-window centered on the random fixation co-ordinate (fx_i, fy_i) is denoted by (x_{1i}, y_{1i}) and (x_{2i}, y_{2i}) respectively. The saliency value at a particular co-ordinate position is defined as the sum of the absolute differences of the pixel intensity value to the patch centroid intensities of the random square sub-windows in which it is contained. The algorithm **Random_Square_Window_Saliency** describes the proposed formulation. A graphical illustration of **PR3** is given in Fig. 8.

Algorithm III.(a) : **Random_Square_Window_Saliency**

Input : (1) $\mathbf{I}$ (Grayscale Image) of size $r \times c$

: (2) n_f - Number of random fixations

: (3) $\mathbf{x_1}$, $\mathbf{y_1}$, $\mathbf{x_2}$, $\mathbf{y_2}$, **fx, fy** each of size n_f

Output : $\mathbf{S}^T$ - Component specific saliency map of size $r \times c$

Method

Step 1 : Set all elements of $\mathbf{S}^T$ to 0

Step 2 : Update $\mathbf{S}^T$

for i= 1 to n_f

for j= x_{1i} to x_{2i}

for k= y_{1i} to y_{2i}

$S^T_{j,k} = S^T_{j,k} + \left|I_{j,k} - I_{fx_i, fy_i}\right|$

end-k

end-j

end-i

We retain the same framework as proposed in **PR1** and **PR2** with regard to processing color images. Thus the complete framework is given in the algorithm **PR3**.

Algorithm III : **PR3**

Input : (1) $\mathbf{I}_{RGB}$ (RGB Image) of size $r \times c \times 3$

: (2) n_f - Number of random fixations

Output : $\mathbf{S}$ - Saliency map of size $r \times c$

Method

Step 1 : Apply Gaussian filter on $\mathbf{I}_{RGB}$

Step 2 : Convert input $\mathbf{I}_{RGB}$ to $L^*a^*b^*$ space

Step 3 : Generate random square window co-ordinates

$[\mathbf{x_1}, \mathbf{y_1}, \mathbf{x_2}, \mathbf{y_2}, \mathbf{fx}, \mathbf{fy}] = \text{Generate\_Square\_Windows}(n_f, r, c)$

Step 4 : Generate saliencies for each component

$\mathbf{S}^L = \text{Random\_Square\_Window\_Saliency}(\mathbf{L}^*, n_f, \mathbf{x_1}, \mathbf{y_1}, \mathbf{x_2}, \mathbf{y_2}, \mathbf{fx}, \mathbf{fy})$

$\mathbf{S}^a = \text{Random\_Square\_Window\_Saliency}(\mathbf{a}^*, n_f, \mathbf{x_1}, \mathbf{y_1}, \mathbf{x_2}, \mathbf{y_2}, \mathbf{fx}, \mathbf{fy})$

$\mathbf{S}^b = \text{Random\_Square\_Window\_Saliency}(\mathbf{b}^*, n_f, \mathbf{x_1}, \mathbf{y_1}, \mathbf{x_2}, \mathbf{y_2}, \mathbf{fx}, \mathbf{fy})$

Step 5 : Compute $\mathbf{S}$ by pixel-wise Euclidean norm

$\mathbf{S} = \text{Fusion}(\mathbf{S}^L, \mathbf{S}^a, \mathbf{S}^b)$

Step 6 : Apply median filter on $\mathbf{S}$

Step 7 : Normalize $\mathbf{S}$ in [0, 255]

Algorithm III.(b) : Generate_Square_Windows

Input : (1) n_f - Number of random fixations

: (2) r - Number of rows

: (3) c - Number of columns

Output : $\mathbf{x_1}$, $\mathbf{y_1}$, $\mathbf{x_2}$, $\mathbf{y_2}$, **fx**, **fy** each of size n_f

Method

Step 1 : Set all elements of $\mathbf{x_1}$, $\mathbf{y_1}$, $\mathbf{x_2}$, $\mathbf{y_2}$, **fx**, **fy** to 0

Step 2 : Generate random fixation window co-ordinates

$i = 1$

while $i \leqslant n_f$

$s = (\text{Random number in } [0, 100]) \cdot 0.01$

$\Delta_i = \lceil s \cdot (r + c) \cdot 0.5 \rceil$

if$(r - \Delta_i > \Delta_i)$ and $(c - \Delta_i > \Delta_i)$

$fx_i = \text{Random number in } [\Delta_i, r - \Delta_i]$

$fy_i = \text{Random number in } [\Delta_i, c - \Delta_i]$

$x_{1i} = fx_i - \Delta_i$

$y_{1i} = fy_i - \Delta_i$

$x_{2i} = fx_i + \Delta_i$

$y_{2i} = fy_i + \Delta_i$

$i = i + 1$

end-if

end-while

3.4 EXPERIMENTAL RESULTS

Experiments were conducted to validate the performance of the proposed saliency maps for two distinct tasks of salient region detection and eye-gaze prediction in free viewing conditions. Salient region detection and eye-gaze prediction are the two most significant applications of saliency maps. Salient region detection is relevant in the context of computer vision tasks like object detection, object localization and object tracking in videos [3, 4]. Automatic prediction of eye-gaze is important in the context of image aesthetics, image quality assessment, human-robot interaction and other tasks which involve detecting image regions that are semantically interesting [51]. The contemporary saliency maps are either employed to detect salient regions as in the case of [3, 88], or are used to predict eye-gaze patterns as in [129, 94]. Only few of the existing saliency approaches like [39, 48] have consistent performance on both of these tasks. Although these

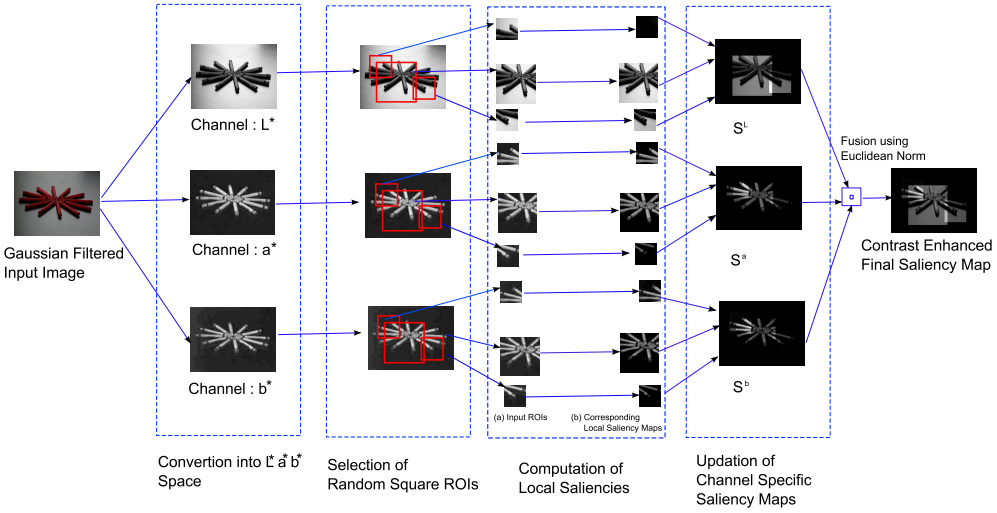


Figure 8: An illustration of the **PR3** approach. Please not that the control flow is similar to that of **PR2** approach as shown in Fig. 7. It can observed that the sampled patches are square shaped, while the patches in **PR2** are not restricted to be square shaped.

two tasks appear similar, there are subtle differences between them. Salient regions of an image are those which are visually interesting. But human eye-gaze which focuses mainly on salient regions are also distracted by semantically relevant regions [89]. The performance on the eye-gaze prediction task were validated on two different datasets from York University [16] and MIT [51]. The experiments to corroborate the performance on salient region detection task were conducted on the popular MSRA dataset [68].

Additional experiments were conducted on video snapshots from Bielefeld Motionese corpus [111]. The Motionese corpus consists of video recordings of parent–child tutoring of manipulative actions. More information about this dataset is presented in Section. 3.4.6. Unlike salient region detection and eye-gaze fixation datasets, the Motionese video snapshots are driven by top-down influence. We are also interested in knowing if the current saliency systems can predict eye-gaze in an interaction or tutoring scenario. This helps selecting and improving those saliency systems that are more appropriate for this task, thus making the human–robot interaction more coherent.

The following parameter settings were used as a standard for all the experiments carried out using **PR1, PR2, PR3** approaches. A rotational symmetric Gaussian low pass filter (size 3×3 with $\sigma = 0.5$; the default Matlab configuration) was used as a pre-processor on the images for noise removal as recommended in [3]. A median filter of size 11×11 was employed to smooth the resultant saliency map. We chose this size of the median filter as it resulted in a stable performance. The parameter n_p was set to $0.1 \cdot r \cdot c$, as the number of pixels or is correlated to input image size. The parameters n_r and n_f

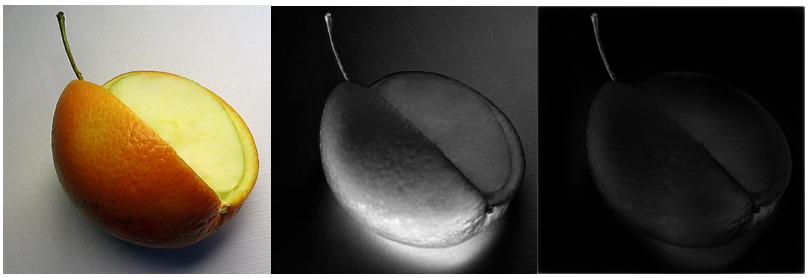
were set to 500 and 1000 respectively as they delivered stable performance. The employed Gaussian filter and median filter are based on the usual straight forward methods. All experiments were conducted using Matlab v7.10.0 (R2010a), on an Intel Core 2 Duo processor with Ubuntu 10.04.1 LTS (Lucid Lynx) as operating system. The inbuilt `srgb2lab` Matlab routine was used to convert the input image from RGB colorspace to $L^*a^*b^*$ colorspace. This results in L^*,a^* and b^* images whose pixel intensity values are normalized in the range of [0, 255]. We selected eight state-of-the-art methods of computing saliency maps to contrast with the proposed methods and shall be referred to as follows - **AC09** [3][2], **BR05** [16][3], **HA07** [39][4] , **IT98** [48][5] , **SE09** [94][6] , **RO09** [88][7] , **ZH08** [129][8] and **AC10** [4][9].

The saliency model which has the best reported performance is HA07. The models AC09 and AC10 are successful global contrast based approaches. IT98 is the standard benchmark for all the existing saliency approaches. BR05 is an information maximization based approach, while ZH08 is an energy minimization based approach. RO09 and SE09 are both driven edges and gradients. In addition the said methods are highly cited.

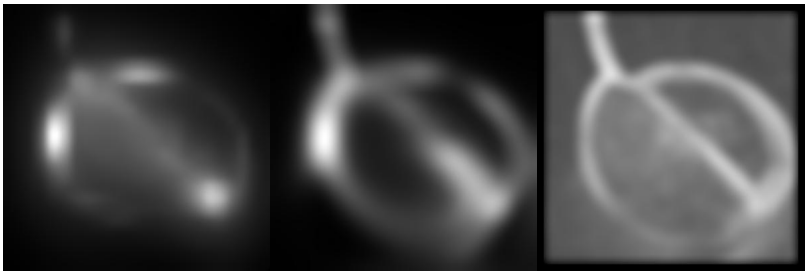
3.4.1 *Qualitative analysis*

We proceed to illustrate three examples where the saliency maps produced by the proposed saliency models are visually compared with those that are produced by other eight state-of-the-art methods in consideration. The visual illustrations are given from Fig. 9 to Fig. 11. We considered images where there is single object, multiple objects in a natural scene and an image which is cluttered with line like objects. The qualitative analysis provides insight to the strengths and shortcomings of the existing state-of-the-art methods in consideration. It also corroborates that the proposed saliency models are effective in producing a saliency map which highlight both boundaries and regions simultaneously.

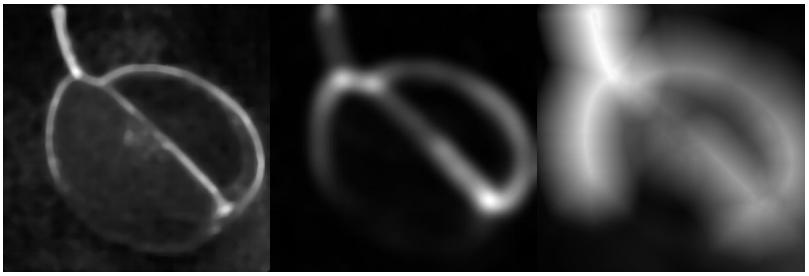
2 http://ivrg.epfl.ch/supplementary_material/RK_CVPR09/SourceCode/Saliency_CVPR2009.m
3 http://www-sop.inria.fr/members/Neil.Bruce/AIM.zip
4 http://www.klab.caltech.edu/~harel/share/gbvs.zip
5 http://www.klab.caltech.edu/~harel/share/simpsal.zip
6 http://users.soe.ucsc.edu/~rokaf/download.php
7 http://users.cs.cf.ac.uk/Paul.Rosin/resources/salience/salience.zip
8 http://cseweb.ucsd.edu/~l6zhang/code/imagesaliency.zip
9 http://ivrg.epfl.ch/supplementary_material/RK_ICIP2010/code/Saliency_MSSS_ICIP2010.m



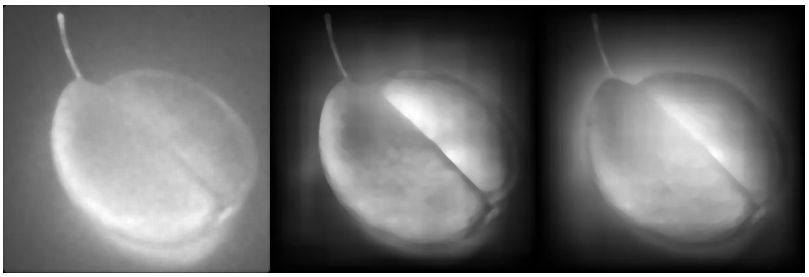
(a) Stimulus (b) AC09 (c) AC10



(d) HA07 (e) IT98 (f) BR05

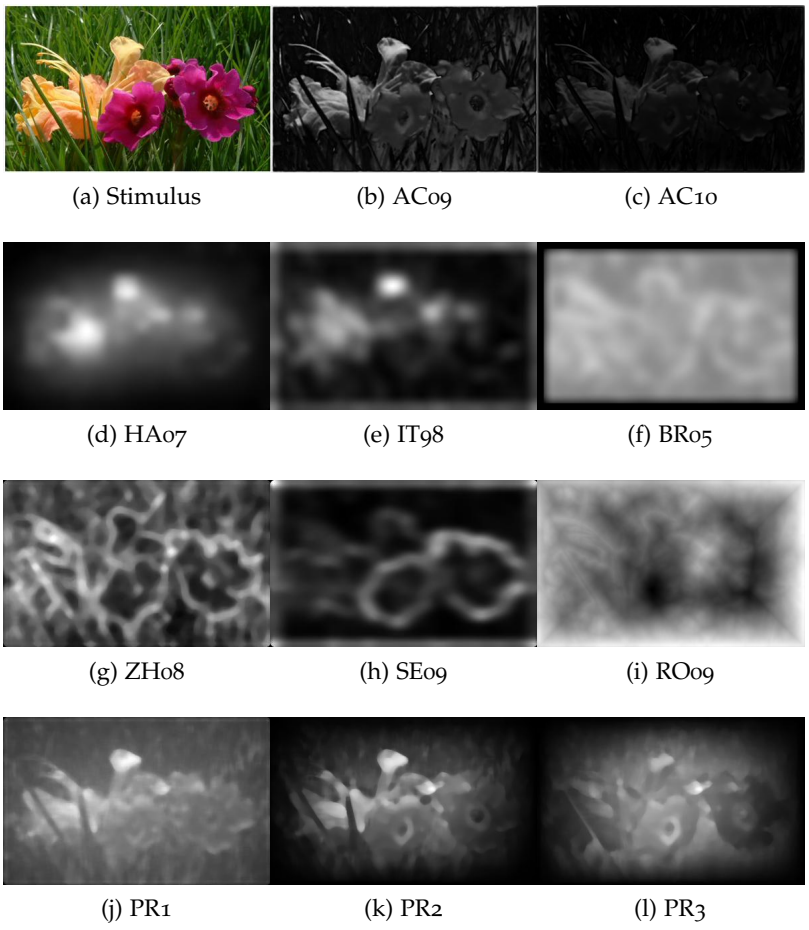


(g) ZH08 (h) SE09 (i) RO09



(j) PR1 (k) PR2 (l) PR3

Figure 9: Saliency maps obtained due to all the methods under consideration on Image 1_48_48173 from the MSRA dataset. Please note that the saliency maps shown in Fig. 9d to Fig. 9i emphasize on edges rather than regions. The saliency maps shown in Fig. 9b and Fig. 9c emphasize regions, but fail to highlight them uniformly. Only the proposed PR1, PR2 and PR3 based approaches produce a saliency map which can highlight both regions and boundaries uniformly.



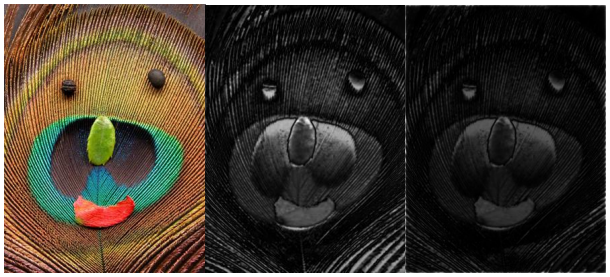
(a) Stimulus (b) AC09 (c) AC10

(d) HA07 (e) IT98 (f) BR05

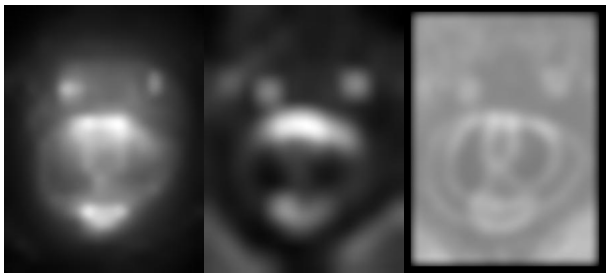
(g) ZH08 (h) SE09 (i) RO09

(j) PR1 (k) PR2 (l) PR3

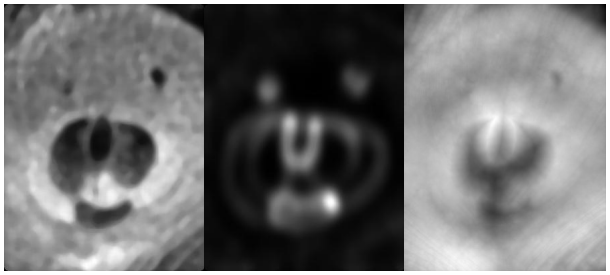
Figure 10: Saliency maps obtained due to all the methods under consideration on Image 0_21_21001 from the MSRA dataset. Unlike the stimulus in Fig. 9, the current image is from a natural scene. In addition it has multiple objects and a rich background. It can be observed from Fig. 10h that the saliency map generated by SE09 highlights only the boundaries. This can be explained by the fact that the said approach is based on gradient self-information of the image. Similarly the saliency map produced by RO09 as shown in Fig. 10i highlights the background instead of objects. The RO09 approach is based on distance transform, which is derived from edge detection at various thresholds. Since the background is cluttered with straight-line like edges, the method weights the background more than the foreground objects. The entropy based BR05, highlights the entire image as it finds all the regions visually interesting (shown in Fig. 10f). BR05 computes ICA bases from training images, and hence all the visually interesting patterns detected from the training base is highlighted. The saliency maps from AC09 and AC10 (Fig. 10b and Fig. 10c), are affected by global contrast of the image and hence fail to highlight the objects completely. As it can be seen from Fig. 10d, the HA07 method highlights all the objects effectively. However HA07 is parametric and computationally expensive as compared to rest of the methods in consideration. The saliency maps from PR1, PR2 and PR3 succeed in highlighting the regions despite a cluttered background.



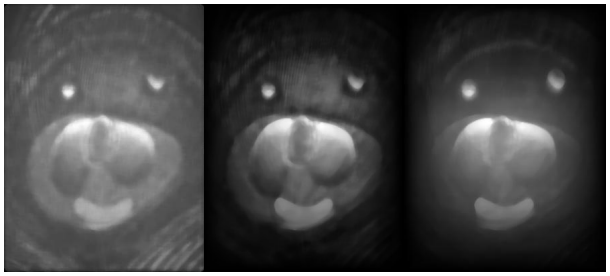
(a) Stimulus (b) AC09 (c) AC10



(d) HA07 (e) IT98 (f) BR05



(g) ZH08 (h) SE09 (i) RO09



(j) PR1 (k) PR2 (l) PR3

Figure 11: Saliency maps obtained due to all the methods under consideration on Image 1_38_38399 from the MSRA dataset. It can be observed that ZH08, RO09 and BR05 fail to highlight any of the visible blobs effectively (as seen from Fig. 11g, Fig. 11i and Fig. 11f). The presence of too many edges forces the distance transform based RO09 to highlight the background than the object blobs. ZH08 and BR05 have similar computational framework and hence fail to highlight the blobs, as edges are detected more saliently by the ICA bases. The HA07 and to a lesser extent IT98 (from Fig. 11d, Fig. 11e) highlight the blobs rather than the background. However HA07 is computationally expensive, and IT98 computes 41 feature maps to compute the master saliency map. The proposed models of PR1, PR2 and PR3 highlight the blobs effectively as compared to the other methods.

