2010 Ninth International Symposium on Distributed Computing and Applications to Business, Engineering and Science

Predictive Saliency Maps for Surveillance Videos

Fahad Fazal Elahi Guraya, Faouzi Alaya Cheikh
Dept of Computer Science and Media Technology
Gjovik University College, HIG
Gjovik, Norway
Email: {fahadg,faouzi}@hig.no

Alain Tremeau, Yubing Tong, Hubert Konik
Laboratoire Hubert Curien
University of Saint Etienne
Saint Etienne, France
Email: {alain.tremeau, yubing.tong,
hubert.konik}@univ-st-etienne.fr

***Abstract*—When viewing video sequences, the human visual system (HVS) tends to focus on the active objects. These are perceived as the most salient regions in the scene. Additionally, human observers tend to predict the future positions of moving objects in a dynamic scene and to direct their gaze to these positions. In this paper we propose a saliency detection model that accounts for the motion in the sequence and predicts the positions of the salient objects in future frames. This is a novel technique for attention models that we call Predictive Saliency Map (PSM). PSM improves the consistency of the estimated saliency maps for video sequences. PSM uses both static information provided by static saliency maps (SSM) and motion vectors to predict future salient regions in the next frame. In this paper we focus only on surveillance videos therefore, in addition to low-level features such as intensity, color and orientation we consider high-level features such as faces as salient regions that attract naturally viewers attention. Saliency maps computed based on these static features are combined with motion saliency maps to account for saliency created by the activity in the scene. The predicted saliency map is computed using previous saliency maps and motion information. The PSMs are compared with the experimentally obtained gaze maps and saliency maps obtained using approaches from the literature. The experimental results show that our enhanced model yields higher ability to predict eye fixations in surveillance videos.**

***Index Terms*—saliency map for videos; motion saliency; video surveillance; predictive saliency maps;**

I. Introduction

Human visual system (HVS) plays an important role in reducing brain's activity to quickly focus on certain regions within a scene. The peripheral sensors in the human visual system continuously generate numerous signals. Treating all of them at the same time is computationally expensive to achieve by the human brain. This results in the selective processing of the available information. The selected stimuli is also prioritized by our nervous system; via a process called selective attention. These select regions form a saliency map which can be used to prioritize the processing of information from them. This may be of crucial importance in surveillance applications for instance where suspicious behavior or unusual objects in a surveillance videos must be detected and analyzed with top priority. These estimated select regions are used to predict where one's attention will be drawn when viewing a video scene or an image.

Human eye movements are found to be tightly coupled with the visual attention [1]. There are two types of cues that humans give direct attention to - one is bottom-up and the other one is top-down [2], [3]. Bottom-up cues rely on the low level features such as intensity, color, orientation to compute the conspicuity maps while the top-down model uses faces, objects, and people as high level features. These can be used to compute the attention model [4]. GBVS [5] used graph theory to concentrate mass on activation maps. Low level features such as color, intensity and orientation are used to form the activation maps. Similarly four low level features are used in GAFFE [6] that uses luminance, contrast, and their bandpass filtered versions to generate saliency map. It has been observed that subjects in free-viewing conditions look at faces 16.6 times more then to similar regions normalized for the size and position of the face [7]. Face detection was introduced in [8] to improve the short comings of both GBVS and GAFFE when computing saliency model. The performance of these models were improved with the addition of face detection and hence correlate better with gaze maps.

In addition to low level and high level features, motion also plays an important role in defining salient regions, when considering videos. It is quite natural for the human visual system to focus on the moving objects in a video sequence. So in case of video sequences it is important to incorporate the motion history information into the saliency model. Motion can be also categorized into background and foreground motion, and a relative motion model like [20] can be added with saliency map.In this paper we propose a predictive saliency map combining motion information with the static saliency information to better model the saliency in video sequences and to predict the position of salient regions already detected in previous frames. A video saliency model based on stationary and motion information had been proposed in [19]. The saliency models could be used in several applications such as perceptual quality evaluation of images [17], [18] and videos [16], video compression, etc.

The rest of the paper is organized as follows: In the next section we discuss our proposed predictive saliency models. Section 3 presents the subjective psychophysical tests followed by the results in Section 4. The last section concludes the paper with some future directions.

