PREDICTIVE VISUAL SALIENCY MODEL FOR SURVEILLANCE VIDEO

Fahad Fazal Elahi Guraya, Faouzi Alaya Cheikh

Faculty of Computer Science and Media Technology, Gjvik University College
P.O. Box 191, N-2802 Gjvik, Norway
phone: + (47) 61135296, email: fahadg@hig.no
web: www.hig.no

ABSTRACT

Visual saliency models(VSM) mimic the human visual system to distinguish the salient regions from the non-salient ones in an image or video. Most of the visual saliency model in the literature are static hence they can only be used for images. Motion is important information in case of videos that is not present in still images and thus not used in most of VSMs. There are very few saliency models which take into account both static and motion information. And there is no saliency model in the literature which uses static features, motion, prediction and face feature. In this paper we propose a predictive visual saliency model for video that uses static features, motion feature and face detection to predict the evolution in time of the human attention or the saliency. We introduce a new approach to compute saliency map for videos using salient motion information and prediction. The proposed model is tested and validated for surveillance videos.

1. INTRODUCTION

Human visual system is attracted by salient objects or events. This is done unconciously and effortlessly in the visual system. Its a challenging task to model such a complex phenomenon of human visual system. Such computational model can be used in many image and video processing applications such as compression, event detection, perceptual quality evaluation, etc.

There are many factors involved in computing these salient regions in a visual scene. These factors or visual cues are categorized mainly into two groups, bottom up, and top down visual cues [1, 2]. In bottom up approaches our visual system computes the salient regions from low level features such as colour, intensity, orientation etc. A famous computational model of bottum-up attention proposed by Itti and Koch [3], that uses low level features such as colour, intensity and orientation. Top down approaches involve more complex visual activity such as object detection, face detection etc. It is done very fastly and efficiently in human visual system. Combining bottom up and top down approaches guide the visual system towards the salient regions or region of interest [5, 6, 7]. It is observed that the human visual system divert the attention to faces 16.6 times more than other similar regions [4]. Therefore, face detection can significantly improve the short commings of static saliency models such as Itti's saliency model [3], GBVS [8] and GAFE [9]. In [10], top-down visual cue of face detection is combined with itti and Koch bottom up saliency computational model [3] which gives promissing results. The bottom up and top down approaches can help us make a model which can detect salient regions in an image, but what about detecting saliency in videos? Videos have an extra dimension, which creates, a perceptual feeling of motion in human brain. Motion has great influence in identifying the salient regions in a complex dynamic visual scene. Many models have been introduced in the literature to detect salient motion such as [11, 12, 13, 14, 25, 26].

Salient motion models combined with prediction, bottom-up and top-down cues can lead to an efficient visual saliency model. A predictive saliency map can be computed on the videos only. It can be generated with the help of static saliency maps of previous(history) frames, and motion vectors between all the previous frames. It helps to maintain the history information of saliency maps that increases the chances of detection of salient features in the current frame. A predictive saliency map is computed is presented in [26].

The saliency model which can be used for visual surveillance has to be modelled considering the information available in surveillance videos. Surveillance videos are most of the time captured at very low resolution. There is always the possibility of people in the field of view of the camera. Human visual system is significantly coupled with eye movements [22] and it easily detect the human faces, a high level visual cue in top down saliency model. To acquire high efficiency of visual attention models for surveillance videos, it is required to combine motion, face and low level features in a single model. To evaluate such saliency model, one approach is the use of eyetracking devices to capture the subjective foveated vision, that gives the positions of a subject eye on a 2-D plane for a given image or video frame [15, 16]. Eye tracking results may give different observation points depending on the observer. These observation points are used to create Gaze Maps, that are finally compared to the saliency maps computed by the saliency model.

This paper is organized as follows. In the next section we will describe the spatio-temporal visual saliency computational model for videos using motion and prediction. The second section will describe the experimental results from eyetracking experiment. Third section will discuss the results and the last section will conclude the paper and point to some future directions.