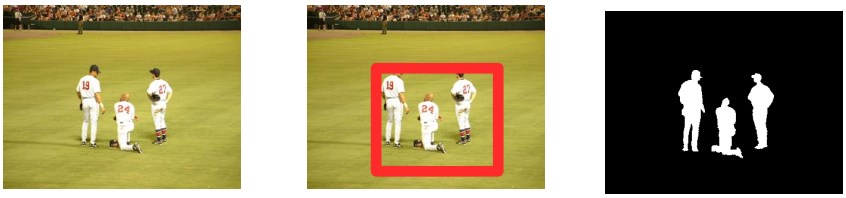


Figure 12: Rectangular and exact segmentation masks. To the left is the sample image from the MSRA dataset [68], the image in the center shows the original rectangular annotation and at right is the accurate-to-contour annotation. Observe the reduction in background information between rectangular and exact annotations.

3.4.2 *Experiments on salient region detection task*

Experiments were conducted on the MSRA dataset [68] in order to evaluate the performance on salient region detection task. The MSRA dataset [68] consists of 5000 images annotated by nine users. The annotators were asked to enclose what they thought was the most salient part of the image with a rectangle. Fig. 12 shows an example of this labeling. Naturally occurring objects do not necessarily have a regular contour which can be enclosed accurately inside a rectangle. As it can be seen from Fig. 12, unnecessary background is enclosed by this kind of annotation. It has been recently shown in [3, 4, 116] that a more precise-to-contour ground truth leads to a more accurate evaluation. Motivated by this a precise-to-contour ground truth was released in [3], for a subset of 1000 images from the MSRA dataset [68]. Thus we consider this subset of 1000 images which have accurate ground truth annotations for our experiments.

In general, the MSRA dataset [68] is significantly different from the eye-gaze datasets in terms of test protocol, content and image size. Fundamentally, the eye movements were not recorded and annotators were required to enclose the most visually interesting region of the image. Such an annotation task involves high-level cognitive mechanisms and is not stimulus driven. Thus there is a weak association between the exact segmentation masks and the saliency of the image. Despite involving a high-level visual task, the MSRA dataset [68] is still relevant to test whether there is an association between the predicted saliency map and the ground truth masks of this dataset. Previous studies [25, 60] have shown that the positions of the principal maxima in a saliency map are significantly correlated to the positions of areas that people would choose to put a label indicating a region of interest.

In order to quantitatively evaluate the performance for the task of detecting salient regions, we followed the method recommended in [3, 88, 4], where the saliency map is binarized and compared with

the ground truth mask. The saliency map is thresholded within [0, 255] to obtain a binary segmentation mask and is further compared with the ground truth. The thresholds are varied from 0 to 255 and recall-precision metrics are computed at each binarizing threshold. The recall (**R**) (also called true positive rate or hit-rate) and precision (**P**) metrics at a given threshold t, where $0 \leqslant t \leqslant 255$ is computed as:

$$R_t = \frac{tp_t}{tp_t + fn_t} \tag{2}$$

$$P_t = \frac{tp_t}{tp_t + fp_t} \tag{3}$$

where tp_t is the number of true positives, fp_t is the number of false positives and fn_t is the number of false negatives at a given threshold t.

In order to evaluate different algorithms both recall and precision at all the thresholds have to be considered simultaneously. This could however be achieved by a measure called *eleven-point average precision*, which takes into account both recall and precision. The eleven-point average precision (AP) [74] is computed as follows:

$$AP = \frac{1}{11} \cdot \sum_{i=0}^{i=i+0.1;i\leqslant 1} \max P_{TH(\mathbf{R},i)} \tag{4}$$

where $TH(\mathbf{R}, \theta)$ is a function which returns all the thresholds between [0, 255] where $\mathbf{R} \geqslant \theta$. AP being equal to 1 describes that the obtained segmentation mask over all the thresholds matches perfectly with the ground-truth mask, while AP being equal to 0 hints that the obtained segmentation masks does not match with ground-truth mask. However, it is unlikely that AP is either 0 or 1, but it assumes a value between them; and AP tending to 1 implies a better segmentation performance. The AP has been used as a benchmarking metric in TREC Video retrieval and PASCAL Visual object detection contests.

The histograms of the AP due to the eight state-of-the-art saliency models in consideration on the MSRA dataset [68] is given in Fig. 13. It can be observed from Fig. 13c and Fig. 13d that HA07 and IT98 have the best performance as compared to the other existing methods in consideration, where in HA07 two hundred and forty images have an AP $\geqslant$ 0.8. It can be seen in Fig. 13f that ZH08 is least effective in terms of accurately detecting the salient regions, as the histogram of APs resembles a uniform distribution. It also implies that the performance of ZH08 on task of detecting salient regions is random, and hence less reliable. Furthermore, the AP histograms of SE09 and RO09 (Fig. 13g and Fig. 13h) resemble a Gaussian distribution, where the majority of APs are 0.5. Despite their simplicity, the AC09 and AC10 (Fig. 13a

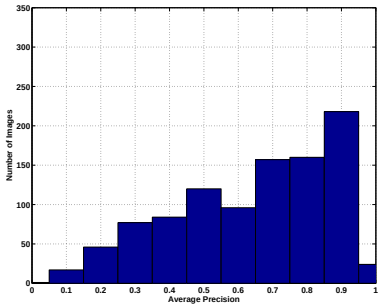
and Fig. 13b) have a better skew in the AP histogram as compared to the other sophisticated methods in consideration.

The AP histograms of the proposed PR1, PR2 and PR3 approaches are given in the Fig. 14 for better visualization. It can be observed from Fig. 14a that the PR1 approach attains the highest performance as compared to other methods in Fig. 13. It should be further noted that more than two hundred and fifty images have an AP $\geqslant 0.9$, and the corresponding AP histogram is skewed-left. The AP histogram of PR2 (Fig. 14b) resembles that of HA07 (Fig. 13c), while PR2 performance being slightly better than that of HA07 as it can be observed from the bin 0.7 in both of their corresponding histograms. PR3 (Fig. 14c) attains a good performance in terms of AP, but it fails to outperform most of the other methods in consideration. It achieves an equivalent performance to SE09, BR05 and ZH08 despite lacking equivalent sophistication.

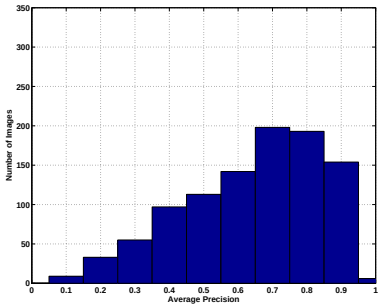
In order to consolidate the evaluation, we compute the mean APs from each of the methods. The resulting mean APs are plotted as a histogram and displayed in Fig. 15. As it can be seen in Fig. 15, PR1 and PR2 outperform all the eight existing state-of-the-art methods in consideration. In particular, PR1 achieves an AP of 0.8 while the best AP among the existing methods is from HA07 at 0.725. PR3 achieves a mean AP of 0.6, thereby outperforming ZH08, SE09 and RO09. PR2 attains an equivalent mean AP to that of HA07. This analysis reinforces the results from Fig. 13 and Fig. 14.

The histogram visualization of the average precisions sheds insight into the degree of effectiveness of each method in consideration. However, it might not be easy to visually compare two methods using the AP histograms as there is an overload of information. Histograms are sensitive to number, width and placement of bins and if the bin width changes within a histogram, the results can be misleading. We therefore use a cumulative distribution function (CDF) plot of the average precisions for better visualization. The CDF displays all the data, and thus portrays the distributions as precisely and completely as they can be known, given the available observations. In addition it does not require arbitrary choices of smoothing or binning parameters. The CDF plots also determine if one algorithm stochastically dominates another. The CDF plots of the APs of all the considered saliency models along with the proposed PR1, PR2, and PR3 approaches on the MSRA dataset [68] is given in Fig. 16. It should be noted that more the CDF plot is towards the left, the better the performance in terms of AP (as AP ranges between 0 to 1, and 1 indicates the best performance). It can be observed from Fig. 16 that PR1 and PR2 achieves state-of-the-art performance and PR3 outperforms ZH08, SE09 and RO09 in terms of AP.

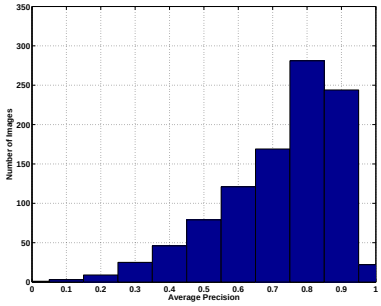
We further wanted to analyze if the proposed saliency models perform consistently on different scales of salient regions. The examina-



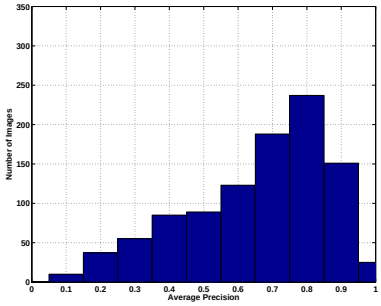
(a) AC09



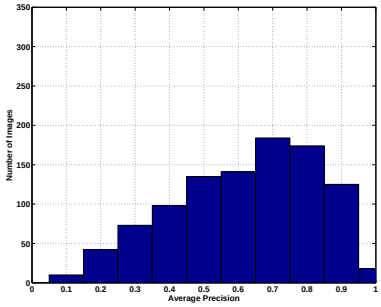
(b) AC10



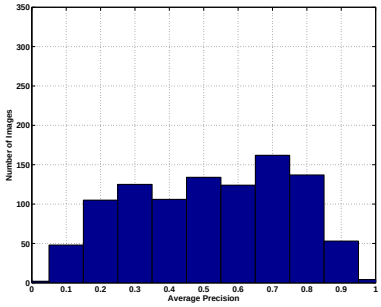
(c) HA07



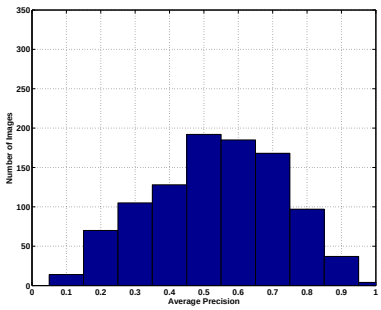
(d) IT98



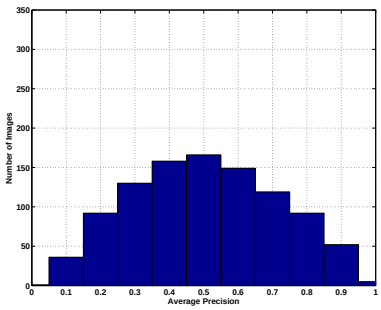
(e) BR05



(f) ZH08

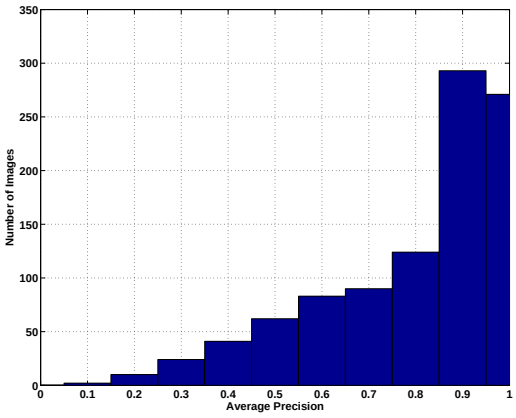


(g) SE09

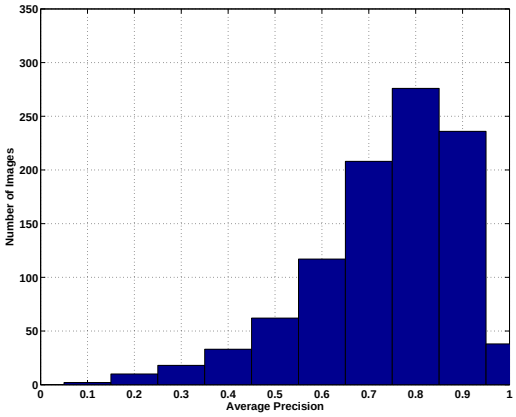


(h) RO09

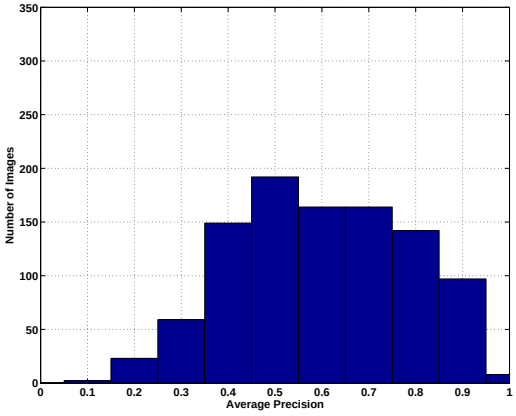
Figure 13: Average Precision of existing methods visualized in terms of histograms. The performance in terms of AP can be considered effective when there is a skew in the histogram and a majority of the samples lie in higher valued bins.



(a) PR1



(b) PR2



(c) PR3

Figure 14: Average Precision of the proposed methods visualized in terms of histograms. It can observed that PR1 attains the highest performance in terms of AP as compared to all the other methods in consideration.

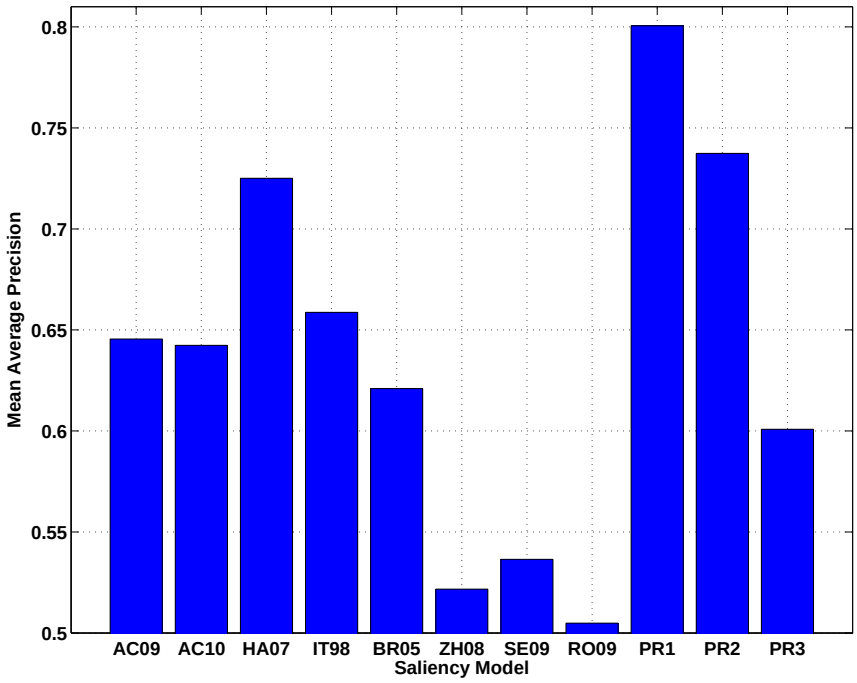


Figure 15: Mean of the Average Precisions obtained from all the methods under consideration. It can be observed that PR1 and PR2 achieve state-of-the-art performance, while PR3 outperforms the RO09, ZH08 and SE09 approaches.

tion of the considered MSRA dataset [68] images has revealed that the annotated salient regions occupied 20% to 30% of the image size in 685 of the 1000 stimulus images present. In about 160 images, the size of the salient regions is less than 10% of the image size, while in the remaining 155 images the size of the salient region is reported to be more than 30% of the image size. These subsets of 160 and 155 images constitute the tail of size distribution, and we further tested the performance of the proposed approaches on these subsets. The CDF plots of the APs on the subsets where the salient region sizes are lesser than 10% and greater 30% of the image size are reported in Fig. 17 and Fig. 18 respectively. From these plots we observe that the performance of PR1 does not diminish despite the size of salient region being extremely small or large. Furthermore, the performance of PR2 and PR3 improves on the subset where the size of the salient region is less than 10% of the image size (Fig. 17).

From the aforementioned average precision based analysis we can conclude that for any given binarizing threshold, the performance of PR1 and PR2 is guaranteed to outperform all the existing state-of-the-art methods on the MSRA dataset [68]. In this case, we assume the selection of the binarizing threshold is common and is done manually. However, a common binarizing threshold may not produce the optimal performance on salient region detection, as an optimal binarizing threshold is always specific to an image. Therefore it is necessary to adopt a method where a map specific binarizing threshold is computed automatically. Several image binarization techniques have been proposed so far, and we select the twelve most popular (in terms of citations) binarization techniques to compute the map specific binarization thresholds. The considered binarization techniques are the Concavity (**TH1**), Entropy (**TH2**), Intermeans (**TH3**), Iterative Intermeans (**TH4**), Intermodes (**TH5**), Maximum Likelihood (**TH6**), Mean (**TH7**), Median (**TH8**), Minimum Error (**TH9**), Minimum Error Likelihood (**TH10**), Minimum (**TH11**), and Moments (**TH12**) based thresholding. Interested readers are requested to consult reference [32] where all of the above mentioned binarization techniques are explained in detail.

In order to qualitatively compare the performance of map specific thresholding, we provide the resulting overlaid segmentation masks resulting from TH12 thresholding technique on three images (the same stimuli images from Fig. 9 to Fig. 11) from the MSRA dataset [68]. The overlaid segmentation masks can be viewed from Fig. 19 to Fig. 21. It can be observed from these three examples, that the proposed saliency models PR1, PR2 and PR3 generate a more accurate segmentation mask as compared to the other saliency models considered.

In order to quantitatively evaluate the segmentation performance, we employ the F-measure metric. Like the average precision, the F-measure also attempts to address the issue of convenience that is

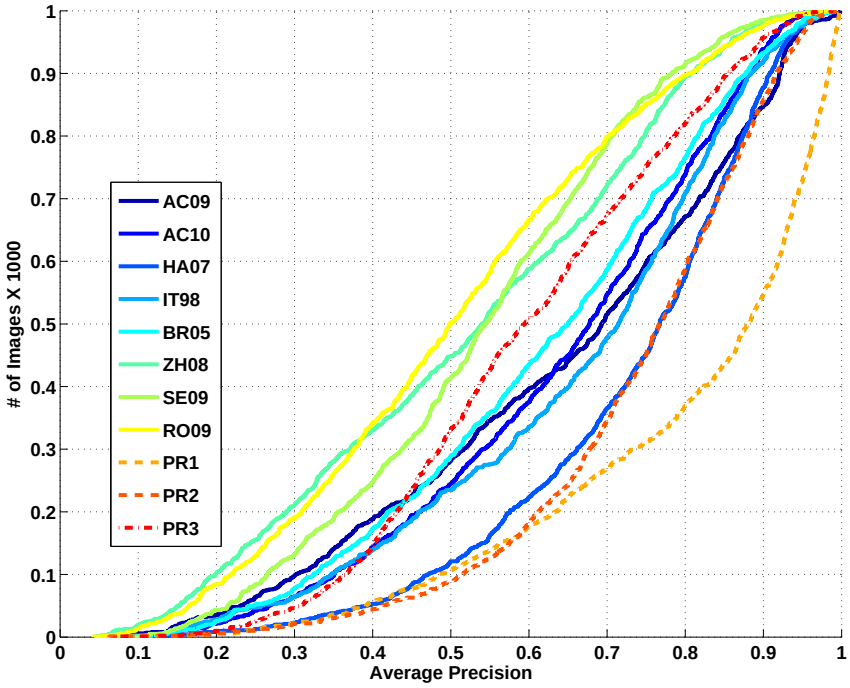


Figure 16: CDF plot visualization of the Average Precisions obtained from all the methods under consideration on MSRA dataset [68]. It can be observed that CDF plot of PR1 is left most and hence has the best performance interms of AP. It can also be seen at 0.5 of the Y-axis (which indicates the median), the AP of PR1 is 0.88. This implies that 50% of the saliency maps generated by PR1 have an AP > 0.88. Similarly it can seen from the CDF plots of PR2 that 50% of the saliency maps generated by this method have an AP > 0.76. The CDF plots of HA07 and PR2 are similar, however PR2 dominates HA07 throughout. Hence we can infer that PR2 has a tendency to perform better than HA07. PR3 outperforms SE09, RO09 and ZH08 saliency models despite lacking equivalent sophistication. The performance of ZH08 and BR05 is constrained by downscaling of the input image and the requirement of training bases. SE09 does not require training bases, but it requires a fixed grid size to process the image. Whenever the salient region or objects are not enclosed in the grid, the saliency detection fails. AC09 and AC10 attain a similar performance and outperform SE09, ZH08 and RO09 methods, thereby confirming the efficacy of the center-surround hypothesis in saliency computation.