II. Spatio-temporal saliency model based on low and high level features

In this paper we propose a predictive method to combine the saliency maps for surveillance videos using static saliency

DOI 10.1109/DCABES.2010.160

and motion information. The saliency computation model for videos is shown in Figure. 4. Our method computes the video saliency map based on stationary and motion information. When we compute saliency maps from still images, we deal with 2-dimension images where we only need a stationary saliency map, whereas in case of videos we also have to consider the third dimension i.e. temporal dimension. The evolution of objects in time in a video sequence gives the illusion of motion of the objects. Moving objects tend to capture our attention and thus is very important to account for in video saliency maps.

In our proposed models PSM is computed by combined saliency map (SM) and motion vectors. Combined saliency map (SM) is a combination of stationary saliency maps and/or motion saliency map based on function f as described in equation (11). In the next two paragraphs we explained how we have computed Predictive Saliency Map (PSM) and Predictive Video Saliency Map (PVSM).

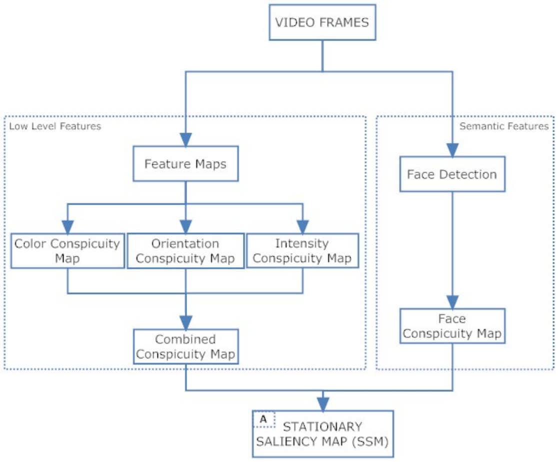

Fig. 1. Stationary saliency map model with face detection.

A. Stationary saliency map

Stationary saliency map (SSM) is composed of two parts, saliency due to low level features such as color, intensity and orientation and that due to high level features such as face as shown in Figure 1. Itti's bottom-up attention model [2], [3] is used to compute low level features (color, intensity, and orientation) conspicuity maps. Seven conspicuity maps, one for intensity (C_i), four for orientations 0, 45, 90 and 135 degrees (C_o), and two for color combinations Red-Green & Blue-Yellow (C_c), are generated. These conspicuity maps are combined, after a normalization step, as shown in the equation (1).

$$C_{itti} = \frac{1}{7}(C_i + 2C_c + 4C_o) \qquad (1)$$

Psychological studies show that faces, heads, and hands attracts human attention [11]. Text also attracts human gaze independently of the task [13]. These are however not considered in Itti's model. Due to the importance of faces in surveillance applications, face conspicuity map will be added to Itti's stationary saliency map. In this paper, we have used Walther et al. face detection model [10] to compute face conspicuity map. This face detection algorithm is based on the computation of a Gaussian model for skin hue color distribution.

Itti's low level feature's conspicuity maps can be combined with face conspicuity maps as in equation (2).

$$SSM = f(C_{itti}, C_{face}) \qquad (2)$$

The f function has been defined empirically. In [18] we proposed to use a linear combination of face conspicuity map and Itti's conspicuity map as shown in the equation (3). We proposed to use the following weighting parameters as for the Itti's model. The most accurate saliency maps that we get from the set of surveillance video sequences that we used was obtained with the following weights in equation 3:

$$SSM = \frac{1}{8}(2C_i + 2C_c + C_o + 3C_F) \qquad (3)$$

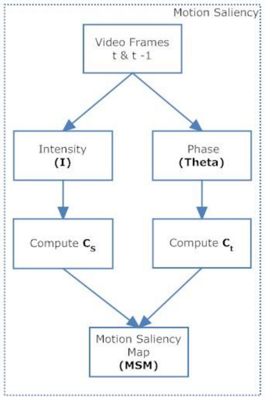

Fig. 2. Motion Saliency Model.

B. Motion saliency map

Motion saliency dominates other low level features' saliency in video sequences [14]. Motion saliency information is thus added to the proposed saliency model. We proposed in [18] to use the motion attention model based on spatial-temporal entropy proposed by [15] to compute the motion saliency map. The motion saliency computational model is described in figure 2.

Motion saliency map is computed using three inductors from motion vectors, i.e. intensity of the motion, spatial coherence and temporal phase coherency, as proposed by [15]. These three inductors are defined by the motion vectors between reference and target frames. Motion vectors are shown in figure 3. They are computed at each location of macro blocks. The Intensity Inductor induces motion energy or activity that can be defined by:

$$I_{i,j} = \frac{\sqrt{dx_{i,j}^2 + dy_{i,j}^2}}{Max(MotionVectorsMagnitude)} \qquad (4)$$

where $(dx_{i,j}, dy_{i,j})$ denote x and y (i.e. horizontal and vertical) components of motion vector.

