2013 *International Conference on Multimedia and Expo* (ICME2013), San Jose, CA, USA, July 2013

VIDEO SALIENCY INCORPORATING SPATIOTEMPORAL CUES AND UNCERTAINTY WEIGHTING

Yuming Fang[1], *Zhou Wang*[2], *Weisi Lin*[1]

[1]School of Computer Engineering, Nanyang Technological University, Singapore
[2]Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON, Canada
fa0001ng@e.ntu.edu.sg, Z.Wang@ece.uwaterloo.ca, wslin@ntu.edu.sg

ABSTRACT

We propose a method to detect visual saliency from video signals by combing both spatial and temporal information and statistical uncertainty measures. The main novelty of the proposed method is twofold. First, separate spatial and temporal saliency maps are generated, where the computation of temporal saliency incorporates a recent psychological study of human visual speed perception, where the perceptual prior probability distribution of the speed of motion is measured through a series of psychovisual experiments. Second, the spatial and temporal saliency maps are merged into one using a spatiotemporally adaptive entropy-based uncertainty weighting approach. Experimental results show that the proposed method significantly outperforms state-of-the-art video saliency detection models.

Index Terms— visual attention, video saliency, spatiotemporal saliency detection, uncertainty weighting

1. INTRODUCTION

Selective visual attention or visual saliency has been an active research topic in the past decades in the fields of biology, psychology and computer vision. It has also attracted a great deal of attention recently in the multimedia field because of its potential applications in the evaluation and improvement of quality-of-experience (QoE) in multimedia communication systems. According to the Feature Integration Theory (FIT) developed by Treisman *et al.* [1] in the 1980s, the early selective attention mechanism leads some image regions to be salient for their different features (color, intensity, orientation, motion, etc.) from their surrounding regions [1]. Koch *et al.*'s visual attention model [2] suggests that selective visual attention includes three stages: elementary parallel feature representation across the visual field; the Winner-Take-All (WTA) mechanism singling out the most salient location; and the routing selection for the next most salient locations. Recently, computer vision researchers proposed various computational saliency detection models for images. Compared with saliency detection in still images, video saliency detection is a more difficult problem due to the complication in the detection and usage of temporal and motion information.

Only a limited number of algorithms have been proposed for spatiotemporal saliency detection from video signals [3, 4, 5, 6, 7, 8]. Itti *et al.* utilized a Bayesian model to detect surprising events as important information attracting human attention, where the surprise is measured by the difference between posterior and prior beliefs of the observer [3]. Ma *et al.* integrated top-down mechanisms into classical bottom-up saliency detection models for video summarization [4], where the top-down information includes semantic cues such as face and speech. Zhai *et al.* linearly combined spatial and temporal saliency maps [5], where the saliency maps are computed based on color histograms and the planar motion between images, respectively [5]. Le Meur *et al.* extended their saliency model for images by adding temporal saliency information into the framework [6]. Mahadevan *et al.* incorporated motion-based perceptual grouping and the discriminant formulation of center-surround saliency [7]. Guo *et al.* represented image pixels using quaternion intensity, color and motion features and employed the phase spectrum of Quaternion Fourier Transform to calculate spatiotemporal saliency [8]. Seo *et al.* introduced the notion of self-resemblance to measure visual saliency from video signals [15].

A key issue in video saliency evaluation is how to quantify the contribution of motion information, for which existing models tend to use ad-hoc methods with little justification from psychological or physiological studies. Our work is inspired by a recent study by Stocker *et al.* regarding human visual speed perception [9], where a set of psychovisual experiments were carried out to measure the prior probability distribution and likelihood function of visual speed perception. These measurements are consistent across human subjects and can be modeled by simple parametric functions. These results allow us to quantify the surprisal or motion information content in a perceptually meaningful way and use it as a predictor of motion visual attention. Another important problem in the development of spatiotemporal saliency models is how to combine spatial and temporal saliency maps when both of them are available. Unlike existing approaches