2. SPATIO-TEMPORAL VISUAL SALIENCY COMPUTATION MODEL

The spatio-temporal saliency model proposed in this paper is based on stationary saliency model that include top-down, bottom-up visual cues and motion information. The bottom-up visual features such as color, intensity and orientation are used. A famous method of computing bottom-up visual cues is presented in [3]. A part from bottom-up visual cues, top-down visual cue such as face is incorporated in our proposed saliency model. As face has significant importance in surveillance videos and grab/attract human visual attention [23]. A model that incorporate bottom-up and top-down visual cues for images has been presented in [10]. In addition to bottom-up and top-down visual cues, our proposed model has also added a new method of computing the salient motion. A method to compute the salient motion in the videos has been proposed in [14]. Apart from top-down, bottom-up visual cues, and motion we have also added the prediction algorithm, that can predict the saliency map of the current frame based on the previous saliency maps from

previous frames, the resultant saliency map is called predictive saliency map (PSM). If we combine all these saliency model with an averaging function, the output will be called predictive visual saliency model(PVSM) as shown in figure 1.

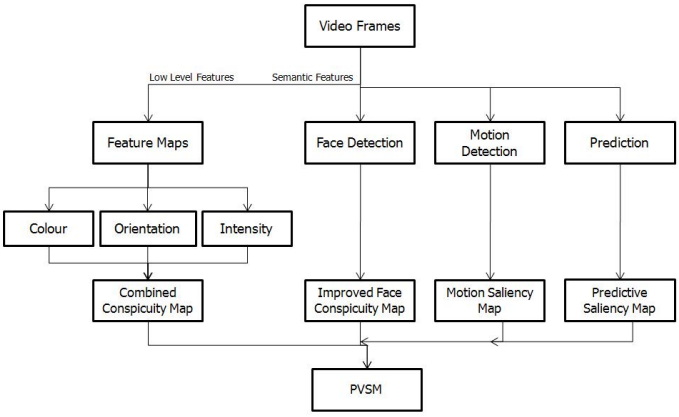

Figure 1: Predictive visual saliency model

2.1 Stationary saliency map

Many saliency models have been proposed in the literature, the famous ones include Itti and koch [3], GAFE [9], GBVS [8]. The most commonly used saliency model with stationary features is the one proposed by [3]. This model generates the saliency map based on the combination of color, orientation and intensity conspicuity maps(color C_c, intensity C_i, and orientation C_o). Itti's saliency model compute the saliency map by taking the average of these three conspicuity maps as shown in equation 1.

$$SM_{itti} = \frac{1}{3}(C_i + C_c + C_o) \tag{1}$$

Experiments show that high level features such as faces attract more attention than other features [17]. As Itti's saliency model is based on low level features and does not consider high level conspicuity maps, it does not perform well for complex scenes such as images with faces, cars, and other familiar objects. To overcome this problem, we need a model which incorporates high level cues such as human faces. There is not so much research done on high level features use in saliency models. A saliency model proposed in [10] use color, intensity, orientation and face detection. This model gives 33% improvement in the saliency detection for images with faces. The authors in [10] have used the face detection model by Walther et al [18]. The face conspicuity map C_{Face} is given four times higher weight than the other conspicuity maps. Equation 2 describes the final saliency model.

$$SM_{Sharma} = \frac{1}{7}(C_i + C_c + C_o + 4C_{Face}) \tag{2}$$

This model provides overall 33% performance improvement over other stationary models [10]. In this paper we propose to used face features, low level features and motion together to create a spatial temporal saliency model as shown in figure 1.

2.2 Motion saliency map

Salient motion is that motion which can grab/attract the attention of the viewer. Motion saliency is a complex phenomenun that depends highly on the specific scene/environment/scenerio. It is also heavily dependent on the viewers interests. In this case we cannot use motion detection methods such as Lukas and Kanade method of detecting motion vectors, to detect salient motion. Because if we use temporal difference of adjacent frames or compute the motion vectors from one frame to another, we might be able to find the moving regions of the image but it cannot distinguish between regions with salient motion and regions with non-salient motion. Let's suppose motion vectors are computed from a video of a tree with moving leaves and some person passing by, it will give us many motion vectors of the tree's region and the passing person's region, here the motion of tree leaves will not be considered salient motion, however the motion of a passing by man will be considered salient. In salient motion detection the non-salient motion has to be filtered out or ignored. In literature we can find much research work done on salient motion detection [14, 19, 24, 25]. A Motion saliency map using spatio-temporal energy accumulation of coherent moving objects by Gabor filtering is proposed in [25]. A salient motion detection algorithm proposed by [19] is using the motion vectors magnitude and phase histograms. These histograms are combined with the help of a proposed formula in such a way that the motion entropy is used to detect the regions with salient motion.