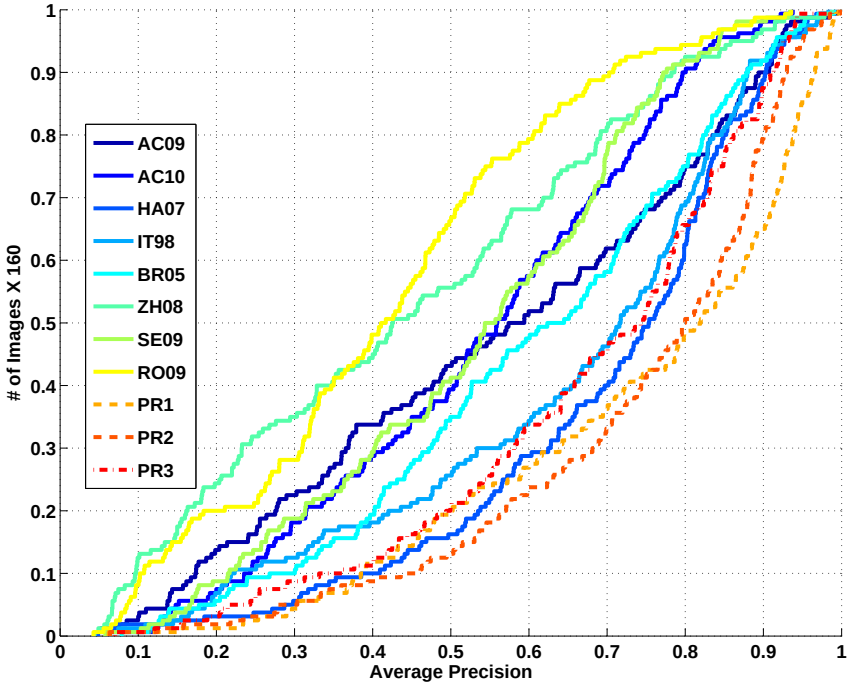


Figure 17: CDF plot visualization of the Average Precisions on those images whose ground-truth mask occupied less than 10% of the image size. It can be observed that PR1 achieves state-of-the-art performance, while PR2 which achieved an equivalent performance to HA07 on the entire dataset (as seen from Fig. 16) outperforms it. The performance of the PR3 is also high as it outperforms IT98 and attains a similar performance to HA07. While the performance of other methods deplete when the size of the salient regions decrease, the proposed (PR1, PR2 and PR3) approaches improve their performances. This capability helps the proposed approaches for being more complaint in applications like localizing small objects on a very large background.

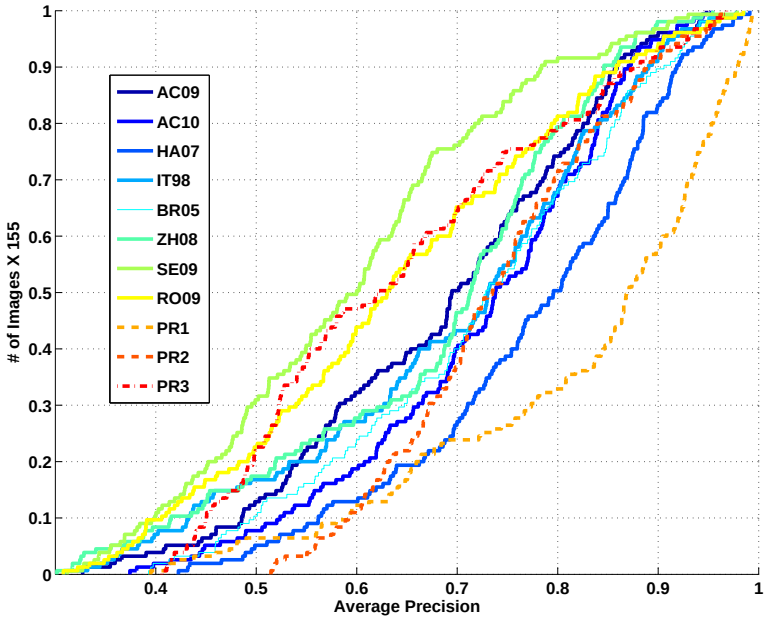


Figure 18: CDF plot visualization of the Average Precisions on those images whose ground-truth mask occupied more than 30% of the image size. It can be seen that CDF plot of PR1 dominates the rest and hence achieves the highest performance. PR2 which achieved a performance equivalent to HA07 on the entire dataset (as seen from Fig. 16), has a diminished performance on this image subset. Similarly PR3 is achieves a better performance than SE09. With these results, we observe that patch based methods (like SE09, PR2, PR3, BR05, ZH08, RO09) have a lower performance as compared to pixel based methods (like PR1, IT98, HA07, AC09, AC10).

brought on by a single metric than a pair of metrics. It combines precision and recall into a single metric. The formula for the F-measure (F_t) at a given binarizing threshold t is given as follows:

$$F_t = \frac{2 \cdot R_t \cdot P_t}{R_t + P_t} \tag{5}$$

High recall with low precision is easy to achieve as it means we highlight most of the regions that are not salient. Similarly, attaining high precision with low recall implies that most of the regions that are highlighted are salient, but a majority of the salient regions are missed. A salient region detection process can achieve either high recall or high precision, but rarely both simultaneously. An effort to improve the performance of either precision or recall causes the performance of the other to drop.

Ideally, we require a salient region detection performance which weighs high on both recall and precision. The F-measure thus attains a high value only when both recall and precision are equally weighted and not skewed towards either of them. A high F-measure is thus synonymous to a more accurate segmentation. In addition, the F-measure is also employed in information retrieval problems which has to deal with the case where negative class examples outnumber the positive examples significantly. Recent research has shown that F-measure has higher correlation with human judgments than the recently proposed alternatives [50].

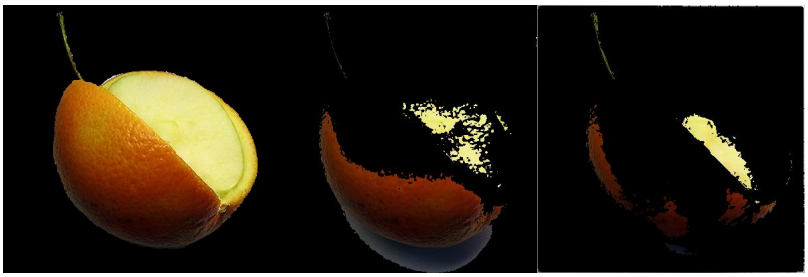
The corresponding recall, precision and F-measures over all the thresholding schemes on the eight state-of-the-art saliency models are given in Fig. 22. It can be observed from Fig. 22 that AC09 (Fig. 22a) and AC10 (Fig. 22b) attain a higher precision than recall on most of the different thresholding schemes. This implies that despite their limited ability to highlight the image regions, most of which what they highlight is salient and accurate. This result is in line with the segmentation illustrations shown in Fig. 19b and Fig. 19c. We have seen from the previous illustrations that the BR05 approach highlights most of the image for a given binarizing threshold. This leads to a high recall and a low precision performance in the salient region detection task. The same can be observed in Fig. 22e where the recall is significantly higher than the precision. Similarly it is also the case with RO09 (Fig. 22h). The F-measure performance of SE09 (Fig. 22g) and ZH08 (Fig. 22f) never exceed 0.5 on either of the twelve different thresholding schemes. This implies that these methods neither attain a high recall or high precision values and hence not suitable for the salient region detection task. The HA07 (Fig. 22c) and IT98 (Fig. 22d) have a more consistent performance as compared to the other methods in consideration. Furthermore, the histograms reveal that HA07 (Fig. 22c) and IT98 (Fig. 22d) have a higher F-measure performance on the T12 thresholding scheme as both recall and precision are equally weighted.

We now similarly illustrate the recall, precision and F-measure performances of the proposed PR1, PR2 and PR3 saliency approaches in Fig. 23. As it can be seen from the F-measure performance of PR1 (Fig. 23a), the said approach outperforms all the other methods in consideration. It should be noted that it achieves high performance on both recall and precision values on majority of the thresholding schemes. This indicates that the segmentation performance is both complete and accurate. This quality of completeness and accuracy was visualized in the segmentation illustrations (Fig. 19j, Fig. 20j, Fig. 21j). The F-measure performance of PR2 (Fig. 23b) is comparable with that of HA07 (Fig. 22c) despite PR2 being a patch based approach. In addition, HA07 is computationally the most expensive scheme, while the complexity of PR2 is indicated by n_r co-efficient which controls the number of image samplings. The segmentation performance of the PR3 (Fig. 23c) approach is similar to that of AC10 (Fig. 22b). In addition, it weighs high on recall than precision over all the thresholding schemes with an exception of T6 thresholding scheme. This indicates that the saliency map produced by PR3 has a low contrast. We further averaged the performance of recall, precision and F-measure due to T1 to T12 on all the saliency models. The averaged performances are thus presented in Fig. 24. This further corroborates that PR1 and PR2 approaches attains state-of-the-art performance in terms of F-measure metric.

3.4.3 *Experiments on eye-gaze prediction task*

An additional way to evaluate the performance of the saliency models is by corroborating their effectiveness in predicting the human eye-gaze on an image in free-viewing condition. In a free-viewing condition, a user is asked to view an image without any specific objective like searching or recognition of objects. The eye-gaze fixation positions on the stimulus image and their durations are further recorded using an eye-tracker. The recorded fixations are subsequently pooled to generate a fixation density map –which is a gray-scale image– where the pixel intensity values are proportional to the fixation duration on that location. The fixation density map is treated as ground-truth and the obtained saliency map on the stimulus image is compared and contrasted with it.

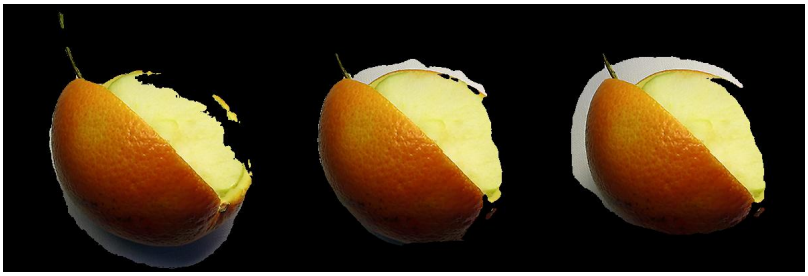
For the purpose of evaluation we have considered eye-gaze fixation datasets from the York University [16] and MIT [51]. The York University [16] dataset consists of 120 images, with eye-fixation recordings from 20 test participants. This dataset is the standard evaluation platform with regard to the eye-fixation correlation experiments. It consists of both indoor and outdoor scenes, and size of the test images are relative small and uniform. Another recent, but large and challenging eye-fixation dataset is available from the MIT [51]. The MIT



(a) Ground-Truth (b) AC09 (c) AC10

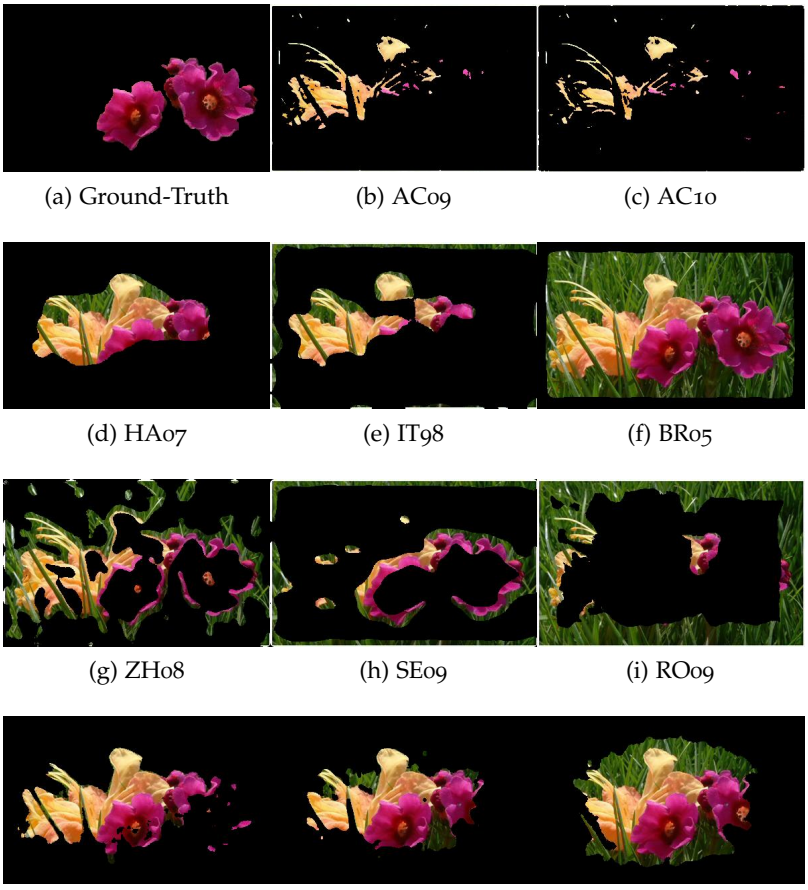
(d) HA07 (e) IT98 (f) BR05

(g) ZH08 (h) SE09 (i) RO09



(j) PR1 (k) PR2 (l) PR3

Figure 19: Overlaid segmentation masks obtained (due to TH12 binarizing scheme) for all the methods under consideration on Image 1_48_48173 from the MSRA dataset [68]. Observe that the proposed PR1, PR2 and PR3 approaches produce a segmentation mask, which is most similar to the ground-truth. ZH08, SE09 and IT98 end up highlighting the boundaries and edges and ignore salient regions. BR05 highlights the salient region, however it encloses redundant backgrounds. AC09 and AC10 fail to sufficiently highlight the salient regions, as the contrast of the stimulus image is insufficient for effective saliency detection by these two methods. HA07 highlights only the salient regions, but does not detect the entire expected regions of interest.



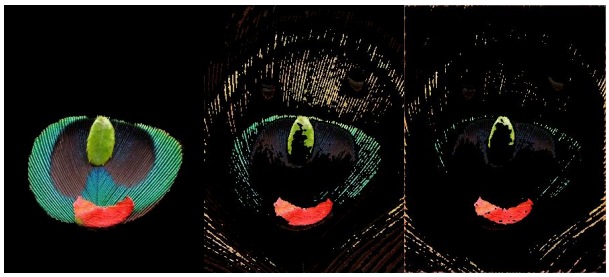
(a) Ground-Truth (b) AC09 (c) AC10

(d) HA07 (e) IT98 (f) BR05

(g) ZH08 (h) SE09 (i) RO09

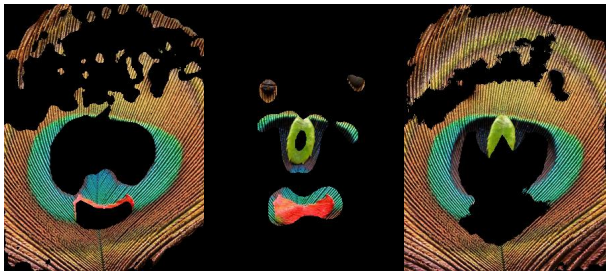
(j) PR1 (k) PR2 (l) PR3

Figure 20: Overlaid segmentation masks obtained (due to TH12 binarizing scheme) for all the methods under consideration on Image 0_21_21001 from the MSRA dataset [68]. The proposed PR1, PR2 and PR3 approaches highlight the salient regions more effectively as compared to the other state-of-the-art methods, though not completely accurate. It should be noted that ZH08, RO09 and SE09 highlights the background instead of the salient region, and is affected by strong edges and boundaries. AC09 and AC10 have a similar performance and highlight high contrast regions instead of salient regions. BR05 highlights the entire image, and hence may not be suitable for salient region detection in cluttered backgrounds. IT98 displays a strong tendency to highlight the boundaries as it can be seen that the borders of the image is highlighted. Similar to the proposed saliency approaches, HA07 smoothly highlights part of the salient region.



(a) Ground-Truth (b) AC09 (c) AC10

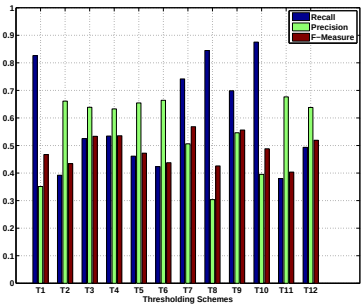
(d) HA07 (e) IT98 (f) BR05



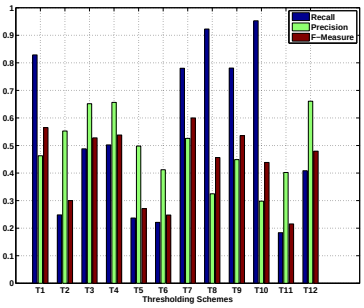
(g) ZH08 (h) SE09 (i) RO09

(j) PR1 (k) PR2 (l) PR3

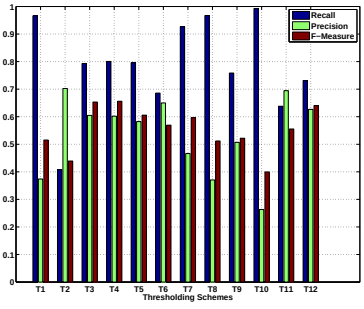
Figure 21: Overlaid segmentation masks obtained (due to TH12 binarizing scheme) for all the methods under consideration on Image 1_38_38399 from the MSRA dataset [68]. The proposed PR1, PR2 and PR3 methods produce a segmentation mask more appropriately than rest of the methods in consideration. The current image has a lot of boundaries, and it is hard to visually distinguish the foreground from the background. Despite this challenge, the proposed approaches perform well. It can be observed that ZH08 and RO09 highlight the line like segments than the intended salient region. Subsequently AC09 and AC10 fail to highlight the salient regions because of the lack in image color contrast. BR05 highlights the entire image as it did in the previous two examples (Fig. 19f and Fig. 20f).



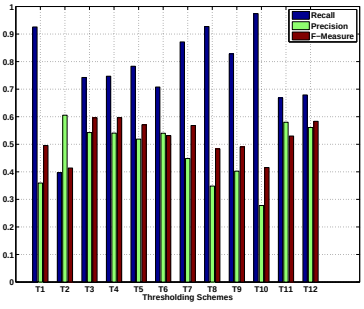
(a) AC09



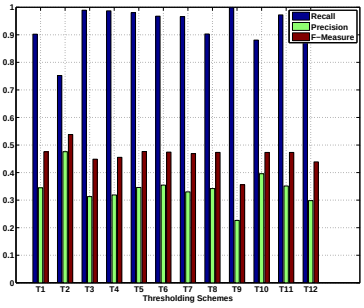
(b) AC10



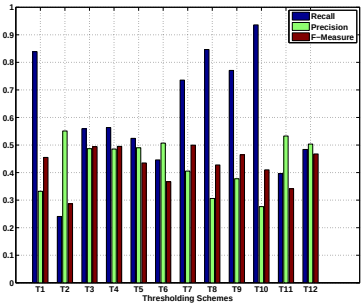
(c) HA07



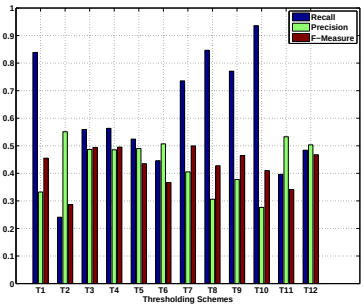
(d) IT98



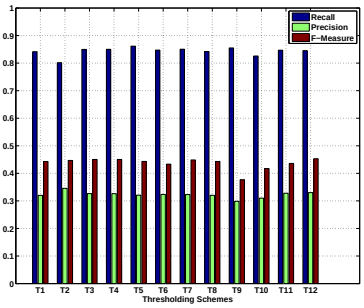
(e) BR05



(f) ZH08

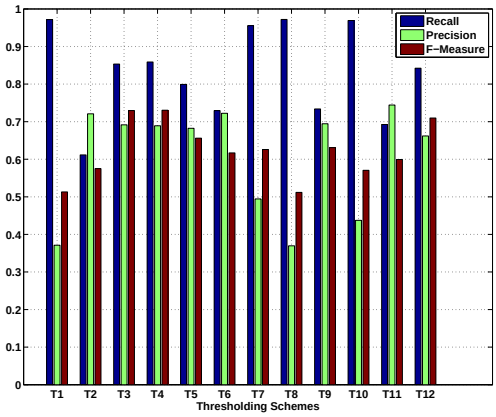


(g) SE09

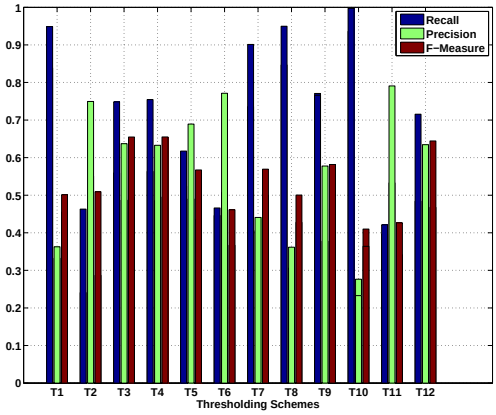


(h) RO09

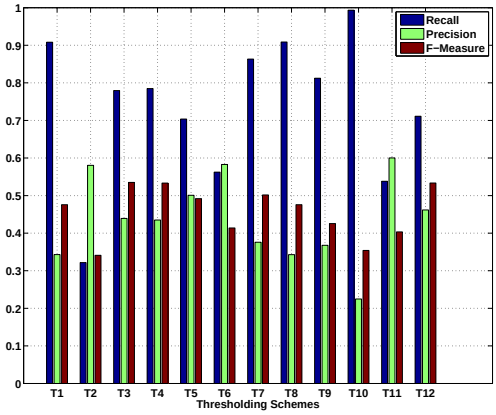
Figure 22: Recall, Precision and F-Measure from twelve different thresholding schemes over existing methods