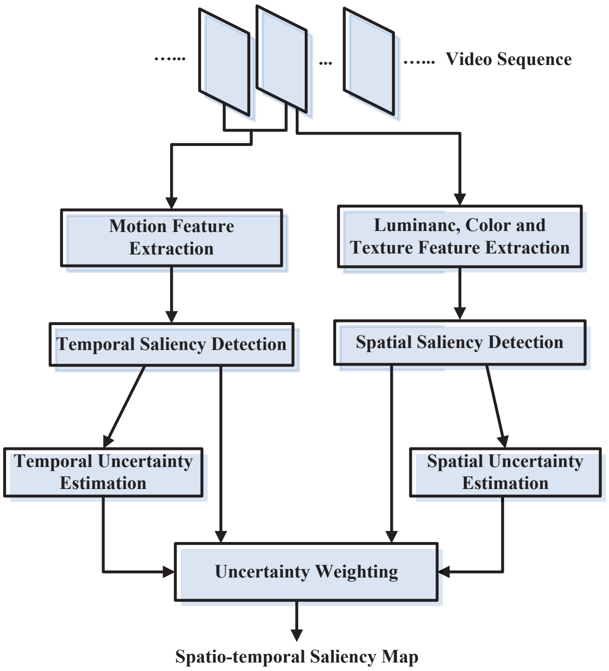
......
...
...... Video Sequence
Motion Feature Extraction
Luminanc, Color and Texture Feature Extraction
Temporal Saliency Detection
Spatial Saliency Detection
Temporal Uncertainty Estimation
Spatial Uncertainty Estimation
Uncertainty Weighting
Spatio-temporal Saliency Map

Fig. 1. Framework of the proposed model.

that often use simple combination rules such as linear combination with fixed weights, we associate each saliency map with a uncertainty map obtained from statistics of human saliency data and merge the saliency maps adaptively based on the local uncertainty measures. In the next two sections, we describe the proposed algorithm and demonstrate its effectiveness in achieving improved accuracy in predicting human saccade in viewing video signals.

2. PROPOSED METHOD

The general framework of the proposed model is depicted in Fig. 1. Low-level spatial and motion features are first extracted from the input video sequence, where the spatial features (including luminance, color and texture) and the motion feature are used to calculate the spatial and temporal saliency maps, respectively. The spatial and temporal uncertainty maps are then calculated to assess the confidence of the corresponding saliency maps. Finally, the spatial and temporal saliency maps are fused using an uncertainty weighting approach, resulting in the final spatiotemporal saliency map.

2.1. Spatial Saliency Evaluation

The spatial saliency detection method basically follows the method for still image saliency estimation introduced in [10] (with modifications) and is briefly described here.

Given a video frame, we first convert all image pixels into the YCbCr color space and divide the frame into non-overlapping 8×8 patches. Four features are extracted from each patch, including one luminance feature L (DC value of the Y component), two color features C_1 and C_2 (DC values of the Cb and Cr components), and one texture feature T (total AC energy of the Y component). These patch-based features extracted across space constitute four feature maps.

Assuming saliency is associated with the surprisal of the current patch against its neighboring patches in terms of certain image features, we use a contrast-of-feature approach to estimate patch saliency. The saliency value S_i^k for patch i based on the contrast of feature k is calculated as:

$$S_i^k = \sum_{j \neq i} \left[\frac{1}{\sqrt{2\pi}\sigma} e^{-l_{ij}^2/2\sigma^2} \right] D_{ij}^k \tag{1}$$

where $k \epsilon \{L, C_1, C_2, T\}$, σ is a width parameter of the Gaussian weighting function, which is used to weight the absolute feature difference D_{ij}^k between patches i and j, and l_{ij} is the spatial distance between patches i and j. The value of σ determines the size of the neighborhood and thus the locality of the feature contrast measure.

Finally, the feature maps are normalized to $[0, 1]$ and the overall spatial saliency map of the video frame is calculated as the average of the four feature maps [10]:

$$S_s = \frac{1}{K} \sum_{k=1}^{K} N(S^k) \tag{2}$$

where N is the normalization operator and K is the number of features ($K = 4$).

2.2. Temporal Saliency Evaluation

Object motion is often highly correlated with visual attention [1]. Our temporal saliency evaluation algorithm starts with optical flow based motion estimation [11], which is more efficient and provides denser and smoother motion vector field compared with block matching-based motion estimation. The optical flow vector field indicates *absolute* local motion, but perceived object motion often corresponds to the *relative* motion between the object and the background. Generally, an object of strong motion with respect to the background would be a strong surprisal to the human visual system (HVS). If we consider the HVS as an efficient information extractor, it would pay more attention to such a surprising event. Therefore, visual attention of motion can be measured by the surprisal of motion, which can be estimated based on the perceptual prior probability distribution about the speed of motion. Recently, Stocker *et al.* measured the prior probability of human speed perception based on a series of psychovisual experiments [9]. The results have been employed in the field of perceptual video quality assessment [12], but have not been exploited in the context of visual saliency estimation. According to their results, the "perceptual" prior distribution of motion speed can be well fitted with a power-law function

$$p(v) = \kappa / v^{\alpha} \tag{3}$$

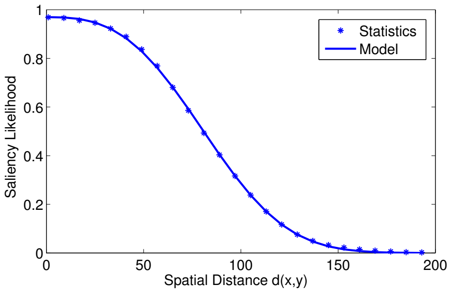

Fig. 2. Likelihood of saliency as a function of spatial distance from saliency center.