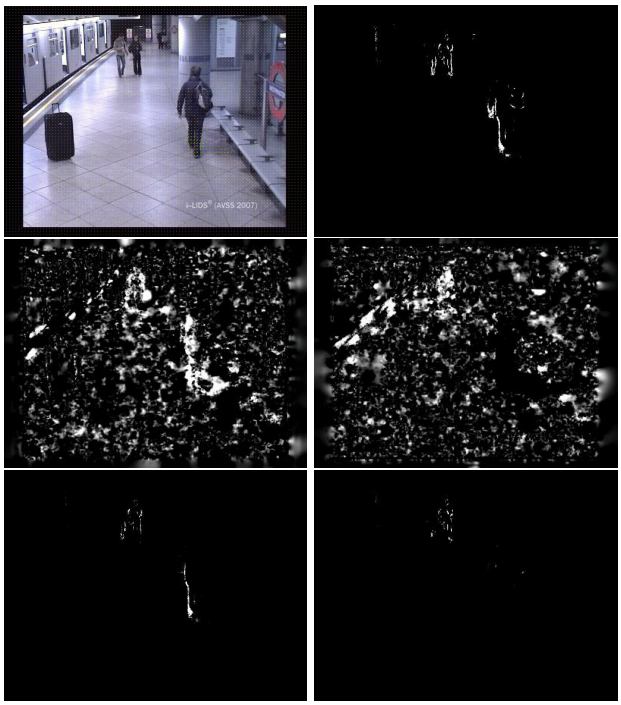

Figure 2: (a) Frame # 155 (top-left), (b) Temporal difference after thresholding (top-right), (c) x-axis motion vector(MV) (middle-left), (d) y-axis MV (middle-right), (e) x-axis MV after filtering (bottom-left), (f) y-axis MV after filtering (bottom-right).

The authors in [14] proposed a method based on temporal differencing, filtering and segmentation. In this paper we propose to use gaussian filter instead of segmentation as the last step of salient motion detection model. The proposed motion model has these main steps: Temporal difference between adjacent frames, motion extraction, temporal filtering, region growing, and multi sources fusion. The region growing and multi sources fusion are not done in our proposed model, for two reasons. First segmentation is a task with high computational complexity, second threshold of region growing is not same for different videos. Rather in the forth

step we included Gaussian filtering of the salient motion pixels to get our motion saliency map. The reasons for using Gaussian filtering instead of segmentation is that segmentation is very slow process and its highly sensitive to region growing threshold values. For example after computing third step of temporal filtering, we get the pixels with salient motion, now if we will do segmentation, there are more chances to get regions which are not salient into this segmentation region. For example if a person wearing black shirt is moving close to its shadow, the segmentation will include the shadow region with the moving person. To avoid these kind of errors we have performed gaussian filtering that gives us a region which is most probable to be salient.

According to [3, 27], a study is conducted to quantify the center bias of the observers in free viewing condition. Hence result show the human vision is center-surround. Thus if there are more than one moving objects in a scene, high priority should be given to the center of each moving object by computing the gaussian filtering on the salient motion points. The further away we go from center, the less salient region we get. The motion saliency maps are filtered by a spatial Gaussian filter of $\sigma = 37$ which was chosen to approximate the size of the viewing field corresponding to the fovea in the gaze map [22].