(a) PR1



(b) PR2



(c) PR3

Figure 23: Recall, Precision and F-Measure from twelve different thresholding schemes over the proposed methods

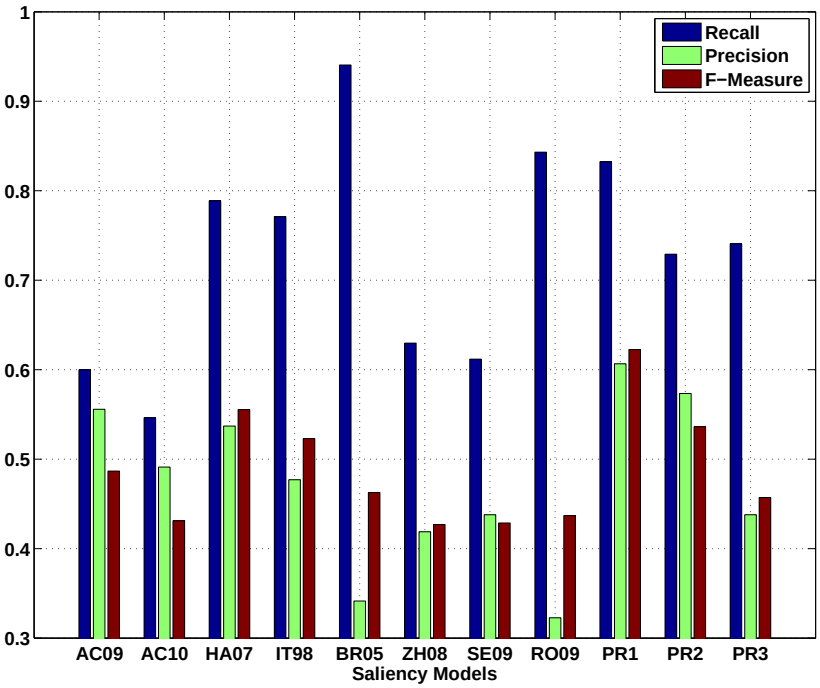


Figure 24: Average of the Recall, Precision and F-Measure obtained due to the twelve thresholding schemes over all the saliency methods under consideration. Please observe that PR1 and PR2 attain the highest F-measure performance, while PR3 outperforms ZH08, SE09, RO09 and AC10 saliency models. It should be noted that BR05 has a higher F-measure performance than PR3. But despite this, PR3 could be considered more reliable than BR05 as it has a higher precision than BR05. A very high recall and a low precision combination as in the case of BR05 could be found effective in detecting the salient regions only under those circumstances where the salient regions significantly occupy the majority of the image. This can be seen in the illustration (Fig. 19f) where the salient region occupies the majority of the image area. The same is also corroborated in the plot (Fig. 18) where the BR05 achieves a higher performance on those images of the MSRA dataset [68] where the salient regions occupied more than 30% of the image. It can be seen that ZH08 and SE09 have low performance in terms of F-measure and are thus unsuitable for salient region detection task. While the performances of IT98 and HA07 are moderately good as compared to the other six state-of-the-art methods, they are hindered by their computational complexity. AC09 also attains a low F-measure performance but could be considered a better alternative to BR05, ZH08, SE09 and RO09 because of its low computational complexity.

[51] dataset consists of 1003 images with fixation recordings from 15 test participants. The images consists of diverse scenes from parties, crowds, wildlife with varying camera angles. In addition the size of the images are large and not uniform throughout the dataset.

In order to empirically evaluate the performance on eye-gaze correlation task, we have employed the receiver operating characteristic (ROC) - area under the curve (AUC) as a benchmarking metric. Several popular and recent works like [39, 53, 60] employ the ROC-AUC metric to evaluate eye-gaze fixation correlation. An ROC graph is a general technique for visualizing, ranking and selecting classifiers based on their performance [50]. The ROC graphs are two-dimensional graphs in which the true positive rate (**TPR**) is plotted on the Y axis and the false positive rate (**FPR**) rate is plotted on the X axis. The TPR_t (also called hit rate and recall) and FPR_t (also called false alarm rate) metrics at a given binarizing threshold t is computed as in [67]:

$$\mathrm{TPR}_t = \frac{tp_t}{tp_t + fn_t} \tag{6}$$

$$\mathrm{FPR}_t = \frac{fp_t}{fp_t + tn_t} \tag{7}$$

where tn_t is the number of true negatives.

An ROC graph depicts relative trade-offs between benefits (**TPR**) and costs (**FPR**). Since the AUC is a portion of the area of a unit square, its value will always be between 0 and 1.0. An ideal classifier would give an AUC of 1.0 while random guessing produces an AUC of less than 0.5. The saliency map and the corresponding ground truth fixation density map are binarized at each discrete threshold in [0, 255]. This results in a predicted binary mask (from the saliency map) and a ground truth binary mask (from the fixation density map) for each binarizing threshold. The TPR_t and FPR_t for each threshold are subsequently computed. The ROC curve is generated by plotting the obtained **FPR**s versus **TPR**s and the AUC is calculated. In our case, the AUC indicates how well the saliency map predicts actual human eye fixations. In general, the AUC represents the performance of the classifier averaged over all possible cost ratios. It has been argued that AUC is one of the good methods to obtain a score of a classifier performance and to compare it with other classifiers as it works well in case of imbalanced data [50]. This property of AUC is suitable in this context as the number of fixation locations in an eye-gaze fixation map is always lesser than the number of unfixated locations. AUC also measures the probability of a classifier associating a higher rank to a randomly chosen positive example than to a randomly chosen negative example. Some of the statisticians also argue that ROC-AUC is equivalent to Wilcoxon's Rank Sum test. This

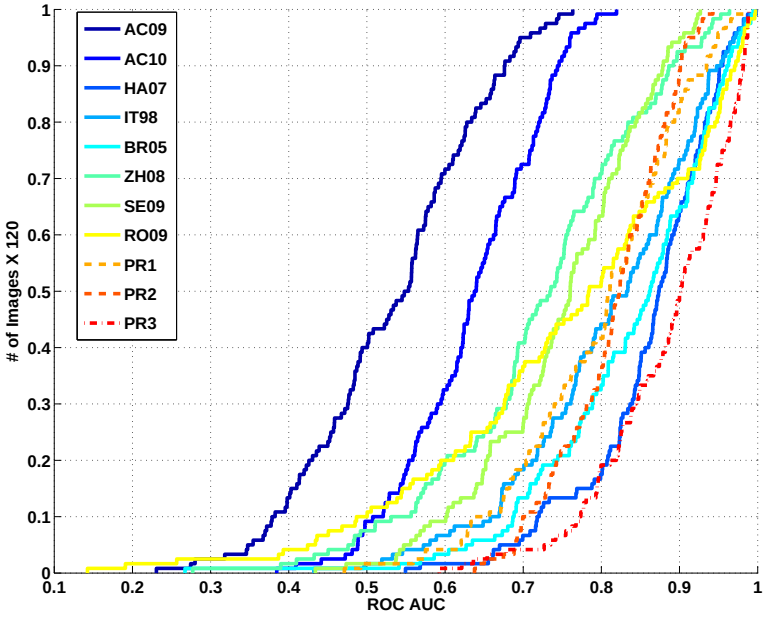


Figure 25: CDF Performance Plot for ROC-AUC on York University Dataset [16]. Note that PR3 attains the highest performance.

measurement also has the desired characteristic of transformation invariance, in that the ROC-AUC does not change when applying any monotonically increasing function (such as logarithm) to the saliency measure [129].

It can be observed from Fig. 25 that the performance of PR3 stochastically dominates rest of the approaches on York University dataset. PR1 and PR2 approaches outperform AC09, AC10, ZH08 and SE09 saliency models. PR3 attains state-of-the-art performance over rest of the computationally expensive approaches.

We can observe from the Fig. 26 that the PR3 approach attains an equivalent performance to HA07 approach. The MIT eye-fixation dataset [51] consist of 1003 images and is nine times bigger than the York University dataset [16]. Despite the high amount of variations in the number of images, and also the variations in the scenes, the performance of PR3 does not diminish. It can also be seen that PR1 outperforms AC09, AC10 and SE09, while PR2 outperforms RO09 in addition to these methods. The good performance of HA07 and BR05 can be attributed to the usage of precomputed training bases. However, the proposed approaches (PR1, PR2 and PR3) do not employ any training bases.

In addition to ROC-AUC, we also employed correlation co-efficient (CC) and mutual information (MI) metrics to evaluate the performances on eye-gaze prediction. We employ the CC metric to test for

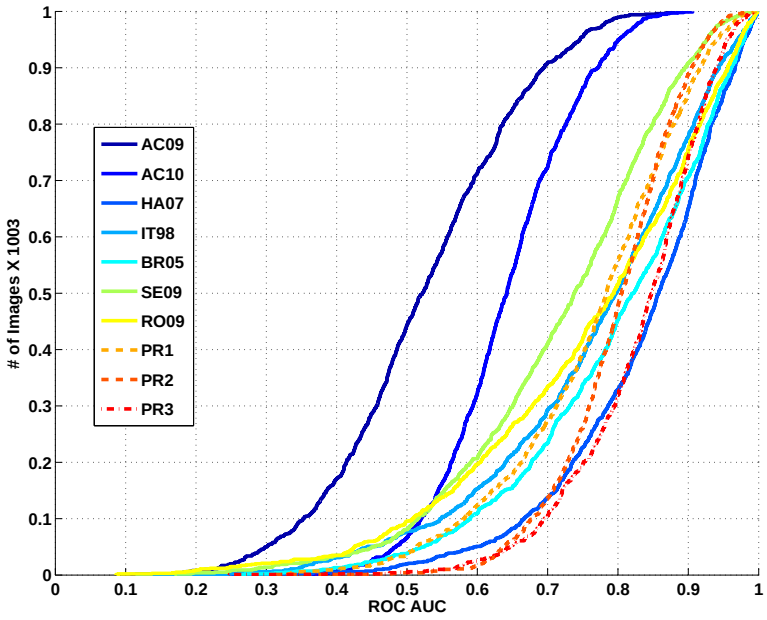


Figure 26: CDF Performance Plot for ROC-AUC on MIT Dataset [51]. Observe that PR3 and HA07 attain state-of-the-art performance.

the presence of outliers in the generated saliency maps. A higher CC implies less number of outliers and vice-versa. The presence of outliers generate false fixation locations while artificially simulating an eye scan path on an image. The CC between the saliency map **S** and fixation density map **F** (where the dimensions of **F, S** is $r \times c$) are computed as given in Eq. 8.

$$CC(\mathbf{F}, \mathbf{S}) = \frac{cov(\mathbf{F}, \mathbf{S})}{\sigma_F \cdot \sigma_S} \tag{8}$$

$$cov(\mathbf{F}, \mathbf{S}) = (rc) \sum_{i=1,j=1}^{i=r,j=c} (S_{i,j} F_{i,j}) - (\sum_{i=1,j=1}^{i=r,j=c} S_{i,j})(\sum_{i=1,j=1}^{i=r,j=c} F_{i,j}) \tag{9}$$

Eq. 9 refers to the co-variance while σ_F and σ_S refer to the variances of **F, S** receptively. The CC CDF performance plots on York University [16] and MIT [51] datasets are given in Fig. 27 and Fig. 28 respectively. We employ the MI metric to evaluate the similarity between the produced saliency maps and the eye-fixation maps. Unlike CC, the MI is not sensitive to outliers. A lower MI value denotes a higher amount

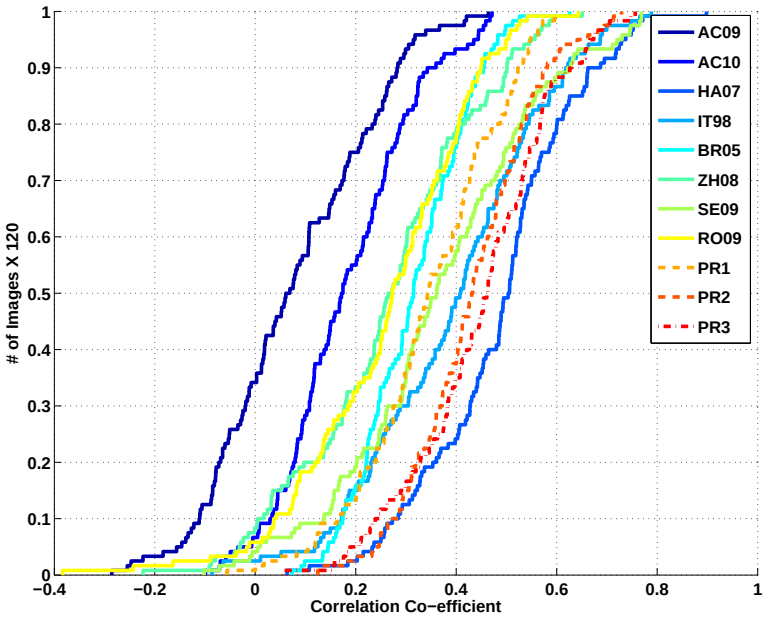


Figure 27: CC performance on York University [16] dataset. PR2 and PR3 attain a higher performance in terms of CC as compared to all methods with an exception of HA07.

of uncertainty to the predicted saliency values in the saliency map. The MI between **F, S** is computed as given in Eq. 10.

$$\mathrm{MI}(\mathbf{F},\mathbf{S}) = \sum_{i=0}^{i=255}\sum_{j=0}^{j=255} p_{i,j}^{\mathbf{F},\mathbf{S}} \log \frac{p_{i,j}^{\mathbf{F},\mathbf{S}}}{p_i^{\mathbf{F}} p_j^{\mathbf{S}}} \tag{10}$$

In Eq. 10 $p_i^{\mathbf{F}}$ and $p_j^{\mathbf{S}}$ represent the probability of the gray levels i and j in $\mathbf{F}$ and $\mathbf{S}$ respectively, while $p_{i,j}^{\mathbf{F},\mathbf{S}}$ denotes the joint probability of gray levels i and j in $\mathbf{F}$ and $\mathbf{S}$. The MI CDF performance plots on York University [16] and MIT [51] datasets are given in Fig. 29 and Fig. 30 respectively.

We also provide the averaged performances in terms of ROC-AUC, CC and MI on York University [16] and MIT [51] eye-gaze datasets in the Table. 1. The ROC-AUC metric evaluates the saliency models in terms of their classification performance. A higher performance in terms of ROC-AUC implies a higher classification accuracy of a saliency model. This is important because a saliency model predicts the fixation probability of a pixel position in an image. However, only a few pixel locations in an image are attended while the rest are ignored. This leads to a high imbalance between positive (fixated) and negative (non-fixated) examples. The ROC-AUC metric evaluates and ranks classifiers appropriately despite the imbalance in the classifica-

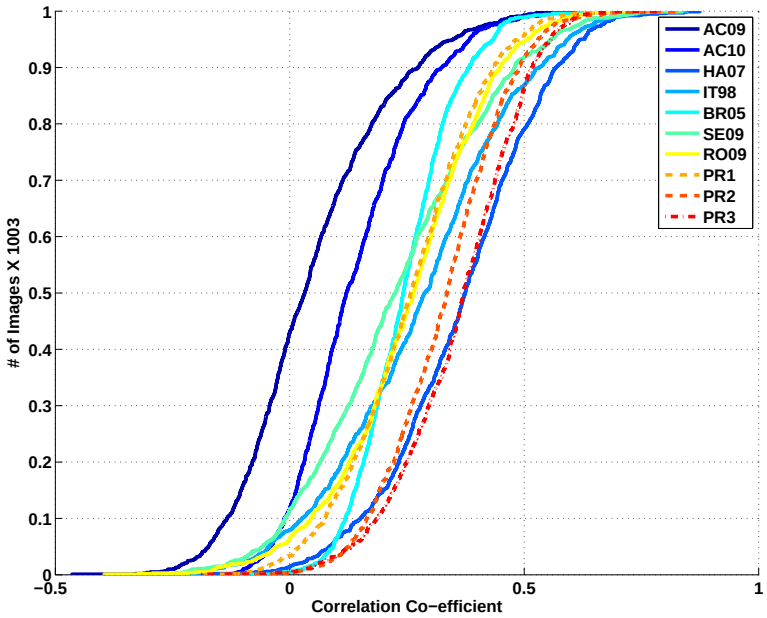


Figure 28: CC performance on MIT [51] dataset. The PR3 approach attains state-of-the-art performance equivalent to HA07.

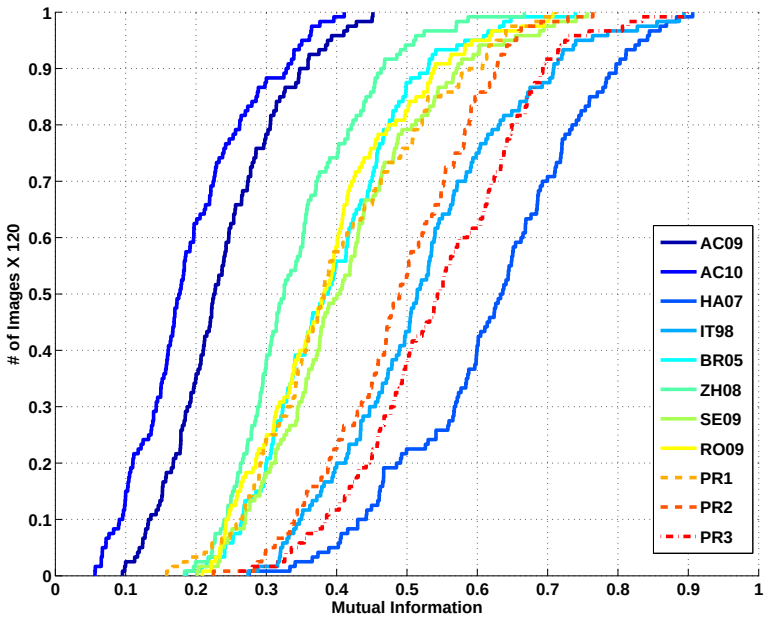


Figure 29: MI performance on York University [16] dataset. It should be observed that PR3 attains a higher performance on all methods except HA07.

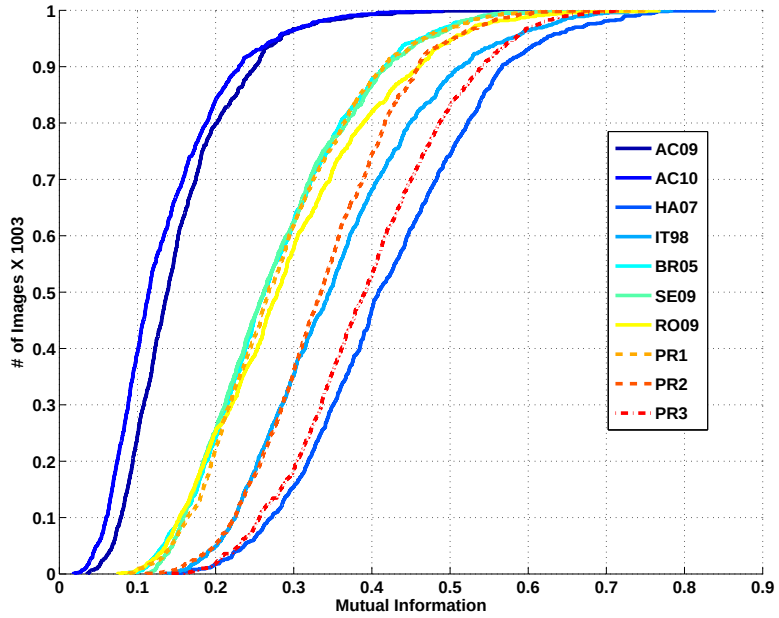


Figure 30: MI performance on MIT [51] dataset. It can be seen that PR3 attains a similar performance to that of HA07.

Table 1: Comparative analysis in terms of ROC-AUC, CC and MI.