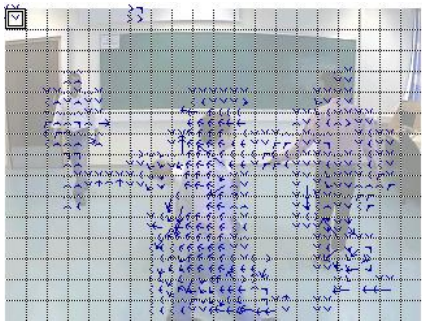

Fig. 3. Representation of Motion Vectors.

Spatial phase coherence is the second inductor that induces spatial consistency of motion vectors in motion saliency map. Spatial phase coherency $C_s(i,j)$ is defined by equation (5).

$$Cs(i,j) = \sum_{s=1}^{n} p_s(t) log(p_s(t)) \qquad (5)$$

where

$$p_s(t) = SH_{i,j}^{w}(t) / \sum_{k=1}^{n} SH_{i,j}^{w}(k) \qquad (6)$$

where $SH_{i,j}^{w}(t)$ is the spatial phase histogram of the probability distribution function $p_s(t)$, and n is the number of histogram bins.

Lastly, the third inductor is defined by the temporal phase coherency $C_t(i,j)$ computed from a temporal sliding window of L frames. This temporal phase coherency is defined by equation (7).

$$Ct(i,j) = \sum_{i=1}^{n} p_t(t) Log(p_t(t)) \qquad (7)$$

and

$$p_t(t) = TH_{i,j}^{L}(t) / \sum_{k=1}^{n} TH_{i,j}^{L}(k) \qquad (8)$$

where $TH_{i,j}^{l}(t)$ is the temporal phase histogram of the probability distribution function $p_t(t)$, and n is the number of histogram bins.

The motion saliency map (MSM) is then computed as in [15] by combining the three motion inductors I, Cs and Ct as in equation 9.

$$MSM = I * C_t(1 - I * C_s) \qquad (9)$$

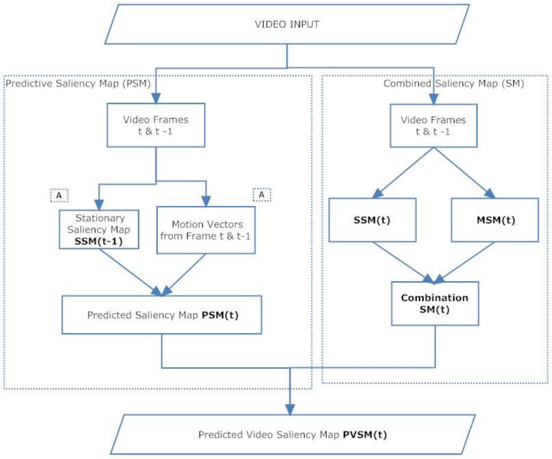
VIDEO INPUT
Predictive Saliency Map (PSM)
Combined Saliency Map (SM)
Video Frames t & t -1
Video Frames t & t -1
A
A
Stationary Saliency Map SSM(t-1)
Motion Vectors from Frame t & t-1
SSM(t)
MSM(t)
Predicted Saliency Map PSM(t)
Combination SM(t)
Predicted Video Saliency Map PVSM(t)

Fig. 4. Flowchart of the Predictive Video Saliency Model (PVSM).

C. Predictive saliency model

Human attention focuses on stationary salient objects as well as to moving objects in a video sequence. Therefore, we propose to combine motion saliency maps (MSM) and stationary saliency maps (SSM) in such a way to minimize the rate of false detection of salient regions and to minimize the rate of false detection of non-salient regions. We propose to compute SSM from low level features and high level features and to compute MSM only from the motion information between consecutive frames (i.e. motion vectors). The problem of stationary saliency maps (SSM) is that when objects evolve in the 3-D space the stationary saliency maps are not consistent. This problem is due to the fact that stationary saliency maps are extracted from each frame separately from the other frames in the sequence. To overcome this problem, motion information can be used to estimate the next position of a salient region in the future frame. An example of such case of study is when the face detector fails to find a face due to a slight rotation. The predictive saliency model (PSM) that we propose here is computed for each frame of the video from motion vectors and stationary saliency map. The motion vectors are computed using motion vector blocks matching algorithm between reference and target frames. The reference video frame is divided into blocks of size 16x16 pixels. Then each of the blocks in the reference frame is searched in the target frame within a search window. Next, the closest block found which matches the current block is used to compute the motion vector between the previous position of the block in the reference frame and the current position of the block in the target frame. These vectors are called motion vectors. An example of motion vectors is shown in Figure 3. The obtained motion vectors show the displacement of a block in the target frame to its origin in the reference frame.