The first step of computing the motion saliency is computing the temporal difference. It is computed between two subsequent frames F(x,y,t) and F(x,y,t+1). First we take the difference of these two frames and then it is thresholded using a threshold value T_d. T_d can be computed based on the image statistics, and its value is equal to 15 [14]. To detect slow moving objects [14] proposed to use the a weighted accumulation with a fixed weight is used for the new observation as described in equations 4 and 5. The weight w_d is 0.5. The I_{diff} for frame 155 is shown in figure 2(b).

$$I_{fr-diff} = I(x, y, t+1) - I(x, y, t) \quad (3)$$

$$I_{diff}(x, y, t+1) = (1 - w_d).I(x, y, t) + w_d.(I_{fr-diff}) \quad (4)$$

$$I_{temp-diff}(x, y, t+1) = \left\{ \begin{array}{ll} 1 & \text{if } I_{diff}(x, y, t+1) > T_d \\ 0 & \text{otherwise.} \end{array} \right\} \quad (5)$$

The next step consists of finding the motion vectors. In this step we used Lucas-Kannade [21] algorithm for computing the optical flow. We computed motion vectors of those pixels that are detected in the first step in $I_{temporal-diff}$. Lucas-Kanade method works for a given set of points in a video frame to find those same points in the next frame, or for given point F(x,y,t) in frame F_t find the point $F(x + x_\delta, y + y_\delta, t + 1)$ in frame F_{t+1} that minimizes error ϵ as shown in equation 6. The magnitude of motion vectors in x and y directions is shown in figure 2(c) and figure 2(d).

$$\epsilon = \sum_x \sum_y \|F(x + x_\delta, y + y_\delta, t + 1) - F(x, y, t)\| \quad (6)$$

After extracting motion vectors, we first multiply the $I_{temporal-diff}$ with Motion vectors x and y component to filter out the unnecessary motion vectors. The result is shown in the figure 2(e) and figure 2(f). After that we do temporal filtering using the filtered motion vector's x and y components F_x and F_y. The x and y components of a motion vector gives us the displacement of a pixel from previous frame t to new frame t+1, in form of x and y displacements.

It is assumed that the periodic motion is non-salient in surveillance videos. Lets take an example of people jogging on a beach along a lake, the lake has slowly moving water and the beach has trees with moving leaves. In this kind of scenerio and under surveillance context, the periodic motion of the water and tree leaves is not salient, however the motion of people jogging is salient. It is assumed in most surveillance scenerios that the object with salient motion will move in a consistent way in the same direction for a considerable period of time [t,t+n], where n is number of frames. This may not be true for videos other than surveillance videos. A positive count P and negative count N is computed by computing the number of times a pixel moved in positive x or positive y direction, similarly negative x or negative y direction over the period [t,t+n]. This gives us the pixels with salient motion information. In the last step the salient motion pixels are filtered with the gaussian filter to simulate the human visual system [27]. The proposed motion saliency map is shown in 3(d).

2.3 Predictive saliency map

In the predictive saliency model, we propose to use prediction which helps to compute saliency map using the previous saliency maps computed for the previous frames. The detailed description of the predictive saliency model is presented in [26]. This helps to increase the probability to detect salient features in the new frame. Prediction model is implemented on the basis of motion vectors. Using the motion vectors between frame t and t+1, we predict which regions could be salient in frame t+1 based on the history of saliency between frames 1 and frame t. We compute a predictive saliency map (PSM), by moving the salient pixels of frame t to the new location in frame t+1, this location is given by motion vectors between frame t and t+1. The proposed saliency map computational diagram is presented in figure 1.

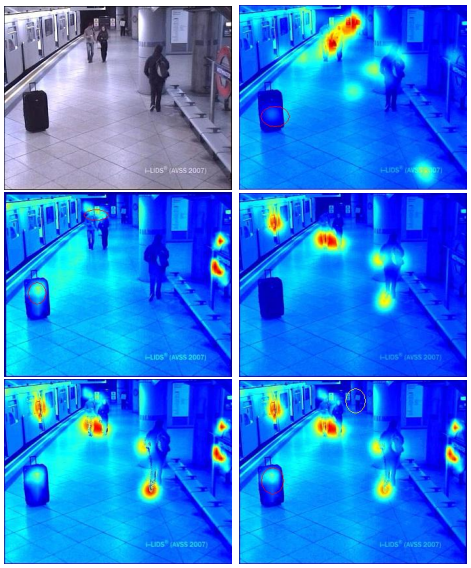

Figure 3: (a) Frame # 175 (top-left), (b) Gaze Map (GM) (top-right), (c) Itti Saliency Map with face detection (SSM) (middle-left), (d) Motion Saliency Map (MSM) (middle-right), (e) Predictive saliency map (PSM) (bottom-left), (f) Mean of PSM and MSM (bottom-right).