	ROC-AUC		CC		MI	
Method	**York**	**MIT**	**York**	**MIT**	**York**	**MIT**
AC09	0.53±0.11	0.52±0.13	0.07±0.15	0.04±0.15	0.23±0.08	0.15±0.07
AC10	0.65±0.08	0.64±0.09	0.18±0.12	0.14±0.13	0.19±0.08	0.13±0.07
HA07	0.86±0.08	0.83±0.12	0.48±0.14	0.36±0.16	0.62±0.13	0.42±0.11
IT98	0.80±0.12	0.76±0.15	0.39± 0.18	0.28±0.19	0.52±0.13	0.35±0.11
BR05	0.83±0.11	0.79±0.14	0.31±0.10	0.24±0.10	0.39±0.10	0.27±0.10
ZH08	0.72±0.13	n/a	0.26±0.17	n/a	0.34±0.09	n/a
SE09	0.74±0.10	0.71±0.15	0.36±0.19	0.23±0.18	0.41±0.11	0.27±0.10
RO09	0.75±0.11	0.75±0.17	0.25±0.16	0.25±0.15	0.38±0.11	0.29±0.11
PR1	**0.79±0.10**	**0.76±0.12**	**0.34±0.13**	**0.25±0.14**	**0.40±0.12**	**0.28±0.09**
PR2	**0.81±0.07**	**0.78±0.08**	**0.43±0.12**	**0.32±0.12**	**0.48±0.10**	**0.33±0.09**
PR3	**0.88±0.08**	**0.82±0.09**	**0.44±0.14**	**0.36±0.13**	**0.54±0.12**	**0.39±0.10**

tion data as it considers TPRs and FPRs and not the absolute number of true positives and false positives.

It can be observed from Table. 1 that PR3 attains state-of-the-art performance on the York University dataset, while it outperforms all the existing methods with an exception of HA07 on MIT dataset. This can also be seen in the Fig. 25 and Fig. 26. The performance of AC09 implies that its performance is random. IT98 and BR05 achieve a good performance on both the eye-fixation datasets, but they are constrained by a large standard deviation in their performances. The performance of a method can be considered more precise, if the standard deviation of the associated performance metric is low. In that respect, the performance of PR2 is precise as it has the lowest standard deviation, and at the same time its averaged performance is next only to HA07. It can be observed that in terms of CC, the PR3 attains state-of-the-art performance on MIT dataset, while on York University dataset it is next only to HA07. A higher performance in terms of CC implies a lower number of outliers in the predicted saliency map. This property is useful while artificially simulating eye-gaze shifts on an image. This can also be seen in the Fig. 27 and Fig. 28. Please note (from Fig. 27 and Fig. 28) that the order of performance of BR05, SE09 and RO09 change with respect to York University and MIT eye-gaze fixation datasets. While the performance ordering of the proposed saliency approaches are consistent for both the data sets.

The standard deviation of the PR2 performance is amongst the lowest as it was the case with the ROC-AUC metric. We can observe that HA07 attains the highest performance in terms of MI. The performance of HA07 sharply depletes on the MIT dataset. However, the performance of PR2 and PR3 do not experience a sharp depletion as compared to HA07. Even on the MI metric, the standard deviation in the performance of PR2 is minimal. It can be observed from Fig. 29 that the performance of PR2 is similar to that of IT98, while the performance of PR1 is similar to that of RO09. Unlike correlation coefficient, mutual information is not sensitive to outliers and is used as an image similarity metric. From Fig. 30 we observe that the performance gap between HA07 and PR3 is reduced as compared to its performance on the York University dataset (Fig. 29). Similarly the performance ordering of PR2 and PR3 approaches do not change despite a different dataset. The results imply that PR3 and HA07 generate saliency maps which is visually more similar to fixation density maps.

3.4.4 *Performance due to change in parameters*

The performances of the proposed methods due to change in n_p, n_r and n_f on the MSRA dataset [68] are presented in Fig. 31. It can be observed from Fig. 31a that PR1 outperforms BR05, SE09, RO09 and ZH08 when n_p is set to as low as 200. This is an interesting result

Table 2: The computational run time(in seconds) of various saliency methods under consideration. Run times were computed using Matlab v7.10.0 (R2010a), on an Intel Core 2 Duo processor with Ubuntu 10.04.1 LTS (Lucid Lynx) as operating system

Saliency Map	**Original Code**	Runtime(in Secs) w.r.t Image Size		
		205×103	308×195	410×259
AC09	Matlab	0.0652	0.0898	0.1539
AC10	Matlab	0.0976	0.1172	0.2050
BR05	Matlab	2.1448	5.0761	10.3915
HA07	Matlab & C++	0.6256	0.4788	0.5577
IT98	Matlab & C++	0.4388	0.3820	0.3661
RO09	Binary	0.1266	0.2806	0.5308
SE09	Matlab	3.0590	3.1187	3.1133
ZH08	Matlab	1.6466	4.0714	7.6242
PR1	**Matlab**	**0.2701**	**0.5939**	**1.2038**
PR2	**Matlab**	**0.3430**	**0.5422**	**1.0766**
PR3	**Matlab**	**0.3600**	**0.3700**	**0.8000**

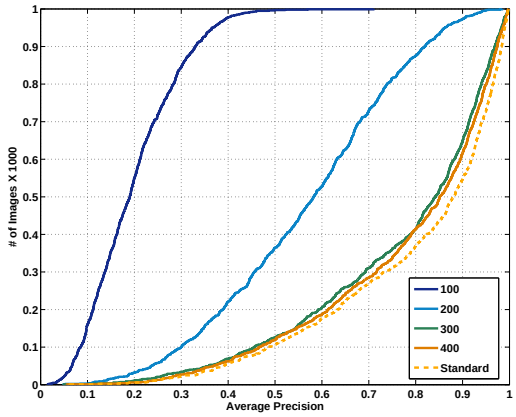
because it involves only forty thousand pixel operations. The MSRA dataset consist of images of size 400×300 pixels, and the saliency map obtained from PR1 when $n_p = 200$ is a result of operating on less than 30% of the pixels in the image. The performance of PR1 increases drastically when $n_p = 300$ and saturates quickly to the standard performance when $n_p = 400$.

We can observe that PR2 (Fig. 31b) nearly attains its top performance when $n_r = 50$. Methods like ZH08 and SE09 sample the image into more than hundred regular patches. Contrastingly PR2 achieves a high performance (even outperforming HA07 , please refer Fig. 16) when $n_r = 50$. The performance of PR2 quickly attains its peak performance for larger values of n_r.

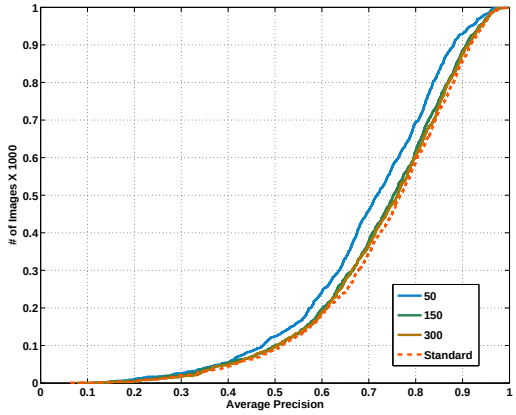
We can observe that when $n_f = 50$, PR3 (Fig. 31c) outperforms RO09, SE09 and ZH08. Unlike PR2, PR3 doesn't even involve computing the mean of the patch and hence its computationally inexpensive. Thus PR3 could be used in those circumstances where computational efficiency is more important than accuracy of the results. PR3 saturates when $n_f > 250$.

3.4.5 *Computational Run-Time*

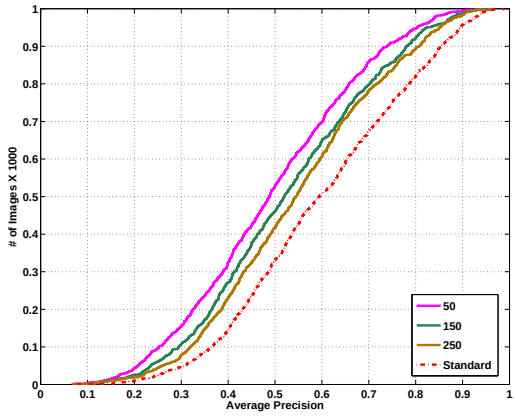
We evaluated the runtime of the proposed saliency approach with reference to the other methods in consideration. The runtime of the various methods were benchmarked on three different scales of a color



(a) PR1–n_p



(b) PR2–n_r



(c) PR3–n_f

Figure 31: CDF plots of the Average Precision performance variations with respect to different parameters.

image as shown in Table 2. The original plugins of AC09, AC10 and BR05 saliency approaches are pure Matlab codes. While the codes pertaining to IT98, HA07 and SE09 are quasi Matlab codes which call C++ functions for run time optimization. The original plugin for RO09 is a binary executable while its original coding language is unknown. The proposed PR1, PR2 and PR3 are programmed in Matlab. An absolute comparison on the basis of runtime might penalize the methods which are coded in Matlab script as they are relatively slower than their C++ counterparts. Nevertheless, it gives a relative overview of runtime performance of the all methods under consideration. It can be observed from Table 2, that the run time of HA07, IT98 and SE09 saliency approaches do not change significantly irrespective of the size of the input image. This is on account that the input image is rescaled to pre-specified dimension as mentioned in their source codes. The rest of the methods process the input image in its original scale and hence the run time changes with the input image dimension.

3.4.6 *Saliency Models for Eye-Gaze Prediction in an Interactive Scenario*

We further investigate the accuracy of the proposed saliency systems in predicting eye-gaze in an interaction scenario. Unlike salient region detection or eye-gaze prediction tasks in a free-viewing condition, the said scenario has top-down influences. The eye-gaze in an interaction scenario is focused not just on visually salient regions but on task relevant semantic regions. In this regard, we are interested to know the degree of effectiveness of the considered saliency systems on this scenario. Many researchers support saliency systems as a bottom-up inspired way to simulate infant-like gazing behavior [81]. The has implications for a cognitive humanoid robot as is often modeled in accordance with human behavior.

The experiments presented in this sub-section is based on video-recordings from the Bielefeld Motionese corpus. The corpus has sixty four pairs of parents who were asked to present a set of ten manipulative tasks their infants. During a task, the parent and child were facing each other while sitting across a table. This was videotaped with two cameras. Several coders have objectively annotated the objects gazed by the infants at any given time-stamp during the course of interaction. An example snapshot and the associated annotation example is given in Fig. 32.

For the current analysis we restrict on parent-infant-interaction during the stacking of cups task. The focus is on the youngest participants comprising 12 families of 8 to 11 months old children, as the main feedback and controlling capabilities of these infants is based on the gazing behavior. The said task consists of sequentially picking up a green, yellow, and a red cup and to subsequently place them

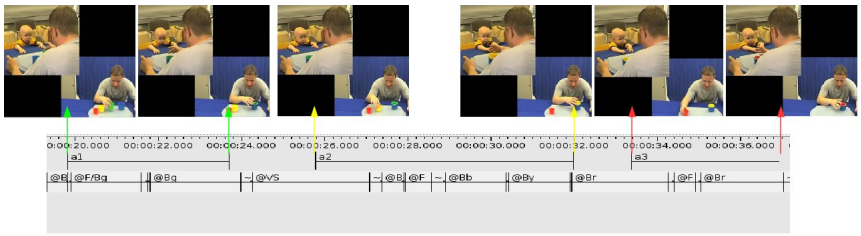


Figure 32: The annotation of Bielefeld Motionese Corpus [111]. The snapshot shows the video stills from both parent as well as the child's view. Please note that this snapshot is from the stacking cup scenario. The colored arrows indicate the color of the cup on which the child's attention is focused during that particular time-stamp of demonstration.

into a blue cup. We were motivated to choose this task, as it involves objects which have strong color saliency and hence could marginally compensate for the lack of top-down information comprehension by the saliency systems. Research carried out by Vollmer et al. [111] has suggested that the bottom-up influence is highest during beginning and ending action of an action sequence. Thus we chose twenty four images which represent the beginning and ending point of the stacking cups task from the considered twelve video recordings.

The test images were subsequently passed to various considered saliency system to produce the corresponding saliency maps. We later employed a WTA network as described in Itti et al.[48] to identify the most salient location in the saliency map. A 15×15 region centered on the maximally salient point was chosen as the focus of attention. The resulting salient regions due to the considered saliency systems are given in Fig. 33. Any overlap of this identified salient region, and child's object of attention was recorded as a success. The resulting child eye-gaze prediction accuracy is presented in Fig. 34. It be observed that the eye-gaze prediction performance of the bottom-up saliency systems is low in an interaction scenario. Eye-gaze shifts in a tutoring situation are a result of top-down influences like dialogue, pointing, gestures and other task related semantics; while saliency systems are driven by salient points, regions, corners and rare visual artifacts which completely ignore task based semantics.

3.5 DISCUSSION AND CONCLUSION

We proposed three different saliency systems which are based on randomized algorithms. Furthermore, we have conducted extensive evaluation on both salient region detection and eye-gaze prediction task. To the best of our knowledge, this is the largest evaluation which involved both of these tasks concurrently. We also show that the proposed models outperform HA07, which is the state-of-the art saliency system.

(a) Stimulus (b) AC09 (c) AC10 (d) BR05

(e) HA07 (f) IT98 (g) ZH08 (h) SE09

(i) RO09 (j) PR1 (k) PR2 (l) PR3

Figure 33: A snapshot from one of the test videos from the Bielefeld Motionese Corpus. The stimulus image is given in Fig. 33a. The child's attention is focused on the red cup according to the annotation. The square white patch is the focus of attention resulting from various saliency models by a WTA network is shown from Fig. 33b to Fig. 33l. It can be observed that only ZH08 and SE09 focus on the red cup while the rest of the saliency systems focus elsewhere.

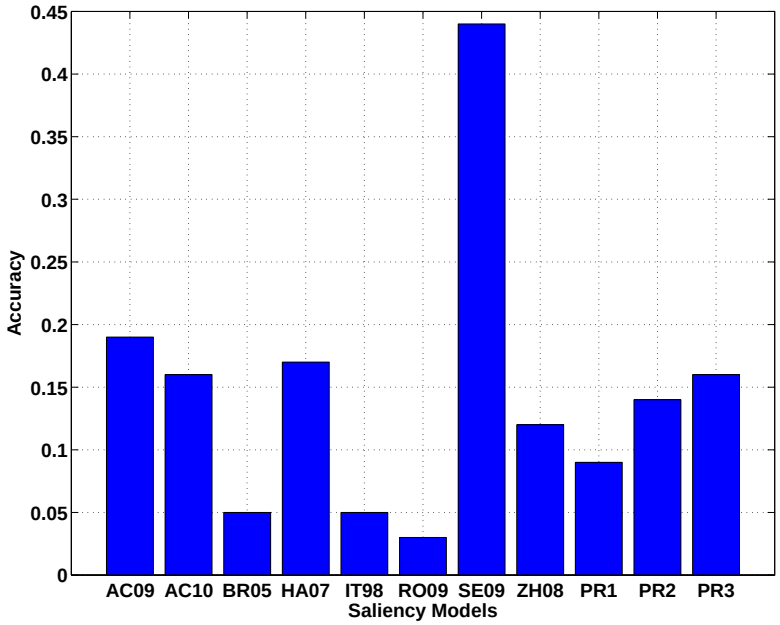


Figure 34: Accuracy in child eye-gaze prediction during interaction. It can be observed that SE09 attains the highest performance at 0.44, while the rest of the methods have an accuracy of less than 0.2

We have seen from the literature review that researchers advocate to employ a large set of different features to compute a saliency map. One plausible explanation to this tendency is that the human visual system employs more than one feature while processing the visual stimulus. This might lead to an improvement in performance, however reduces the computational efficiency. We have shown from our results that center-surround contrast alone is sufficient to achieve the state-of-the-art performance.

By sampling the image into random pixels and patches, we have solved the issue of a pre-specified grid size. In addition, fixing the number of pixels or patches to be sampled does not require rigorous cross-validation. Most of the saliency systems are either efficient in detecting salient regions or in predicting eye-gaze fixations. The proposed PR2 model is perhaps one of the saliency systems apart from HA07 which has consistent performance on both of these tasks.

The proposed saliency models work on color contrast and does not recognize saliency in terms of corner or dominant points, orientations or shapes. This issue could be addressed by factoring in other features like orientation, gradient, texture, etc. into the proposed framework. The additional experiments involving image snapshots from tutoring videos has shown that most of the existing saliency systems in their current format are ineffective in predicting eye-gaze which are driven

by top-down influences. This result is intuitive as the existing saliency systems are driven by low-level bottom-up image features, while the eye-gaze in an interaction scenario is driven by high-level top-down semantics. This motivates us to incorporate task relevant information to the proposed saliency models and evaluate if they could predict the eye-gaze in an interaction or demonstration scenario. A more detailed discussion about the future works and enhancement are postponed to the concluding chapter of the thesis. The proposed saliency models were published in [106, 107, 109]. Subsequent to the proposed works, we have witnessed the usage of random sampling based approaches for saliency computation as it can be seen from works like [63, 47].

For practising engineers we recommed to use PR3 for eye-gaze prediction and PR1 for salient region detection applications. In cases where both these tasks are required to be executed concurrently, PR2 could be employed. To obtain a good segmentation mask, we recommend to used TH12 threshold scheme as it provides the best map specific segmentation threshold over all the approaches.

4

SALIENCY MODEL IN THE CONTEXT OF GOAL-DIRECTED ACTION

We have emphasized from the previous chapters that computational modeling of visual attention has profound implications on developing intelligent interactive robots. Models of visual attention are being actively used in robotic eye-gaze control for automatic detection of salient regions and objects in a visual scene. Saliency based visual attention models can also help social robots to nearly mimic human eye-gaze while freely viewing static images. Despite this, the existing models of visual attention have not been successful so far in guiding a robot's attention during an interaction or tutoring scenario. We have inferred this from our experiments on Bielefeld Motionese corpus presented in Section. 3.4.6. The two fundamental reasons behind this shortcoming are:

1. The image based saliency models cannot be directly used to analyze video streams
2. Saliency models are driven mainly by image features, while the human attention is also driven by history, context, anticipation, task based semantics and other high level priors

Developing an attention system which could reason the intentions of humans and infer their next actions, is thus challenging.

Humans have an innate ability to deploy and shift attention to appropriate locations in a spatio-temporal scene while being tutored a goal-directed action. The experiments conducted by Falck-Ytter et al. [26] have revealed that 12 months old infants could judge the destination of an object being manipulated by the caregiver during a goal-directed action, and hence were able to shift their eye-gaze to the destination even before the object arrived there. Due to increased cognitive development by repeated exposure to goal-directed actions, humans develop the ability to predict where the objects occur next while viewing a goal-directed action. This helps in deploying the attention to the particular region even before the object arrives there, and prepare further for a motor reaction if necessary. In this context, we propose a modification to the **PR2** saliency system which can be used to control the robotic eye-gaze while viewing a goal-directed action. The performance of the proposed system is validated on five different goal-directed action videos and benchmarked along with the performance of original **PR2** and **SE09**. The **PR2** based saliency computation relies on fusion of the local saliencies of random image patches, while **SE09** saliency model relies on image self infor-

mation. The **SE09** in particular is tuned to handle spatio-temporal data as compared to rest of the saliency models we considered previously. The goal-directed action is defined within the limited context of transporting a particular object from a source to a destination in an object-specific trajectory. The relevant research on this topic is further described in the next section.

4.1 RELATED WORK

The associated work on this research area is sparse and is available from disparate fields. We describe the relevant works from the computer vision, reinforcement learning and interaction studies discipline.

4.1.1 *Computer Vision*

The principal thrust of attention modeling has been focused on developing saliency maps that predict human eye-gaze on static images and scenes. Some of the prominent examples were explained in Chapter. 2. These are essentially bottom-up models that are driven by low level image features. Furthermore, they can only predict the human eye-gaze shifts for a free-viewing task, i.e viewing images without an objective.

For a visual task like searching or recognizing an object in a scene, the attention is driven by expectancy. Such an attention is called the top-down attention, where there is a prior knowledge of particular features or objects to look for in the scene. Elazary and Itti [25] proposed an object recognition and localization mechanism based on Bayesian learning of object specific feature maps. Moosmann et al. [78] also introduced an object specific saliency map, which is based on the conspicuity of histogram of oriented gradients descriptors. Frintrop et al. [29] introduced the VOCUS framework to search for a target object in complex indoor scenes. These works rely on efficient learning of object specific features and enforce a brute force search on the input image to localize the object.

Humans do not perform a point-to-point template matching in order to search for an object in a scene. They rather optimize the searching behavior by using feature and location specific cues. An eye-tracking study carried out by Torralba et al. [100] has revealed that the eye saccadic behavior of a participant was different while searching different objects. Object specific location priors were further used to re-weight the bottom-up saliency maps to obtain a better prediction of the eye-gaze by Torralba et al. [100] and Chikkerur et al. [21]. Navalpakkam et al. [82] proposed an ontology based search for objects or regions in a scene which relies on higher level semantic information. Contextual information was also used to detect the presence of objects by Torralba et al. [100]. A neural network equivalent of

such models which encodes object and location priors was proposed by Wang et al. [115].

The majority of the existing image based saliency maps cannot be directly extended to handle video streams. However, saliency systems which are capable of handling video streams are highly consequential for robotic vision, as they help in sustaining a coherent human-robot interaction. Modeling visual attention on videos is mostly achieved by decomposing the video into individual image frames, and applying the bottom-up saliency models independently on each frame. To a certain extent they have shown to be effective, despite not utilizing the temporal and contextual information that is available in a video data.

A few methods like [94, 31, 37, 38], were specifically designed to predict visual attention on videos. They principally employ motion history to compute visual saliency and have been successfully applied for anomaly detection in videos. However, these spatio-temporal attention models are not useful in handling real time video streams, as they require the successor image frame to compute the saliency of the current image frame.