To compute the PSM of the frame t of a video sequence, we need to compute firstly the final saliency map of the previous frame $FSM(t-1)$ and the motion vectors between the frames $F(t-1)$ and $F(t)$. We propose to compute $PSM(x, y, t)$ by changing the position of the 16x16 block of previous $FSM(x, y, t-1)$ to the new position defined by the motion vector. Thus the predicted saliency map for the current frame at time t is based on the computation of the previous FSM saliency map and of the motion vectors. This predicted saliency map gives the new position of each block in the current frame. Next, the $PSM(x, y, t)$ is combined linearly with $SSM(x, y, t)$ to account for the motion saliency. This gives us a predictive video saliency map (PVSM) as shown in figure 4. We propose to compute PVSM as the linear combination of PSM with SM as in equation (10):

$$PVSM(x, y, t) = \alpha * PSM(x, y, t) + (1-\alpha) * SM(x, y, t) \tag{10}$$

where $\alpha = 0.5$, We propose to compute SM as a combination of SSM and MSM as in equation (11).

$$SM(x, y, t) = f(SSM(x, y, t) + MSM(x, y, t)) \tag{11}$$

where f could be $MEAN, MAX, AND$ or a linear combination function. In this paper we have used mean function to combine the SSM and MSM.

III. Experimental Setup

We have conducted an experiment in order to see where observers look when they are viewing images and videos under standard viewing conditions. The results from the experiment have been analyzed by computing the average Gaze Map (MP) of observers. The experiment details are given in the next section.

A. Gaze maps

In our experiments we mainly used indoor surveillance videos recorded by ourselves with people moving inside a static background. These experiments have confirmed that indeed the attention of observers is in general strongly attracted by faces. These gaze map were then compared to the results obtained from our visual perception model. The goal of this comparison is to study if the video saliency maps computed from our model are properly correlated to the gaze map derived from subjective experiments. To compute the gaze maps we did subjective experiments with an eye tracker. 20 observers, aged between 25 and 42, participed to the experiments done with a 50 Hz infra-red SMI eye tracker. During the experiments observers were asked to watch surveillance videos on a 17 inch CRT display as they normally would do under normal viewing conditions. The subjects were asked to watch the videos as they normally would do. The resolution of the display was of 1024x768 pixels. The distance between the monitor and the observer was between 60-70 cm. Before each experiment a test was performed to detect the dominant eye of the observer. During experiments observers' dominant eye was tracked and tracking data were saved with a system processing with the SMI IView software. Gaze maps were computed from fixation points of the dominant eye. Firstly, a fixation frequency map was computed for each frame of each video by adding up all the fixation positions of each observer. As with the Human Visual System the fixation frequency map was next filtered by a spatial Gaussian filter. It is important to find a suitable standard deviation σ for the Gaussian filter. These frequency maps were filtered by a spatial Gaussian filter of σ = 37 which was chosen to approximate the size of the viewing field corresponding to the fovea in the gaze map. All fixation points were taken into account. The size of the Gaussian window was of 40x40 pixels. Next, the average of these Gaussian maps for all observers was computed, then normalized and surimposed to the original frame with a colormap of 64 color values, where blue colors correspond to lowest gaze map values and red colors correspond to the highest gaze map values, i.e. the most salient regions of a video frame. An illustration of gaze maps and saliency maps is given in the figure 6; where figure 6 (a) is a video frame extracted from a surveillance video, figure 6 (b) is the corresponding gaze map derived from subjective experiments, figure 6 (c - f) correspond to the same video frame with different saliency maps surimposed.

IV. Results and Discussion

To study the performance of the predictive model we computed SSM, MSM, PSM and PVSM of indoor surveillance videos with people moving inside a static background. In this paper results shown concern one surveillance video sequence of 75 frames. These saliency maps have been compared to the results of the gaze maps obtained with the subjective experiment. The comparison was done by computing the area under the curve (AUC) and the mean correlation between computed a saliency map and the gaze map. The proposed saliency map PVSM was compared with SSM [8] which is a static saliency map model and with the motion saliency map (MSM) proposed by [15]. The mean area under the curve (AUC) and the mean correlation results are shown in tables I and II respectively. These results show the scores obtained with the MEAN and the AND functions (see columns at left and at right, respectively) used for combining SSM and MSM as in (11). As it can be seen from these tables the MEAN function performs better than the AND function, for MEAN function we get higher values with AUC for PSM and for PVSM, and when we use the AND function between SSM and MSM we get higher AUC values but lower correlation values.