In this paper we have used history based on one frame only, however more frames can be used which will make the process more robust at the cost of being complex and slow. Lastly we combine the stationary saliency map with faces, motion saliency map and predictive saliency map into a predictive video saliency map(PVSM) using the average func-

tion. The PVSM is normalized in range of 0 to 1, where 0 represents no saliency and 1 represents most salient pixel.

The saliency model needs to be verified with psychophysical tests. The best way is to get data from subjective eye-tracking experiments after presenting the videos to many subjects. In our experiment we presented three surveillance videos to 30 subjects, and recorded their eye movements using SMI high speed eye tracker with a frequency of 500 samples per second. The video is displayed on a CRT monitor. The eye tracks are acquired and used to generate gaze maps. The gaze maps are obtained by averaging the eyemovements of 30 subjects. In the last step of gaze map production gaze maps are filtered with gaussian filter to depict the human visual system [22]. The gaze maps of frame 389 and frame 175 of video 1 are shown in figure 3. The gaze maps data is later used for the comparison with the saliency maps using Area under the curve and cross correlation metrices.

3. RESULTS

The results have been computed on 3 surveillance videos from surveillance cameras. Video 1 and 2 are from the iLIDS database of AVSS 2007 conference and video 3 is from store survellance camera. Video 1 and 2 contain the videos of surveillance camera on a train station. Video 1 has less activity of passengers on the train station and video 2 has more activity on the train station. Video 3 has a view of a store with a man coming into the scene, picking something and lefting the scene. The frame size we used is 600x800, for the fast computation of ROC curve, we have downsampled the gaze maps and saliency maps by 4. For illustration, results of one video frame is shown in figure 3 for frame no 175 of the video 1.

Area under the curve(AUC) is computed between the saliency maps (SM, MSM, PSM and PVSM) and gaze maps for all videos as shown in the figures 4 and 5. Table 1 shows the mean AUC value computed between saliency maps and gaze maps for the same surveillance video.

The table 2 shows the correlation coefficient computed between corresponding saliency maps and gaze maps. The mean correlation coefficient value is computed by averaging the correlation coefficient values for the whole video sequence.

Table 1: Mean Area Under the curve for Saliency maps with Gaze maps.

Saliency Map	Mean AUC for Video 1
Itti SM with face detection(SSM)	0.5058
Motion SM (MSM)	0.6750
Predictive SM (PSM)	0.8089
Average of SM, MSM,and PSM (PVSM)	0.8087

Table 2: Mean Correlation of Saliency maps with Gaze maps.

Saliency Map	Correlation for Video 1
Itti SM with face detection(SSM)	0.1097
Motion SM (MSM)	0.2910
Predictive SM (PSM)	0.2898
Average of SM, MSM,and PSM (PVSM)	0.3031

3.1 Discussion

The average area under the curve (AUC) for MSM (Motion saliency map) shows that the motion saliency model is performing well. However there are some issues with the prediction saliency model. PSM and PVSM are overlapping in almost all the graphs of figures 4 and 5. Video 2 AUC graph shows that PSM and PVSM are better than static and motion saliency as shown in figure 5. However it is not the case in videos 1 and 3 AUC graphs as shown in figure 4.

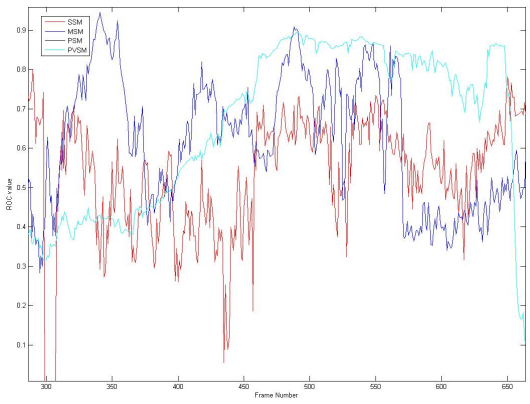

Figure 4: ROC Curve Video sequence - 1(frame 290 - 680).