In general, tutoring related video streams consist not only of events which freely manipulate objects, but also goal-directed actions which are based on certain semantics. Developing a saliency system which has the capability to predict visual attention specific to a goal-directed action is essential for the development of intelligent robotic systems. Recent works by Yuen et al. [127] and Rodriguez et al. [87] attempt to predict the future trajectory of a moving object in an image. A set of videos from a scene are decomposed into individual frames and are clustered to form a knowledge base. The motion history from the reference frames most similar to the query image is retrieved as the expected motion trajectory. Though these works are tested on real life videos, they are not based on the semantics underlying the task. As a result they process all the pixels in the image, while only a few of them which constitute an object of interest are relevant.

4.1.2 *Reinforcement Learning*

Reinforcement learning techniques have also been used to model anticipatory eye movements during a human-machine interaction scenario. Reinforcement based attention control models employ history and spatial memory and hence are more robust than saliency based attention control. The first reinforcement learning based anticipatory system was proposed by Balkenius and Johansson [9]. A simulation to predict the future trajectory of a particle moving in a sinusoidal trajectory has shown its efficacy. Fix et al. [28] have presented a continuous attractor network model that is able to anticipate the visual scene as it is supposed to be after an action execution. The predic-

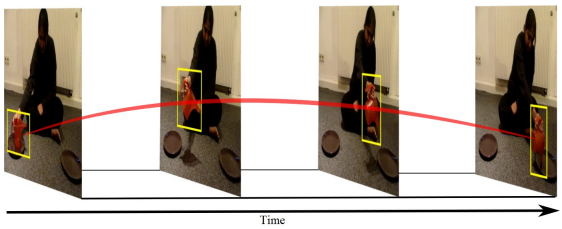


Figure 35: Task driven attention. The illustration involves the demonstration of a plant water jug usage. Experienced observers can predict future positions (marked in red) of the water jug and reorient their attention correspondingly. Redundant background is suppressed and only the water jug (enclosed by yellow box) is focused.

tive information is further used in the context of a serial search of a target. Though the aforementioned models are appealing, they have been tried and tested only on synthetic images and not on real life videos.

4.1.3 *Interaction Studies*

Yi and Ballard [124] proposed a Bayesian network for recognizing and anticipating the steps in a sandwich making task. The study has revealed that the eye-gaze shifts are temporally correlated with the task. The model is highly abstract, conceptually complex, and assumes that the object recognition and localization problems are solved by default.

4.2 MOTIVATION AND CONTRIBUTIONS

The selective tuning theory of visual attention proposed by Tsotsos [102] views visual attention as a cognitive component which concurrently solves the search and recognition of an object in a scene optimally. Attention is viewed not just as a reactive object tracker, but also as an anticipatory system that can make predictions about where the object of interest could appear next in a given task scenario (see Fig. 35).

Some of the prominent bottle-necks for saliency detection in the context of goal-directed actions are:

1. Video saliency models require the successor frame to predict the saliency of the current frame
2. Image saliency models do not integrate history or memory components
3. Saliency models which can search and recognize objects cannot handle rotational and scale changes
4. Image saliency models cannot be scaled to handle video sequences

5. Video saliency models fail to handle goal-directed actions, as they operate without knowing the location, temporal duration, and the spatial scale of the manipulated task relevant object

6. Absence of a video dataset pertaining to goal-directed actions

In order to address these challenges we propose a task based visual attention model. It computes visual saliency only on a task relevant spatio-temporal window rather than entire scene. The task relevant spatio-temporal windows are obtained from a training phase. We convert the input color image stream into a gray scale and further operate on it. This reduces the computational load on the saliency model. This also further facilitates a fair comparison with other existing models of video saliency, as they all operate on gray scale images. The proposed model has both top-down and bottom-up components. The saliency maps from the top-down and bottom-up components are computed only on the task relevant spatio-temporal window and later fused to produce a master saliency map. The top-down component is based on Stentiford's similarity measure [97]. This similarity measure solves the search and recognition of an object concurrently by image template matching. The bottom-up component is based on the **Random_Sub_Window_Saliency** algorithm. This is technically applying the **PR2** algorithm on a gray scale image. We chose this bottom-up saliency model as it can be applied on an image sub-window without modifications, while most of the existing bottom-up saliency models can be applied only on the entire image and not on a pre-specified region of interest. In order to test the proposed model we have created a dataset of videos demonstrating goal-directed actions. An illustration of the proposed framework is given in Fig. 36.

4.3 PROPOSED MODEL

The proposed model computes saliency maps $\mathbf{S}_1, \cdots, \mathbf{S}_t$ for a goal-directed action sequence $\mathbf{I}_1, \cdots, \mathbf{I}_t$. We define a goal-directed action as transporting an object – where $\mathbf{O}$ is the image template of the object – from a source to a destination point in a given trajectory with a velocity specific to each time stamp. Only the spatio-temporal window where the target object is expected, is considered for further processing while the rest of the background is ignored.

The task specific spatial window corresponding to each time stamp of the action is computed from the training set. The training set consists of several video sequences where a specific goal-directed action is being demonstrated. Each training video sequence for a particular goal-directed action consists t frames. The co-ordinates of the region of interest (ROI) that encloses the target object was manually recorded for all frames of the training videos. Subsequently, the best enclosing rectangle (BER) which encompasses all the target enclosing ROIs for

a given $\mathbf{I}_i^{th}$ frame instance across all the training videos specific to a given task is stored as the task relevant spatio-temporal window. The ROI on the $\mathbf{I}_i^{th}$ frame of a given video sequence is bounded by the co-ordinates (u_{1i}, v_{1i}) and (u_{2i}, v_{2i}) while the BER which encompasses all the temporally corresponding ROIs is bounded by the co-ordinates (a_{1i}, b_{1i}) and (a_{2i}, b_{2i}).

During the testing phase, only the sub-image within the BER corresponding to the respective time stamp is processed further. The object specific top-down attention map $\mathbf{T}$ based on Stentiford's similarity measure [97] and the stimulus driven bottom-up attention map $\mathbf{B}$ based on the **Random_Sub_Window_Saliency** algorithm are computed within the BER of frame $\mathbf{I}_i$. Finally, the task based saliency map $\mathbf{S}_i$ is computed by pixel-wise multiplication of $\mathbf{T}$ and $\mathbf{B}$. The algorithm **Task_Based_Saliency** (TBS), describes the proposed model.

Algorithm IV : **Task_Based_Saliency**
Input : (1) $\mathbf{O}$ of size $g \times h$
: (2) $\mathbf{I}_1, \cdots, \mathbf{I}_t$ each of size $r \times c$
: (3) $\mathbf{a}_1$, $\mathbf{b}_1$, $\mathbf{a}_2$, and $\mathbf{b}_2$, each of length t
: (4) $\mathbf{x}_1$, $\mathbf{y}_1$, $\mathbf{x}_2$, and $\mathbf{y}_2$, each of length n_r
Output : $\mathbf{S}_1, \cdots, \mathbf{S}_t$ each of size $r \times c$
Method
Step 1 : Set all elements of $\mathbf{S}_1, \cdots, \mathbf{S}_t$ to 0
Step 2 : Compute frame-wise saliency:
for $i = 1$ to t
$[\mathbf{x}_1, \mathbf{y}_1, \mathbf{x}_2, \mathbf{y}_2] =$
Generate_Random_Sub_Windows$(n_r, a_{1i}, b_{1i}, a_{2i}, b_{2i})$
$\mathbf{B}$ = Random_Sub_Window_Saliency($\mathbf{I}_i$,$n_r, x_{1i}, y_{1i}, x_{2i}, y_{2i}$)
$\mathbf{T}$ = Top_Down_Attention($\mathbf{O}$,$\mathbf{I}_i$,$a_{1i}, b_{1i}, a_{2i}, b_{2i}$)
$\mathbf{S}_i$ = Re_Weight($\mathbf{B}$, $\mathbf{T}$)
end-i

The object of interest is localized in a given target image using Stentiford's similarity measure [97]. The similarity measure relies upon matching f random pairs of pixels (fork) taken from reference object image $\mathbf{O}$ and target image $\mathbf{I}_i$. The match is considered a success when the difference between the corresponding pairs of pixels is less than a given threshold (Λ). The process of generating and matching forks is repeated a large number (Γ) of times and the final top-down saliency map is obtained. The said approach is robust to linear and non-linear transformations as illustrated in [97]. The algorithm

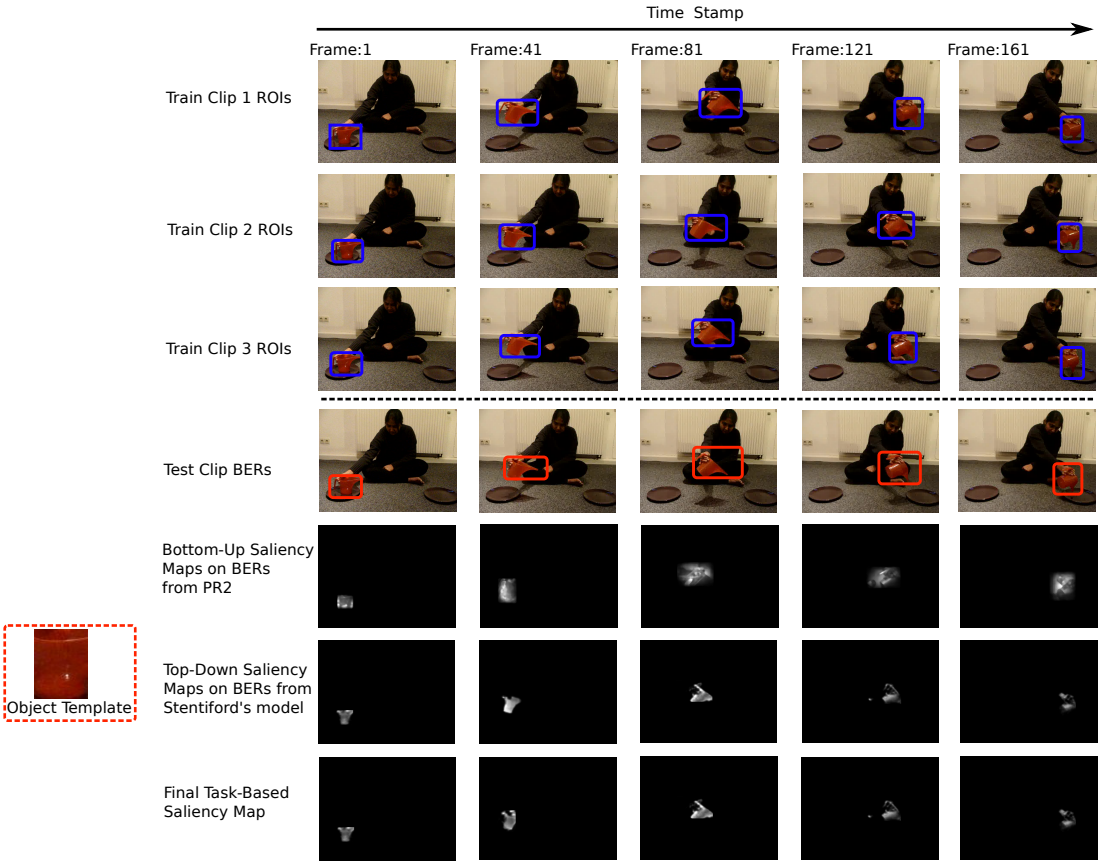


Figure 36: Illustration. We describe the training (with three video samples) and testing phase of the proposed saliency model pertaining to an action demonstration. The train and test videos involve transporting a red jug from a source to destination point in a particular trajectory. The first three rows contains the snapshots at a particular time-stamp from the training videos. Observe that the object of interest (red jug) is enclosed in a blue region of interest (ROI) in the training videos. A best enclosing rectangle (BER), which encompasses the all the ROIs at a particular time-stamp for the all training videos is further computed. It should be noted that while computing the saliency on the test video frames, we process only the region enclosed in the BERs corresponding to a particular time-stamp, as rest of the image is treated as background. Please notice the BERs enclosed within red rectangles on the test video frames in the fourth row of the illustration. The bottom-up saliency is obtained by applying **Random_Sub_Window_Saliency** algorithm only on that part of the test frame which is enclosed by the BERs. The resulting bottom-up saliency maps are shown in the fifth row. Stentiford's [97] algorithm is further employed to localize the object of interest within the BERs. The Stentiford's [97] model requires a template of the target object as a prior. The object template and the resulting top-down saliency maps are given in sixth row of the illustration. Notice that the localization is successful despite the change in pose and orientation of the target object over different time-stamps. The corresponding master saliency maps shown in the final row is obtained by point-wise multiplication of the computed bottom-up and top-down saliency maps.

Top_Down_Attention, describes the Stentiford's model [97] to compute the top-down saliency map.

Algorithm IV.(a) : Top_Down_Attention
Input : (1) **O** of size $g \times h$
: (2) **I** of size $r \times c$
: (3) (p_1, q_1) upper left co-ordinates of the BER
: (4) (p_2, q_2) lower co-ordinates of the BER
Output : **T** of size $r \times c$
Method
Step 1 : Set all elements of **T** to 0.
Step 2 : Generate forks:
for $i = 1$ to f
α_i = Random number in $[-\Delta, +\Delta]$
β_i = Random number in $[-\Delta, +\Delta]$
end-i
Step 3 : Concurrent search, detection and recognition of object:
while $\Gamma > 0$
l = Random number in $[-\Delta, +\Delta]$
m = Random number in $[-\Delta, +\Delta]$
for $k = p_1 + \Delta$ to $p_2 - \Delta$
for $j = q_1 + \Delta$ to $q_2 - \Delta$
$\theta = 1$
for $i = 1$ to f
$\varphi = O(l + \alpha_i, m + \beta_i)$
$\phi = I(k + \alpha_i, j + \beta_i)$
if$(|\varphi - \phi| > \Lambda)$
$\theta = 0$
end-if
end-i
$T(k, j) = T(k, j) + \theta$
end-j
end-k
$\Gamma = \Gamma - 1$
end-while

The algorithm **Re_Weight**, describes the pixel-wise multiplication of top-down and bottom-up saliency maps which results in the master saliency map.

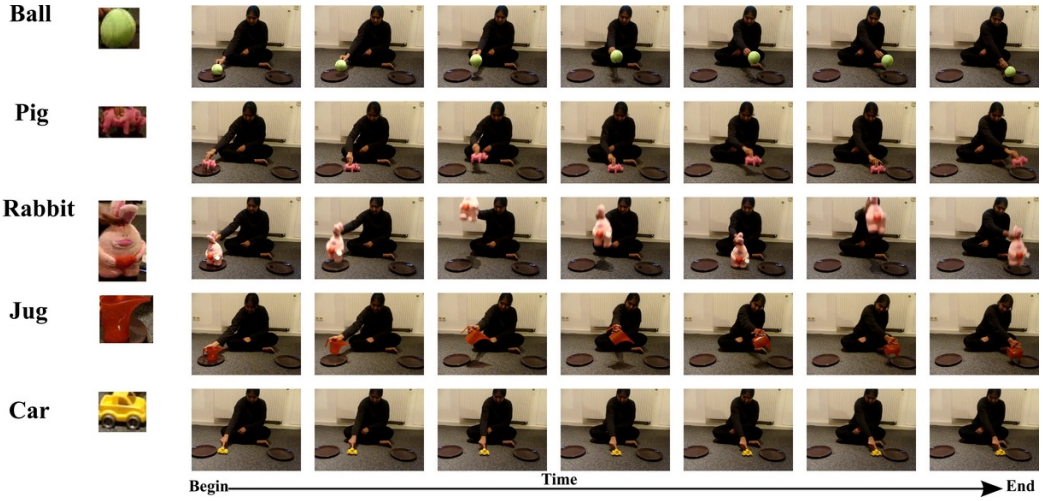


Figure 37: Goal directed action database (available on request). Examples of sequences corresponding to different types of actions are given. The first image consists of the template of the target object (**O**). Please observe the differences in the movement trajectories of the objects in different action sequences. Also note the differences in size, color and aspect ratio of the object templates pertaining to each action sequence.

Algorithm IV.(b) : Re_Weight

Input : (1) **B** of size $r \times c$

: (2) **T** of size $r \times c$

Output : **S** of size $r \times c$

Method

Step 1 : Set all elements of **S** to 0.

Step 2 : Point-wise multiplication:

for $i = 1$ to r

for $j = 1$ to c

$S(i,j) = B(i,j) \cdot T(i,j)$

end-j

end-i

4.4 EXPERIMENTS

For the evaluation, we created a video database containing five types of goal-directed actions. These are demonstrations of ball movement, pig walking, rabbit jump, jug usage and car movement. Each of these actions were demonstrated twelve times by the same demonstrator

in an identical indoor scenario (see Fig. 37). The dataset was created within a controlled environment, and particular care was taken to avoid drastic variations in speed while demonstrating an object motion. Currently the database contains 60 video sequences. All sequences were recorded in front of homogeneous backgrounds with a static camera. The frame rate of the video capture was set to 25fps. The sequences were downsampled to the spatial resolution of 640×480 pixels and have an average length of four seconds. The length of each training video sequence for a particular goal-directed action is set to t frames. This was achieved by manually removing the redundant frames at the beginning and ending of the videos. Further, the co-ordinates of the ROI that encloses the target object was manually recorded for all frames of the video sequences. All sequences from a given action were divided into a training set (4 videos), a validation set (2 videos) and a test set (6 videos). The BERs were computed on the training set while the validation set was used to optimize the parameters $(\Delta, \Lambda, \Gamma, f, g, h, n)$ for each class of action. The presented performance results were obtained on the test set.

We considered the PR2 and SE09 along with the proposed task based saliency model (TBS) for comparative evaluation. An illustration of the resulting saliency maps for an example action sequence is given in Fig. 38. The ability of the considered saliency models to predict the visually interesting areas on the goal-directed videos is evaluated by the CONF metric proposed by LeMeur and Chevet [60]. The metric CONF standing for confidence on a resultant saliency map $\mathbf{S_i}$ is given by:

$$\text{CONF} = \frac{\sum_{p=u_{i1}}^{u_{i2}} \sum_{q=v_{i1}}^{v_{i2}} S_i(p,q)}{\sum_{p=1}^{r} \sum_{q=1}^{c} S_i(p,q)} \tag{11}$$

This is essentially the ratio of the cumulated salience within the annotated ROI and the entire saliency map. CONF tends to 1 when all the predicted salience is inside the ROI. The worst case (0) would suggest that the predicted salience is not in agreement with the BER (the predicted salience would be outside the BER) or there could be a genuine failure in highlighting the target object. It can be observed from Fig. 39 that TBS outperforms the SE09 and the PR2 saliency models by a large margin in terms of CONF. It can be seen that the median performance of TBS on the pig (Fig. 39b) and car (Fig. 39e) sequences is almost 0.9, while the SE09 and PR2 have low performance values. This shows the effectiveness of the TBS model. All plots (except Fig. 39d), show that the SE09 always performs better than the PR2. This converges with our intuition that spatio-temporal saliency based methods perform better than image based saliency methods. To explain the low performance of SE09 on jug (Fig. 39d) sequences, the readers are requested to look at Fig. 38 where the SE09 fails to highlight the jug. The SE09 saliency model is driven by gradients

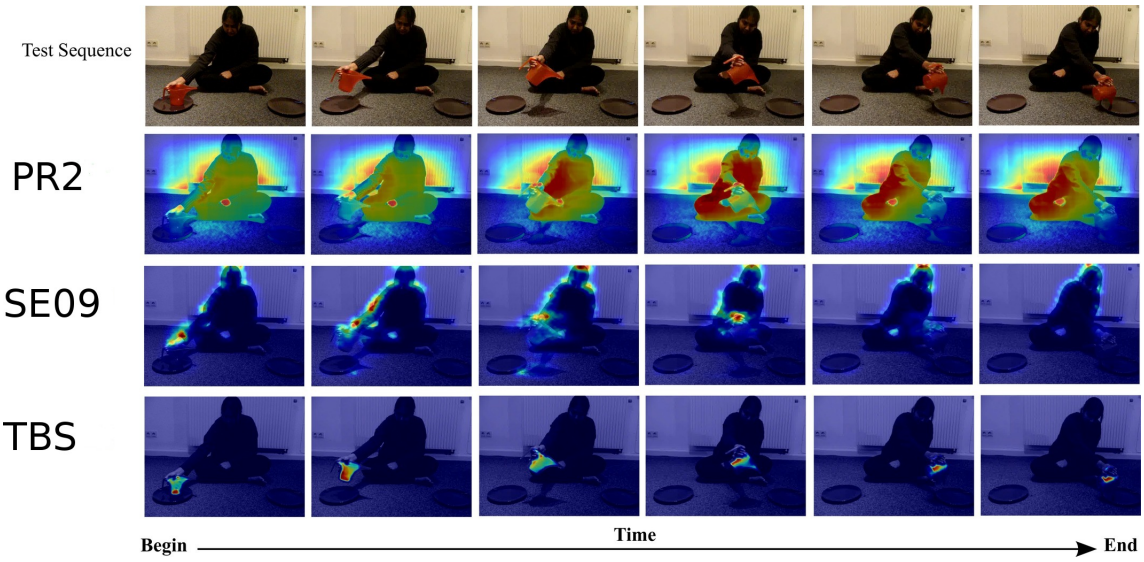


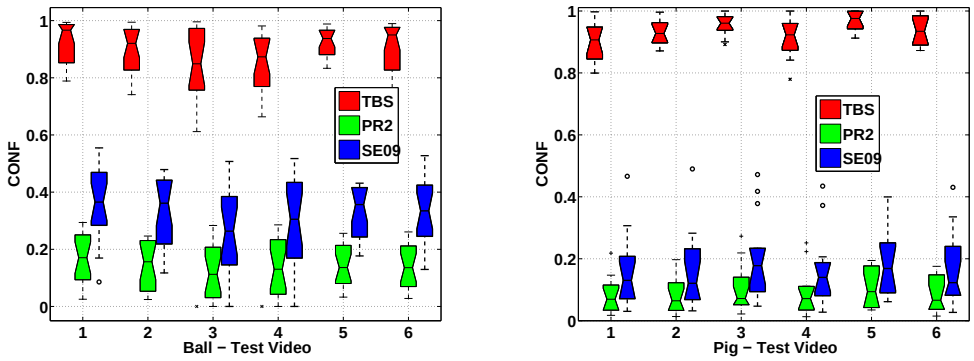
Figure 38: Overlaid saliency maps. Observe that the PR2 highlights contrast rich background, while SE09 highlights only those regions which have motion saliency. Only the TBS effectively highlights the target region as compared to SE09 and PR2.

and hence the contrast free texture of the jug is ignored. The performance of TBS on the rabbit (Fig. 39c) sequence has high variation. This implies that the number of training sequences (4 videos) used to compute the BERs are inadequate as the motion trajectory has high variations. Despite this, TBS fares better than the SE09 and PR2.