The individual plots of AUC for SSM, MSM, PSM and PVSM are shown in Figure 5. In this graph, the x-axis shows the number of frames and y-axis shows the AUC value. This graph shows that our predictive saliency maps, i.e. PSM and PVSM, outperform the results of SSM and MSM for most of the frames. Similarly in table I, the mean AUC value for PVSM and PSM is almost 10% to 13% higher than the mean AUC of SSM or MSM.

Figure 6 shows respectively the original frame, gaze map, Itti's saliency map with face information, motion saliency map, predicted saliency map and predicted video saliency

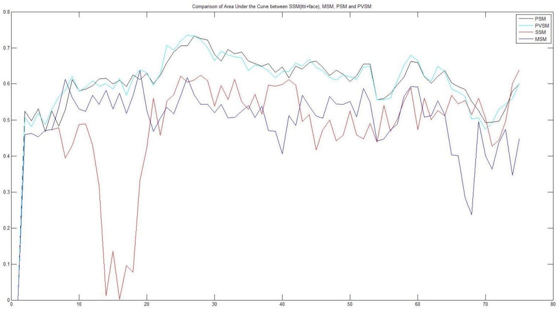

Fig. 5. Graph of Area Under the Curve (AUC) for SSM, MSM, PSM and PVSM.

TABLE I
MEAN AREA UNDER THE CURVE FOR SALIENCY MAPS.

Saliency Map	AUC for Mean	AUC for AND
Stationary SM(SSM)	0.4776	0.4776
Motion SM (MSM)	0.4994	0.4994
Predictive SM (PSM)	0.6047	0.563
Predictive Video SM (PVSM)	0.6046	0.5606

TABLE II
MEAN CORRELATION FOR SALIENCY MAPS.

Saliency Map	Correlation for Mean	Correlation for AND
Stationary SM(SSM)	0.0568	0.0568
Motion SM (MSM)	0.0531	0.0531
Predictive SM (PSM)	0.0886	0.043
Predictive Video SM (PVSM)	0.0898	0.0441

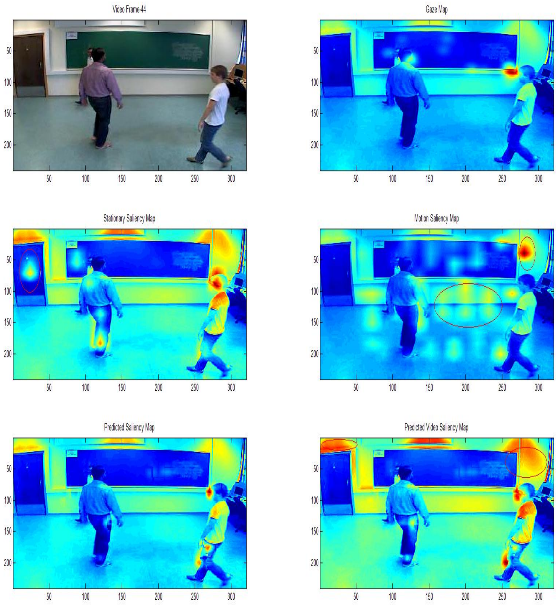

Fig. 6. Computed saliency maps.

map of one frame of the surveillance video used to illustrate this paper. The gaze map and the saliency maps have been surimposed to the original frame to highlight areas of interest. We have drawn on the SSM image a red ellipse (at left) in order to show a salient area in the background which is due to illumination variations. Similarly we have drawn on the MSM image a red ellipse (near the center) in order to show a salient area based a small motion which is due to background illumination changes. These false salient regions in SSM and MSM are not present when using the PSM, in this case only true salient regions are detected. The results shown in this Figure are computed with AND function between SSM & MSM. And due to this history information we managed to predict the next frame saliency. Let us note here that in case of PVSM we get also some salient regions in the background due to illumination changes. However these background false salient regions are successfully removed in case of PSM. Currently we are predicting PSM based on only one previous frame's SM. It may be a good idea to predict the PSM from few more previous SM. Whatever, our result show that PSM performs better than SSM and MSM, in case of computing video saliency maps.