As PVSM is computed by averaging all the saliency maps, it looks like average is not a good function to combine different saliency maps. In some cases it works for example in case of video 2 and some part of video 1 it works but not in the case of video 3. The major issue is how to combine the different saliency maps, most of the time in videos, motion and faces are more important than stationary low level features as shown in the gaze maps in figure 3(b). As can be seen in figure 3(c) the saliency map is highlighting the board on the right side of the frame. But in the gaze map its not salient. It might be because the board is not so salient or the viewer looked at it in some different frame. But if we combine by averaging the stationary saliency with motion saliency maps, this board will remain salient through out the video, however the gaze maps of the video show that motion and faces always attract the attention of the viewer. This saliency model can be used to detect some unusual events, for example detecting some un-attneded bag on a platform as encircled in the Gaze map figure 3(b), is detected as salient, encircled in the figure 3(c,f). Motion saliency computation

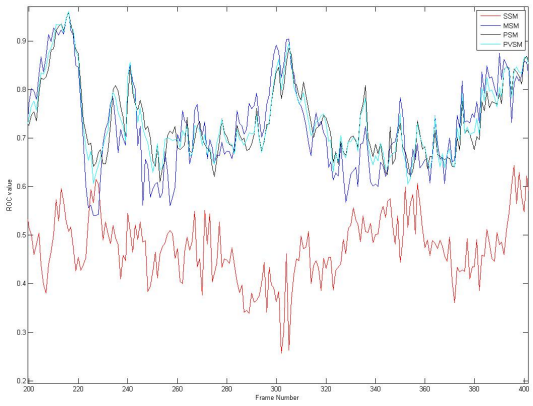

Figure 5: ROC Curve Video sequence - 2(frame 200 - 400).

is also a challenge: which motion is salient and which is not? For example if we see the frames in figures 3(a), there are people coming towards the camera and one person is going to the right side away from camera. In this case the motion vectors show high motion on the feet of these people, which

is quite natural. However if we look at the gaze maps of these frames, subjects are more interested in the upper part of the body or faces. The faces are also detected with the help of face detection model used in our model. When the people are getting closer to the camera then we will get high motion but its not the case when people are far away from the camera. That is why the more prominent object motion dominates the less prominant ones.

Furthermore, as in most surveillance scenerios the background scene is static. So it could be possible to reduce the saliency computation time by computing the background saliency once after few hundreds frames. While motion, face and prediction should be done for each frame. There are fast face detection and motion vector computing algorithms. Prediction is not very complex if we are working with a limited number of history frames.

4. CONCLUSION AND FUTURE DIRECTIONS

A predictive visual saliency model is proposed in this paper for surveillance videos, that use low-level as well as high-level features, and motion information. A modified model for motion saliency map is combined with the proposed predictive saliency model. The results show good correlation of the proposed saliency model with the gaze maps. However there are still some issues to be addressed for example how to combine the different modules of this model to achieve better results as average function does not seems to be optimal. In the next paper we are targetting to use neural networks as a method to combine the static conspicuity maps, motion saliency map and the predictive saliency map. We need to investigate further the predictive saliency map, by using multiple frames from the history to get better prediction. The gaze maps show motion saliency can be important but needs more effort to compute all the salient parts of a video frame while avoiding non-salient motion.

REFERENCES

[1] L. Itti, Ph.D. thesis, Models of Bottom-Up and Top-Down Visual Attention, California Institute of Technology, Pasadena, California, 2000.

[2] L. Itti and C. Koch, Computational modeling of visual attention., Neuroscience 2001 2(3), 194 (2001).

[3] L. Itti, C. Koch, E. Niebur, A Model of Saliency-Based Visual Attention for Rapid Scene Analysis, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 20, No. 11, pp. 1254-1259, Nov 1998.

[4] M. Cerf, E. P. Frady, C. Koch, Faces and text attract gaze independent of the task: Experimental data and computer model. Journal of Vision 2009, 9(12):10, 1-15.

[5] J. M. Wolfe, KR. Cave, SL. Franzel, Guided search: an alternative to the feature integration model for visual search, J Exp Psychol Hum Percept Perform. 1989 Aug;15(3):419-33.

[6] J. M. Wolfe, Visual Search in Continuous, Naturalistic Stimuli. Vision Research. 1994, 34 (9), 1187-1195.

[7] J. M. Wolfe, Visual Memory: What do you know about what you saw? Current Biology, 1998, 8: R303-R304.