4.5 DISCUSSION AND CONCLUSION

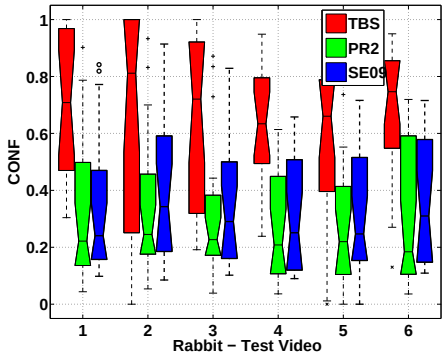
We proposed a task based saliency model that can effectively highlight a target object in a goal-directed action. The proposed saliency model is not just reactive, but employs spatial location priors to localize the object of interest. Please note that the BERs are computed only from the training part of the dataset. The experiments are carried out on the test part of the dataset, using the spatial priors i.e the BERs obtained from the training phase. The actual position of the target object (in the test image) is never used during the test phase, as it would trivialize the entire process. The BERs are also alternatively referred to as symbolic interval representation, and has been successfully utilized for many pattern recognition applications like shape recognition, online signature recognition, clustering etc. [24].

The existing saliency models scan the entire image since they do not have any prior where the object will be moved. We would like to point out that the SE09 saliency model with which we made a comparison requires 15 successor frames in addition to the current frame on which saliency is being computed. On the other hand, the TBS approach does not compute priors from successor or predeces-

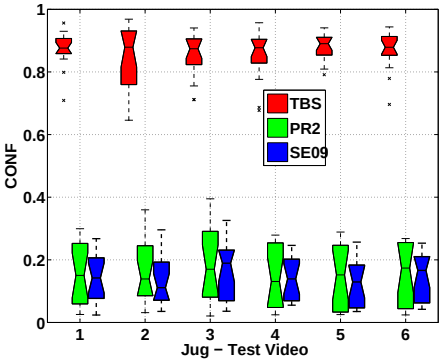


(a)

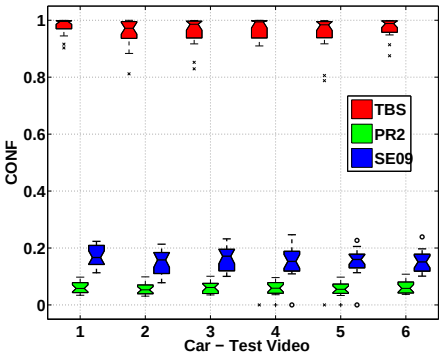
(b)



(c)



(d)



(e)

Figure 39: Performance evaluation in terms of CONF. The results are displayed using a Box-and-Whisker plot. In general, it can be observed that the overall performance of TBS is always better than PR2 and SE09 based saliency approaches for all the test videos.

sor frames. Real-life action demonstrations might also involve pauses, where there is no motion for a specified interval of time. In such no-motion intervals, the existing video saliency models produces a blank saliency map. However, our method does not produce a blank saliency map even during pauses, because the spatio-temporal priors are independent of motion information.

The usage of BERs not only enhances computational efficiency of the proposed architecture, but also reduces the probability of false alarms while detecting the target object. To justify this argument, we provide a visual illustration in Fig. 40 where the **TBS** algorithm is applied to the entire image and not restricted to the region enclosed within the BER.

Our architecture does not emphasize any specific low level image feature, but computes saliency by employing random pixel contrasts which is concordant with human cognitive neurobiology [97]. To the best of our knowledge, this is the first visual saliency model designed to compute saliency for viewing a goal-directed action. Our model can also be integrated with the top-down attention framework of Yi and Ballard [124] for predicting eye-gaze while performing complex tasks. It can also be scaled up to handle a complex action by decomposing it into a combination of atomic actions and handling each atomic action independently.

In the current configuration, the proposed model needs to know the object (whether it is a red jug, yellow car or a pink rabbit etc.) that is being manipulated apriori. However, this issue can be addressed by applying Stentiford's [97] algorithm on a given image frame to localize all the objects under consideration, and choose object that produces the highest response on the top-down saliency map. The appropriate BER which maximally encloses the response of the Stentiford's [97] algorithm can be further used to identify the current time-stamp.

At present, our model works only within a specific camera view. This issue could be resolved by using depth information along with the two-dimensional images to compute the camera calibration matrix. The proposed model can also handle actions with linear changes in speed if the training and test videos are pre-processed using temporal segmentation algorithms such as [12]. In the current work, the training videos are manually aligned. However, temporal alignment of video sequences can be automated by the use of dynamic time warping as shown in [71].

It should be noted that proposed model requires only one training example of the target object. The experimental results have shown that the proposed saliency architecture successfully localizes the target object, despite wide changes in pose and orientation. Our model works reliably even if the target object gets occluded during the course of demonstration. This is because it does not estimate motion in real-

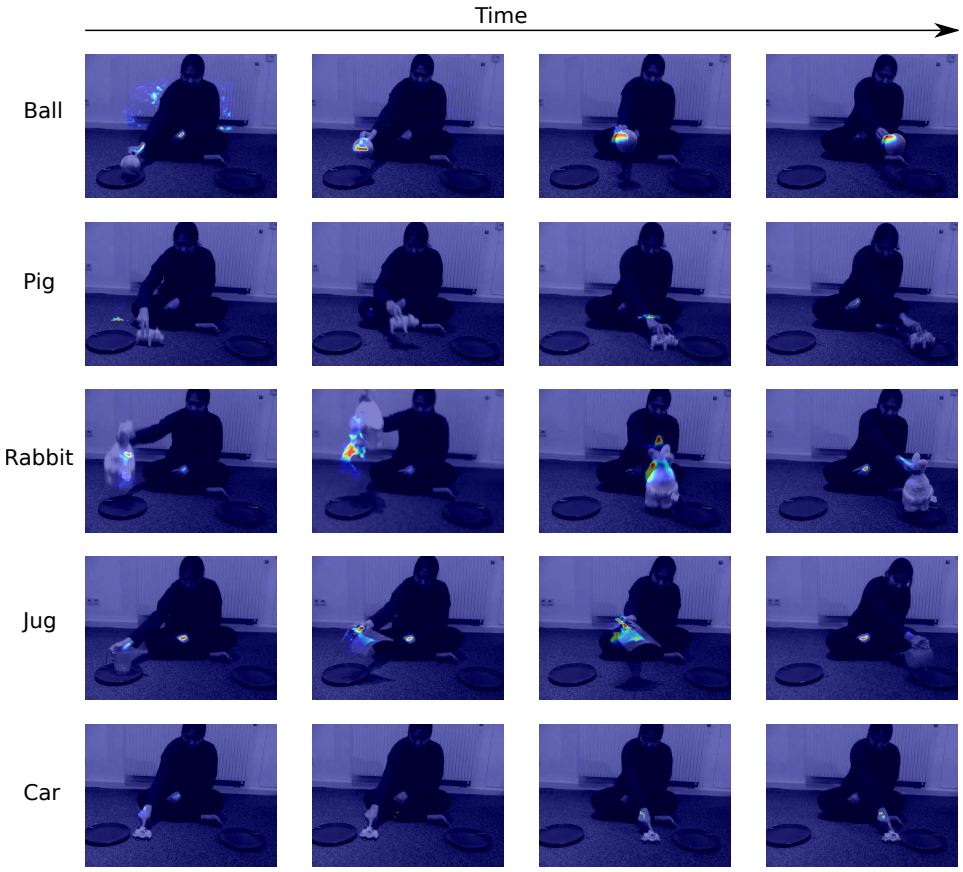


Figure 40: **TBS** algorithm applied on the entire image and not just on BERs. The resulting overlaid saliency maps are displayed. Observe the first image of Ball transportation example. The background is shown salient while the target object is ignored. Please notice the samples from Pig and Car (second and final row) transportation demonstrations. The target object is extremely small, and are faintly highlighted. Background and hand are highlighted instead, as they might resemble the target object templates. It can be further observed in the videos pertaining to Rabbit and Jug (third and fourth rows) that the target object is sometimes missed completely. This illustration corroborates the our assumption that BERs reduce false alarms.

time but instead relies only on the BERs obtained from the training data. The proposed **TBS** model was published in [110, 108].

5

CONCLUSION AND FUTURE WORK

In this thesis, we investigated and proposed the use of randomized algorithms for saliency computation. The proposed saliency models were shown to be effective on eye-gaze prediction task (pure bottom-up task) and also on salient region detection (which has top-down influences). We further improved the state-of-the-art on both of these tasks without any increase in computational complexity or the increase in the size of the feature set.

Several saliency models have been proposed which uses computational paradigms other than the center-surround approach. However, the proposed approaches are driven by the center-surround paradigm and thus reinforces the original proposition that the computational process of the human visual attention system is driven by surround-suppression.

Many successful pattern classification techniques like mixture of Gaussian, mean shift algorithms for blob detection, random forest for classification, adaboost, etc., are all examples of successful randomized algorithms. The success of these algorithms are mainly driven by their ability to obtain quick approximations rather than exact solutions. Parallelly, the selective tuning theory of attention hypothesized that the brain implements approximations through optimization to solve the vision problem. The limitations in the computational power available in the neurons act as the constraints for this optimization problem. Attention is thus seen as a controller which schedules and sequences the usage of various available resources in the brain to suppress the irrelevant stimuli that is present in the visual space. The optimization hypothesis is justified through the presence of optical illusions and other veridicalities which inturn is a result of bad convergence. Thereby, we can see parallels in the random algorithms like mixture of Gaussians and random forests which are driven by approximations to solve a pattern classification problem and the selective tuning theory which explains the necessity for approximations in solving the vision problem. The proposed methods sample random pairs of pixels or patches and obtains a convergence by repeating this process a large number of times. This helps us in solving problems like pre-specifying an initial condition, fixing a grid space, without knowing the location, scale or presence of an object. Furthermore, the vision problem is computationally complex and sometimes is even ill-conditioned and ill-posed. These are the very problems which the proposed saliency computation approaches address by random sampling of the visual space. Another important factor which affects the

computational efficacy is the size of the parameter set. A majority of the methods have a large set of tunable parameters and sometimes may even include meta parameters. However, the proposed bottom-up saliency models have only one tunable parameter and the experiments have revealed that it does not require rigorous cross-validation to fine tune them.

Human attention system is driven not just by bottom-up features but also by context, memory, anticipation, task relevant semantics, etc. The bottom-up saliency approaches model the response of the visual system on an exposure to a stimuli. The bottom-up attention is expected to hand over the controls to other cognitive mechanisms while interacting with the stimuli on a long term. This is corroborated by our experiments on the Bielefeld Motionese Corpus [111] (in Section. 3.4.6) where most of the saliency systems fail to predict the annotated eye-gaze locations. As a result, we tried to model task dependent semantics by employing spatio-temporal priors. We thereby address the issue of learning where and when to attend in the limited context of simple goal directed actions. In order to keep the discourse in near proximity to cognitive sciences, we employed Stentiford's [97] concurrent search and localization algorithm [97]. The algorithm also has inherent random process and is robust to linear transformation. The final objective of attention is to guide the search and recognition mechanisms. This phenomenon is explained in the partial and full recurrence binding aspects in the selective tuning theory. The Stentiford's algorithm [97] thus models these aspects thereby rendering it more attractive than other existing object detection and recognition techniques.

We conducted experiments on MSRA [68], York University [16] and MIT eye-fixation [51] datasets. But during the present times we lack a dataset which has both eye-gaze fixation recordings and salient region detection annotations. The existence of such a unified dataset would enable us to understand the relationship between bottom-up eye-gaze fixations and the sub-sequent eye-gaze re-orientation to a specific image region. We also have to take into account all of the existing datasets are taken by photographers. These images are captured using regular and ideal camera orientations and center bias. But the developed saliency systems are envisaged to work on robotic platforms both in civilian and industrial areas where there can be irregular camera movement, ego-motion, change of illuminations etc. A robust saliency system can thus be developed by testing it on hard datasets like the one envisaged above. In the current work, we focus primarily on contrast and pixel based features. With the Kinect revolution, depth information along with the pixel intensity values are available for a low cost. The future test datasets should perhaps include the depth information so that researchers can analyze the many latent factors which have not been examined so far.

The limitations in the variety of the datasets that are available places a restriction on the evaluation of the saliency models. The reaction of the visual system to a stimuli is captured in the EEG recordings. No such systematic analysis of correlating the EEG data, the fixation density values and the predicted saliency values are in existence. Such a study will the bridge the gap between the computer vision and the cognitive vision disciplines. In addition, a majority of the experiments involve adults as gazing subjects. But in order to understand the developmental process involved, we need to factor in the results when the subjects are children. Note that we analyze this subtly during the experiments on Bielefeld Motionese Corpus (in Section. 3.4.6). The eye-gaze data obtained was from children, and in parallel the saliency systems were not able to predict them properly. It is thus necessary to ascertain the band width of the age for whom which the current saliency systems are able to predict the eye-gaze appropriately.

The proposed saliency systems can also be enhanced in several aspects. In the current format, they do not recognize orientation or shape based saliencies. This can be alleviated by considering orientation, gradient, shapelet maps etc. Currently, the random locations and scales are sampled from a uniform distribution. Other distributions like Weibull's distribution and power law based models have also been found in detecting and modeling saliency. Further investigation is necessary to replace uniform distribution based sampling to a more appropriate distribution. We have deliberately not incorporated machine learning techniques for sampling a location or fixing the size and shape of the patch. Saliency approaches which incorporate machine learning models are computationally complex but are highly effective within certain limitations. In our future work, we will attempt to understand the machine learning models which are more appropriate to enhance the proposed saliency systems.

In this thesis, we utilized information retrieval metrics like F-measure, ROC-AUC, average precision along with information theoretic measure like mutual information and a statistical measure like correlation coefficient. All of these measures are from diverse disciplines and can capture specific attributes like performance, accuracy, reliability, etc. of a classifier. As we have seen through the course of this research, a high performance on a single metric alone is not sufficient in deciding the best saliency model. We deem it necessary to evaluate the saliency systems on a variety of metrics and identify the best saliency model by consensus. In the proposed saliency systems, we fix the number of image patches to be sampled manually. This can be automated by incorporating a stochastic gradient ascent or descent algorithm which automatically senses the convergence of a saliency map.

The proposed patch based saliency models also have similarities to the spotlight theory of attention [101]. Like in the spotlight theory of

attention, the proposed saliency models focus on a particular patch of an image and subsequently move on to another patch. The spotlight theory of attention also hypothesizes that attention focuses on a specific region in the visual space and subsequently moves to another. Active research is still being pursued to understand the size of the attention spotlight, if the region between two spotlights are processed, the sequence in which the spotlights are chosen etc. Our model can be seen as a special case based on the spotlight theory where the location and the size of the spotlight is random. The area between two spotlights is not processed immediately when the attention shifts from one location to another. This unprocessed area is not immediately attended, but saliency values are attributed to these locations due to the processing of overlapping spotlights which are image patches in our case. This might be another direction of research which requires investigation.

Our saliency models sample random patches and pixels sequentially. But the evidence from neurobiology states that the visual system employs a ensemble of serial and parallel information processing systems. We are thus required to test parallel algorithms and visualize the impact on the performance of the proposed saliency systems. The visual system is also driven by experience. This aspect can be incorporated into the saliency systems through Bayesian statistics, as this gives a good prior for choosing patch sizes and locations.

It is our opinion that an artificial attention system should have all the capabilities without having the drawbacks of the human visual attention. In the case of human beings, evolution decides the break even point between computational efficiency and accuracy in solving the vision problem. Any natural improvement will require significant amount of time as it requires the genetic information to mutate. On the contrary, the artificial attention systems can benefit from the advances in mathematics. The selective tuning theory of attention is one of the few systems which incorporates lattice algebra to explain the various sub-functions involved in attention. Saliency is one small aspect in the framework of the selective tuning theory. We are curious to understand if the proposed saliency models fit into the lattice algebra framework of selective tuning. Insofar, we have seen several saliency models which are specialized for a particular task. The general solution for the visual attention problem is thus presented as an aggregation of several specialized solutions. We have attempted to reduce the degree of specialization wherein the proposed saliency models have consistent performance on both eye-gaze fixation and salient region detection task. However, it is still a long way from the ideal visual attention module.

Improvements on saliency models can cascade to several computer vision applications. In the era of small touch screens, where users prefer to watch movies and large images, saliency models come into

play. As we know, saliency models can automatically detect salient regions and thereby crop the redundant background in the videos for a better viewer experience. Large images can automatically be zoomed location wise by computing the saliency map and finding out the top salient blobs. Saliency models will also enhance image thumbnailing, resizing, collage creation and other photo editing functions.

We hope that these discussions inspires new research in some of the discussed directions. Practicing engineers in the industry could perhaps gain hints about the examined saliency systems for their strengths and shortcomings and use them in the appropriate context. With greater integration of computer sciences and cognitive sciences in the context of saliency research, a more inter-disciplinary dialogue is set to evolve. We further hope that our proposed saliency models are further advanced and help in finding an answer to the larger attention problem over the next few years.

BIBLIOGRAPHY

[1] G. Abdollahian, C. M. Taskiran, Z. Pizlo, and E. J. Delp. Camera motion-based analysis of user generated video. *IEEE Transactions on Multimedia*, 12(1):28–41, January 2010.

[2] Radhakrishna Achanta, Francisco Estrada, Patricia Wils, and Sabine Süsstrunk. Salient region detection and segmentation. In *International Conference on Computer Vision Systems*, pages 66–75, 2008.

[3] Radhakrishna Achanta, S. Hemami, F. Estrada, and S. Susstrunk. Frequency-tuned salient region detection. In *IEEE conference on Computer Vision and Pattern Recognition*, pages 1597–1604, 2009.

[4] Radhakrishna Achanta and Sabine Süsstrunk. Saliency detection using maximum symmetric surround. In *International Conference on Image Processing*, pages 2653–2656, 2010.

[5] Bogdan Alexe, Thomas Deselaers, and Vittorio Ferrari. What is an object? In *IEEE conference on Computer Vision and Pattern Recognition*, pages 73–80, 2010.

[6] Coduta O. Ancuti, Cosmin Ancuti, and Philippe Bekaert. An effective grayscale conversion with applications to image enhancement. In *ACM SIGGRAPH ASIA*, 2009.

[7] L. Aryananda. Attending to learn and learning to attend for a social robot. In *IEEE-RAS International Conference on Humanoid Robots*, pages 618–623, 2006.

[8] T. Avraham and M. Lindenbaum. Esaliency (extended saliency): Meaningful attention using stochastic image modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(4):693–708, 2010.

[9] Christian Balkenius and Birger Johansson. Anticipatory models in gaze control: a developmental model. *Cognitive Processing*, 8(3):167–174, 2007.

[10] Sang-Woo Ban, Bumhwi Kim, and Minho Lee. Top-down visual selective attention model combined with bottom-up saliency map for incremental object perception. In *International Joint Conference on Neural Networks*, pages 1–8, 2010.

[11] M. Begum and F. Karray. Visual attention for robotic cognition: A survey. *IEEE Transactions on Autonomous Mental Development*, 3:92–105, 2011.

[12] A. Bobick and J. Davis. Real-time recognition of activity using temporal templates. In *IEEE Workshop on Applications of Computer Vision*, pages 39–42, 1996.