V. CONCLUSION AND FUTURE WORK

In this paper we have proposed a saliency detection model that accounts for the motion in the surveillance video sequence and predicts the positions of the salient objects in future frames. This novel technique based on attention models that we call Predictive Saliency Map (PSM) improves the consistency of the estimated saliency maps for video sequences. PSM uses the static information provided by static saliency maps (SSM) and motion vectors to predict the future salient regions in the next frame. In this paper we focused on surveillance videos, therefore, in addition to low-level features such as intensity, color and orientation we consider high-level features such as faces as faces are salient regions that attract easily viewer's attention. Furthermore, saliency maps computed based on these static features are combined with motion saliency maps to account for saliency created by the activity in the scene. The proposed PSM has been compared with the experimentally obtained gaze maps and saliency maps obtained using approaches from the literature. The experimental results show that our predictive model combined with motion vectors yields higher performance to predict eye fixations in surveillance videos. The next step of our study will consist to test and to extend this video saliency model on other sets of videos such as for example outside videos or inside videos with camera motion or with other moving objects than peoples. It is also proposed as future work to test different types of fusion techniques between SSM and MSM. And these combined saliency maps should be used to compute PSM. In this paper we have used only one frame history, however longer history information may improve the results.

REFERENCES

[1] T. Jost, N. Ouerhani, R. V. Wartburg, R. Muri, and H.Hugli, Computer Vision and Image Understanding, Elsevier 100,107 (2005).

[2] L. Itti, Ph.D. thesis, California Institute of Technology,Pasadena, California (2000).

[3] L. Itti and C. Koch, Neuroscience 2001 2(3), 194 (2001).

[4] P. Sharma, F. A. Cheikh, and J. Y. Hardeberg, in Sixteenth Color Imaging Conference (The Society for Imaging Science and Technology, 2008), vol. 16, pp. 332-337.

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

[6] U. Rajashekar, I. van der Linde, A. C. Bovik, and L. K. Cormack, IEEE Transcations on Image Processing 17, 564 (2008).

[7] Cerf, M., Frady, E. P., Koch, C. (2009). Faces and text attract gaze independent of the task: Experimental data and computer model. Journal of Vision, 9(12):10, 1-15, http://journalofvision.org/9/12/10/, doi:10.1167/9.12.10

[8] Puneet, Sharma, Saliency Maps & Eye Tracking, Master's thesis, Gjvik University College, Norway, 2008.

[9] Brian Michacel Scacellat. "Theory of Mind for a Humanoid Robot", Autonomous Robert, vol. 12, No.1, pp.13-24, 2002.

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

[11] 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.

[12] Yufei Ma, Hongjing Zhang. A model of motion attention for video skimming. Vol.1, pp.22-25, ICIP 2002.

[13] Cerf, M., Frady, E. P., and Koch, C. (2009), "Faces and text attract gaze independent of the task: Experimental data and computer model". Journal of Vision, 9(12):10, 1-15, http://journalofvision.org/9/12/10/, doi:10.1167/9.12.10.

[14] Dwarikanath Mahapatra, Stefan Winkler, and Shih-Cheng Yen, "Motion saliency outweighs other low-level features while watching videos". Proc. SPIE 6806, 68060P (2008), DOI:10.1117/12.766243.

[15] Yu-Fei Ma; Hong-Jiang 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.

[16] Fahad F. E. Guraya, A. Shariq, Y. Tong, F. Alaya Cheikh, "A non-reference perceptual quality metric based on visual attention model for videos," Information Science and Signal Processing, ISSPA. 2010, International Conference on.

[17] Ali Shariq, Fahad F. E. Guraya, F. Alaya Cheikh, "A visual attention based reference free perceptual quality metric," Accepted for publication in European workshop on visual information processing, EUVIP. 2010, Paris, France.

[18] Y. Tong, F. Alaya Cheikh, A. Tremeau and H. Konick, "Full Reference Image Quality Assessment Based on Saliency Map Analysis," Accepted for publication in the International Journal of Imaging Systems and Technology, in 2010.

[19] Y. Tong, F. Alaya Cheikh, Fahad F. E. Guraya and A. Tremeau, "A Visual Saliency Model for Perception-based Video Surveillance," Accepted to Visual Communications and Image Processing VCIP 2010, China.

[20] Zhou Wang and Qiang Li, "Video quality assessment using a statistical model of human visual speed perception," Journal of the Optical Society of America A 24, B61-B69 (2007)