[8] J. Harel, C. Koch, and P. Perona, Graph-based visual saliency, Advances in Neural Information Processing Systems (NIPS 2006), pp. 545-552.

[9] U. Rajashekar, I. van der Linde, A. C. Bovik, and L. K. Cormack, GAFFE: A Gaze-Attentive Fixation Finding Engine, IEEE Transcations on Image Processing 17, 564 (2008).

[10] P. Sharma, F. A. Cheikh, and J. Y. Hardeberg, Face Saliency in Various Human Visual Saliency Models, Proc. of 6th Int. Sym. on Image and Signal Proc. and Analysis (2009), vol. 16, pp. 332-337.

[11] L. Itti, Quantifying the contribution of low-level saliency to human eye movements in dynamic scenes, Visual Cognition 2005, vol. 12, pp. 1093-1123.

[12] L. Wixson. 2000. Detecting Salient Motion by Accumulating Directionally-Consistent Flow. IEEE Trans. Pattern Anal. Mach. Intell. 22, 8 (August 2000), 774-780.

[13] L. Wixson and M. Hansen. 1999. Detecting Salient Motion by Accumulating Directionally-Consistent Flow. In Proceedings of the International Conference on Computer Vision-Volume 2 - Volume 2 (ICCV '99), Vol. 2.

[14] Y. L. Tian and A. Hampapur, Robust Salient Motion Detection with Complex Background for Real-Time Video Surveillance. In Proceedings of the IEEE Workshop on Motion and Video Computing (WACV/MOTION'05) - Vol. 2.

[15] B. C. Motter, and E. J. Belky, The guidance of eye movemnets during active visual search. Vision Research 1998, 38(12), 1805-1815.

[16] J. Shen, E. M. Reingold, and M. Pomplun, Distractor ratio influences the patterns of eye movements during visual search. Perception 2000, 29(2), 241-250.

[17] R. Desimone, TD. Albright, CG. Gross and C. Bruce, Stimulus selective properties of inferior temporal neurons in the macaque, Journal of Neuroscience, vol4, 2051-2062, 1984.

[18] D. Walther, Koch, Modeling Attention to Salient Proto-objects, Neural Networks 19, 1395-1407, 2006.

[19] Y. F. Ma, H. J. Zhang, A model of motion attention for video skimming, Image Processing. 2002. Proceedings. 2002 International Conference on , vol.1, no., pp. I-129-I-132 vol.1, 2002.

[20] Q. Ma and L. Zhang. Saliency-Based Image Quality Assessment Criterion, Proceedings of ICIC 2008, LNCS 5226, pp. 1124-1133, 2008.

[21] B. D. Lucas and T. Kanade, An iterative image registration technique with an application to stereo vision. Proceedings of Imaging Understanding Workshop 1981, pages 121–130

[22] T. Jost, N. Ouerhani, R. von Wartburg, R. Mri and H. Hgli. Assessing the contribution of color in visual attention, Compute. Vis. Image Underst. Vol.100, No.1, pp.107-123, 2005.

[23] M. Cerf, E. P. Frady, and C. Koch, Faces and text attract gaze independent of the task: Experimental data and computer model. Journal of Vision 2009, 9(12):10, 1-15.

[24] D. Mahapatra, S. Winkler, and S. C. Yen, Motion saliency outweighs other low-level features while watching videos. Proc. SPIE 2008, 6806, 68060P , DOI:10.1117/12.766243.

[25] A. Belardinelli, F. Pirri, and A. Carbone, Motion Saliency Maps from Spatiotemporal Filtering. In Attention in Cognitive Systems, Lecture Notes In Artificial Intelligence, Vol. 5395. Springer-Verlag, 112-123.

[26] F.F.E. Guraya, F.A. Cheikh, A. Tremeau, Y. Tong, H. Konik, Predictive Saliency Maps for Surveillance Videos, In Conference of Distributed Computing and Applications to Business Engineering and Science DCABES, pp.508-513, 10-12 Aug. 2010

[27] P. H. Tseng, R. Carmi, I. G. M. Cameron, D. P. Munoz and L. Itti, Quantifying center bias of observers in free viewing of dynamic natural scenes, Journal of Vision, 9(7), pp 1-16, 2009.