[13] Iva Bogdanova, Alexandre Bur, Heinz Hügli, and Pierre A. Farine. Dynamic visual attention on the sphere. *Computer Vision and Image Understanding*, 114(1):100–110, 2010.

[14] Ali Borji, Majid N. Ahmadabadi, and Babak N. Araabi. Learning sequential visual attention control through dynamic state space discretization. In *International conference on Robotics and Automation*, pages 2294–2299, 2009.

[15] N. D. B. Bruce and J. K. Tsotsos. Saliency based on information maximization. *Advances in Neural Information Processing Systems*, pages 155–162, 2005.

[16] Neil D. B. Bruce. Features that draw visual attention: an information theoretic perspective. *Neurocomputing*, 65-66:125–133, 2005.

[17] Timothy J. Buschman and Earl K. Miller. Shifting the spotlight of attention: evidence for discrete computations in cognition. *Frontiers in Human Neuroscience*, 4, 2010.

[18] Y. Caron, P. Makris, and N. Vincent. Use of power law models in detecting region of interest. *Pattern Recognition*, 40(9):2521–2529, 2007.

[19] Moran Cerf, Jonathan Harel, Wolfgang Einhäuser, and Christof Koch. Predicting human gaze using low-level saliency combined with face detection. In *Advances in Neural Information Processing Systems*, 2007.

[20] Kai-Yueh Chang, Tyng-Luh Liu, Hwann-Tzong Chen, and Shang-Hong Lai. Fusing generic objectness and visual saliency for salient object detection. In *International Conference on Computer Vision*, pages 914–921, 2011.

[21] Sharat Chikkerur, Thomas Serre, Cheston Tan, and Tomaso Poggio. What and where: a bayesian inference theory of attention. *Vision Research*, 50(22):2233–2247, 2010.

[22] Xinyi Cui, Qingshan Liu, Shaoting Zhang, Fei Yang, and Dimitris N. Metaxas. Temporal spectral residual for fast salient motion detection. *Neurocomputing*, 86:24–32, 2012.

[23] R. Descartes. *Les Passions de l'âme*. Le Gras, 1649.

[24] Diday Edwin and Esposito Floriana. An introduction to symbolic data analysis and the sodas software. *Intelligent Data Analysis*, 7(6):583–601, 2003.

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

[26] Terje Falck-Ytter, Gustaf Gredebäck, and Claes von Hofsten. Infants predict other people's action goals. *Nature Neuroscience*, 9(7):878–879, 2006.

[27] Yuming Fang, Weisi Lin, Chiew Tong Lau, and Bu-Sung Lee. A visual attention model combining top-down and bottom-up mechanisms for salient object detection. In *International Conference on Acoustics, Speech and Signal Processing*, pages 1293–1296, 2011.

[28] Jérémy Fix, Nicolas P. Rougier, and Frédéric Alexandre. A Top-down attentional system scanning multiple targets with saccades. In *From Computational Cognitive Neuroscience to Computer Vision : CCNCV 2007*, Bielefeld, Germany, 2007.

[29] Simone Frintrop. *VOCUS: A Visual Attention System for Object Detection and Goal-Directed Search*, volume 3899 of *Lecture Notes in Computer Science*. Springer, 2006.

[30] Simone Frintrop, Ro Erich, and Henrik I Christensen. Computational visual attention systems and their cognitive foundations: A survey. *ACM Transactions on Applied Perception*, 7:6:1–6:39, 2010.

[31] Dashan Gao, Vijay Mahadevan, and Nuno Vasconcelos. The discriminant center-surround hypothesis for bottom-up saliency. In *Advances in Neural Information Processing Systems*, 2007.

[32] C. A. Glasbey. An analysis of histogram-based thresholding algorithms. *Graphical Models and Image Processing*, 55:532–537, 1993.

[33] Stas Goferman, Lihi Zelnik-Manor, and Ayellet Tal. Context-aware saliency detection. In *IEEE conference on Computer Vision and Pattern Recognition*, pages 2376–2383, 2010.

[34] V. Gopalakrishnan, Yiqun Hu, and D. Rajan. Random walks on graphs to model saliency in images. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1698–1705, 2009.

[35] V. Gopalakrishnan, Yiqun Hu, and D. Rajan. Salient region detection by modeling distributions of color and orientation. *IEEE Transactions on Multimedia*, 11(5):892–905, 2009.

[36] V. Gopalakrishnan, Yiqun Hu, and D. Rajan. Random walks on graphs for salient object detection in images. *IEEE Transactions on Image Processing*, 19(12):3232–3242, 2010.

[37] Chenlei Guo, Qi Ma, and Liming Zhang. Spatio-temporal saliency detection using phase spectrum of quaternion fourier transform. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2008.

[38] Chenlei Guo and Liming Zhang. A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression. *IEEE Transactions on Image Processing*, 19(1):185–198, 2010.

[39] Jonathan Harel, Christof Koch, and Pietro Perona. Graph-based visual saliency. In *Advances in Neural Information Processing Systems*, pages 545–552, 2007.

[40] J. F. Herbart. Psychologie als wissenschaft neu gegründet auf erfahrung. *Metaphsyik und Mathematik*, 1824.

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

[42] Xiaodi Hou and Liqing Zhang. Dynamic visual attention: searching for coding length increments. In *Advances in Neural Information Processing Systems*, pages 681–688, 2008.

[43] Yiqun Hu, Deepu Rajan, and Liang-Tien Chia. Adaptive local context suppression of multiple cues for salient visual attention detection. In *International Conference on Multimedia and Expo*, pages 346–349, 2005.

[44] Yiqun Hu, Deepu Rajan, and Liang-Tien Chia. Robust subspace analysis for detecting visual attention regions in images. In *ACM international conference on Multimedia*, pages 716–724, 2005.

[45] Chaobing Huang, Quan Liu, and Shengsheng Yu. Regions of interest extraction from color image based on visual saliency. *The Journal of Supercomputing*, pages 1–14, 2010.

[46] Rui Huang, Nong Sang, Leyuan Liu, and Qiling Tang. Saliency based on multi-scale ratio of dissimilarity. In *International Conference on Pattern Recognition*, pages 13–16, 2010.

[47] Zhiyong Huang, Fazhi He, Xiantao Cai, Zhengqin Zou, Jing Liu, Mingming Liang, and Xiao Chen. Efficient random saliency map detection. *SCIENCE CHINA Information Sciences*, 54(6):1207–1217, 2011.

[48] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11):1254–1259, 1998.

[49] W. James. *Principles of Psychology*. Holt, 1890.

[50] Nathalie Japkowicz and Mohak Shah. *Evaluating Learning Algorithms: A Classification Perspective*. Cambridge University Press, New York, NY, USA, 2011.

[51] Tilke Judd, Krista Ehinger, Frédo Durand, and Antonio Torralba. Learning to predict where humans look. In *IEEE International Conference on Computer Vision*, pages 2106–2113, 2009.

[52] Pattaraporn Khuwuthyakorn, Antonio Robles-Kelly, and Jun Zhou. Object of interest detection by saliency learning. In *European conference on Computer vision*, pages 636–649, 2010.

[53] Wolf Kienzle, Felix A. Wichmann, Bernhard Schölkopf, and Matthias O. Franz. A nonparametric approach to bottom-up visual saliency. In *Advances in Neural Information Processing Systems*, pages 689–696, 2006.

[54] Gunhee Kim, Daniel Huber, and Martial Hebert. Segmentation of salient regions in outdoor scenes using imagery and 3-d data. In *IEEE Workshop on Applications of Computer Vision*, pages 1–8, 2008.

[55] Hyundo Kim, H. Jasso, G. Deak, and J. Triesch. A robotic model of the development of gaze following. In *International Conference on Development and Learning*, pages 238–243, 2008.

[56] D.A. Klein and S. Frintrop. Center-surround divergence of feature statistics for salient object detection. In *IEEE International Conference on Computer Vision*, pages 2214–2219, 2011.

[57] Ulf Knoblich, Maximilian Riesenhuber, David J. Freedman, Earl K. Miller, and Tomaso Poggio. Visual categorization: How the monkey brain does it. In *Biologically Motivated Computer Vision*, pages 273–281, 2002.

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

[59] O. Le Meur, P. Le Callet, D. Barba, and D. Thoreau. A coherent computational approach to model bottom-up visual attention. *IEEE transactions on pattern analysis and machine intelligence*, 28(5):802–817, 2006.

[60] Olivier Le Meur and Jean-Claude Chevet. Relevance of a feed-forward model of visual attention for goal-oriented and free-viewing tasks. *IEEE Transactions on Image Processing*, 19(11):2801–2813, 2010.

[61] KangWoo Lee, H. Buxton, and Jianfeng Feng. Cue-guided search: a computational model of selective attention. *IEEE Transactions on Neural Networks*, 16(4):910–924, 2005.

[62] Wen-Fu Lee, Tai-Hsiang Huang, Su-Ling Yeh, and Homer H. Chen. Learning-based prediction of visual attention for video signals. *IEEE Transactions on Image Processing*, 20(11):3028–3038, 2011.

[63] Ren Lei, Shi Chaojian, and Ran Xin. Small salient target detection using overlapped sub window. In *International Congress on Image and Signal Processing*, pages 1448–1451, 2011.

[64] Jia Li, Yonghong Tian, Tiejun Huang, and Wen Gao. Probabilistic multi-task learning for visual saliency estimation in video. *International Journal of Computer Vision*, 2010.

[65] Jia Li, Yonghong Tian, Tiejun Huang, and Wen Gao. Multi-task rank learning for visual saliency estimation. *IEEE Transactions on Circuits and Systems for Video Technology*, 21(5):623–636, 2011.

[66] Jian Li, Martin Levine, Xiangjing An, and Hangen He. Saliency detection based on frequency and spatial domain analyses. In *British Machine Vision Conference*, pages 86.1–86.11, 2011.

[67] Yuewei Lin, Bin Fang, and Yuanyan Tang. A computational model for saliency maps by using local entropy. In *AAAI Conference on Artificial Intelligence*, 2010.

[68] Tie Liu, Jian Sun, Nan-Ning Zheng, Xiaoou Tang, and Heung-Yeung Shum. Learning to detect a salient object. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2007.

[69] Tie Liu, Nanning Zheng, Wei Ding, and Zejian Yuan. Video attention: Learning to detect a salient object sequence. In *International Conference on Pattern Recognition*, pages 1–4, 2008.

[70] Zhi Liu, Yinzhu Xue, Liquan Shen, and Zhaoyang Zhang. Non-parametric saliency detection using kernel density estimation. In *International Conference on Image Processing*, pages 253–256, 2010.

[71] C. Lu and M. Mandal. A robust technique for motion-based video sequences temporal alignment. *IEEE Transactions on Multimedia*, 15(1):70–82, 2013.

[72] Yu-Fei Ma and Hong-Jiang Zhang. Contrast-based image attention analysis by using fuzzy growing. In *ACM international conference on Multimedia*, pages 374–381, 2003.

[73] V. Mahadevan and N. Vasconcelos. Automatic initialization and tracking using attentional mechanisms. In *IEEE Computer Vision and Pattern Recognition Workshops*, pages 15–20, 2011.

[74] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schtze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[75] Sophie Marat, Tien Ho Phuoc, Lionel Granjon, Nathalie Guyader, Denis Pellerin, and Anne Guérin-Dugué. Modelling spatio-temporal saliency to predict gaze direction for short videos. *International Journal of Computer Vision*, 82(3):231–243, 2009.

[76] David Marr. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. Henry Holt & Company, 1982.

[77] H. Maruta, M. Ishii, and M. Sato. Salient region extraction based on local extrema of natural images. In *International Conference on Image Processing*, pages 1113–1116, 2010.

[78] Frank Moosmann, Diane Larlus, and Frédéric Jurie. Learning saliency maps for object categorization. In *ECCV International Workshop on The Representation and Use of Prior Knowledge in Vision*, 2006.

[79] Jan Morén, Ales Ude, Ansgar Koene, and Gordon Cheng. Biologically based top-down attention modulation for humanoid interactions. *International Journal of Humanoid Robotics*, pages 3–24, 2008.

[80] N. Murray, M. Vanrell, X. Otazu, and C.A. Parraga. Saliency estimation using a non-parametric low-level vision model. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 433–440, 2011.

[81] Y. Nagai. From bottom-up visual attention to robot action learning. In *International Conference on Development and Learning*, pages 1–6, 2009.

[82] Vidhya Navalpakkam and Laurent Itti. Modeling the influence of task on attention. *Vision Research*, 45(2):205–231, 2005.

[83] R. M. Nosofsky. Stimulus bias, asymmetric similarity, and classification. *Cognitive Psychology*, 23(1):94–140, 1991.

[84] Francesco Orabona, Giorgio Metta, and Giulio Sandini. A proto-object based visual attention model. pages 198–215. 2008.

[85] Chaoke Pei, Li Gao, Donghui Wang, and Ying Hong. A pft visual attention detection model using bayesian framework. In *International Conference on Image and Graphics*, pages 816–820, 2011.

[86] Zhixiang Ren, Yiqun Hu, Liang-Tien Chia, and Deepu Rajan. Improved saliency detection based on superpixel clustering and saliency propagation. In *Proceedings of the international conference on Multimedia*, pages 1099–1102, 2010.

[87] Mikel Rodriguez, Josef Sivic, Ivan Laptev, and Jean-Yves Audibert. Data-driven crowd analysis in videos. In *International Conference on Computer Vision*, pages 1235–1242, 2011.

[88] Paul L. Rosin. A simple method for detecting salient regions. *Pattern Recognition*, 42(11):2363–2371, 2009.

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

[90] Han S and Vasconcelos N. Biologically plausible saliency mechanisms improve feedforward object recognition. *Vision Research*, 50(22):2295–22307, 2010.

[91] Peter A Sandon. Simulating visual attention. *Journal of Cognitive Neuroscience*, 2(3):213–231, 1990.

[92] Nong Sang, Longsheng Wei, and Yuehuan Wang. A biologically-inspired top-down learning model based on visual attention. In *International Conference on Pattern Recognition*, pages 3736–3739, 2010.

[93] Hae Jong Seo and P. Milanfar. Nonparametric bottom-up saliency detection by self-resemblance. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 45–52, 2009.

[94] Hae Jong J. Seo and Peyman Milanfar. Static and space-time visual saliency detection by self-resemblance. *Journal of vision*, 9(12), 2009.

[95] F. Shahbaz Khan, J. van de Weijer, and M. Vanrell. Top-down color attention for object recognition. In *International Conference on Computer Vision*, pages 979–986, 2009.

[96] Hang Shi and Yu Yang. A computational model of visual attention based on saliency maps. *Applied Mathematics and Computation*, 188(2):1671–1677, 2007.

[97] Fred Stentiford. Attention-based similarity. *Pattern Recognition*, 40(3):771–783, 2007.

[98] Xiaoshuai Sun, Hongxun Yao, Rongrong Ji, Pengfei Xu, Xianming Liu, and Shaohui Liu. Saliency detection based on short-term sparse representation. pages 1101–1104, 2010.

[99] Xiaoshuai Sun, Hongxun Yao, Rongrong Ji, Pengfei Xu, Xianming Liu, and Shaohui Liu. Visual saliency as sequential eye fixation probability. In *International Conference on Image Processing*, pages 1093–1096, 2010.

[100] A. Torralba, A. Oliva, M. Castelhano, and J.M. Henderson. Contextual guidance of attention in natural scenes: The role of global features on object search. *Psychological Review*, 113:766–786, 2006.

[101] Anne M. Treisman and Garry Gelade. A feature-integration theory of attention. *Cognitive Psychology*, 12(1):97–136, 1980.

[102] J. K Tsotsos. A computational perspective of visual attention. 2011.

[103] Roberto Valenti, Nicu Sebe, and Theo Gevers. Image saliency by isocentric curvedness and color. In *International Conference on Computer Vision*, pages 2185–2192, 2009.

[104] Eduard Vazquez, Theo Gevers, Marcel Lucassen, Joost van de Weijer, and Ramon Baldrich. Saliency of color image derivatives: a comparison between computational models and human perception. *Journal of the Optical Society of America*, 27(3):613–621, 2010.

[105] Milan Verma and Peter W McOwan. A semi-automated approach to balancing of bottom-up salience for predicting change detection performance. *Journal of Vision*, 10(6):3, 2010.

[106] Tadmeri Narayan Vikram, M. Tscherepanow, and B. Wrede. A random center surround bottom up visual attention model useful for salient region detection. In *IEEE Workshop on Applications of Computer Vision*, pages 166–173, 2011.

[107] Tadmeri Narayan Vikram, Marko Tscherepanow, and Britta Wrede. A visual saliency map based on random sub-window means. In *Iberian Conference on Pattern Recognition and Image Analysis*, pages 33–40, 2011.

[108] Tadmeri Narayan Vikram, Marko Tscherepanow, and Britta Wrede. Integrating habituation into saliency maps. In *IEEE International Conference on Development and Learning and Epigenetic Robotics*, pages 1–6, 2012.

[109] Tadmeri Narayan Vikram, Marko Tscherepanow, and Britta Wrede. A saliency map based on sampling an image into random rectangular regions of interest. *Pattern Recognition*, 45(9):3114–3124, 2012.

[110] Tadmeri Narayan Vikram, Marko Tscherepanow, and Britta Wrede. A saliency model for goal directed actions. In *IEEE International Conference on Development and Learning and Epigenetic Robotics*, pages 1–6, 2012.

[111] Anna-Lisa Vollmer, Karola Pitsch, Katrin Solveig Lohan, Jannik Fritsch, Katharina Rohlfing, and Britta Wrede. Developing feedback: How children of different age contribute to a tutoring interaction with adults. In *International Conference on Development and Learning*, pages 76–81, 2010.

[112] H. von Helmholtz. Sustained and transient components of focal visual attention. *Vision Research*, pages 1631–1647, 1896.

[113] Min Wang, Jia Li, Tiejun Huang, Yonghong Tian, Lingyu Duan, and Guochen Jia. Saliency detection based on 2d log-gabor wavelets and center bias. In *International Conference on Multimedia*, pages 979–982, 2010.

[114] Wei Wang, Yizhou Wang, Qingming Huang, and Wen Gao. Measuring visual saliency by site entropy rate. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2368–2375, 2010.

[115] Yuekai Wang, Xiaofeng Wu, and Juyang Weng. Synapse maintenance in the where-what networks. In *International Joint Conference on Neural Networks*, pages 2822–2829, 2011.

[116] Zheshen Wang and Baoxin Li. A two-stage approach to saliency detection in images. In *International Conference on Acoustics, Speech and Signal Processing*, pages 965–968, 2008.

[117] Longsheng Wei, Nong Sang, and Yuehuan Wang. A biologically inspired object-based visual attention model. *Artificial Intelligence Review*, 34:109–119, 2010.

[118] Matthew H. Wilder, Michael C. Mozer, and Christopher D. Wickens. An integrative, experience-based theory of attentional control. *Journal of Vision*, 11(2), 2011.

[119] Jinhua Xu, Zhiyong Yang, and Joe Z. Tsien. Emergence of visual saliency from natural scenes via context-mediated probability distributions coding. *PloS one*, 5(12):e15796+, 2010.

[120] Tingting Xu, K. Kuhnlenz, and M. Buss. Information-based gaze control adaptation to scene context for mobile robots. In *International Conference on Pattern Recognition*, pages 1–4, 2008.

[121] Junchi Yan, Jian Liu, Yin Li, Zhibin Niu, and Yuncai Liu. Visual saliency detection via rank-sparsity decomposition. In *International Conference on Image Processing*, pages 1089–1092, 2010.

[122] Victoria Yanulevskaya, Jan Bernard B. Marsman, Frans Cornelissen, and Jan-Mark M. Geusebroek. An image statistics-based model for fixation prediction. *Cognitive computation*, 3(1):94–104, 2011.

[123] A. L. Yarbus. *Eye Movements and Vision*. Plenum. New York., 1967.

[124] Weilie Yi and Dana H. Ballard. Recognizing behavior in hand-eye coordination patterns. *International Journal of Humanoid Robotics*, 6(3):337–359, 2009.

[125] Haonan Yu, Jia Li, Yonghong Tian, and Tiejun Huang. Automatic interesting object extraction from images using complementary saliency maps. In *International Conference on Multimedia*, pages 891–894, 2010.

[126] Yuanlong Yu, G.K.I. Mann, and R.G. Gosine. A goal-directed visual perception system using object-based top-down attention. *IEEE Transactions on Autonomous Mental Development*, 4(1):87–103, 2012.

[127] Jenny Yuen and Antonio Torralba. A data-driven approach for event prediction. In *European Conference on Computer Vision*, pages 707–720, 2010.

[128] A. Zajonc. *Catching the light: The entwined history of light and mind*. Bantam, 1993.

[129] Lingyun Zhang, Matthew H. Tong, Tim K. Marks, Honghao Shan, and Garrison W. Cottrell. Sun: A bayesian framework for saliency using natural statistics. *Journal of Vision*, 8(7), 2008.

[130] Wei Zhang, Q.M.J. Wu, Guanghui Wang, and Haibing Yin. An adaptive computational model for salient object detection. *IEEE Transactions on Multimedia*, 12(4):300–316, 2010.