where κ and α are two positive constants. This suggests that with the increase of object speed, the probability decreases and thus the visual surprise increases. This also allows us to compute a motion speed-based temporal saliency value using its self-information or surprisal as

$$S_t = -\log p(v) = \alpha \log v + \beta \tag{4}$$

where $\beta = -log\kappa$ is a constant. The parameters α and β are chosen based on the study in [12].

It remains to compute v, which is the relative motion speed of the current position with respect to the background. To be aligned with the spatial saliency map, here we evaluate the relative speed v_i of the i-th patch as

$$v_i = \sum_{j \neq i} \left[\frac{1}{\sqrt{2\pi}\sigma} e^{-l_{ij}^2/2\sigma^2} \right] D_{ij}^v \tag{5}$$

where D_{ij}^v is the length of the vector difference between the mean absolute motion vectors of patches i and j. As in (1), a Gaussian weighting function is applied, which determines the impact of neighboring patches based on their distances to the current patch.

2.3. Uncertainty Evaluation

Depending on the visual content, the detected saliency based on spatial and motion features may have different levels of confidence or certainty across space and time. For example, a single moving object in a static background scene and with sharp color contrast with respect to the background may be detected as a salient object with high certainty, while the certainty drops dramatically when multiple objects with similar color and texture are moving at a similar speed. Here we propose to estimate such uncertainty in saliency evaluation and demonstrate its value in improving the accuracy of saliency detection.

Our uncertainty measure is based on two intuitive observations. First, the spatial location that is closer to the most concentrated saliency regions in an image is more likely to be a salient location. Second, a spatial location that is more connected to other saliency regions are more likely to be a salient location. These observations are justified by our empirical statistics of an image database created by Achanta *et al.* [13], which includes 1000 images and their corresponding ground truth salient objects selected by human subjects. Specifically, given an image and its ground truth saliency map S, we first compute the expected center location of its saliency map by

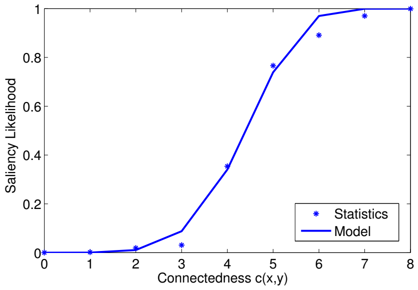

Fig. 3. Likelihood of saliency as a function of connectedness.

$$x_c = \frac{1}{M} \sum_{(x,y)\in R_S} x S_{x,y} \tag{6}$$

$$y_c = \frac{1}{M} \sum_{(x,y)\in R_S} y S_{x,y} \,. \tag{7}$$

where R_S is the set of all ground truth salient pixels and M is their total count. We can then compute the spatial distance d from the expected saliency center (x_c, y_c) to any location (x, y) in the image, and carry out statistics of the likelihood of being a salient pixel as a function of d. The statistical results are shown in Fig. 2. As expected, with the increase of d from the saliency center, the likelihood decreases. To describe this relationship efficiently, we find that the statistical data can be very well fitted with the following function

$$p(s|d) = \alpha_1 \exp\left[-\left(\frac{d}{\beta_1} \right)^{\gamma_1} \right] \tag{8}$$

where $p(s|d)$ stands for the likelihood of a pixel being salient given its distance d from the saliency center (x_c, y_c). α_1, β_1 and γ_1 are fitting parameters for the model and are found to be $\alpha_1 = 0.9694$, $\beta_1 = 93.30$, and $\gamma_1 = 2.8844$, respectively, based on the image database [13]. The fitting curve is also shown in Fig. 2. Given this likelihood model, a natural way to quantify the level of perceptual uncertainty is to compute the entropy of the likelihood:

$$U^d = H_b(p(s|d)) \tag{9}$$

where $H_b(p)$ is the binary entropy function computed as $-p \log_2 p - (1-p) \log_2 (1-p)$.

Another aspect that could have a significant impact on the saliency likelihood of a pixel is how it is connected to other salient pixels. For each pixel, we calculate its connectedness as

$$c = \sum_{(x,y)\in R_N} S_{x,y} \tag{10}$$

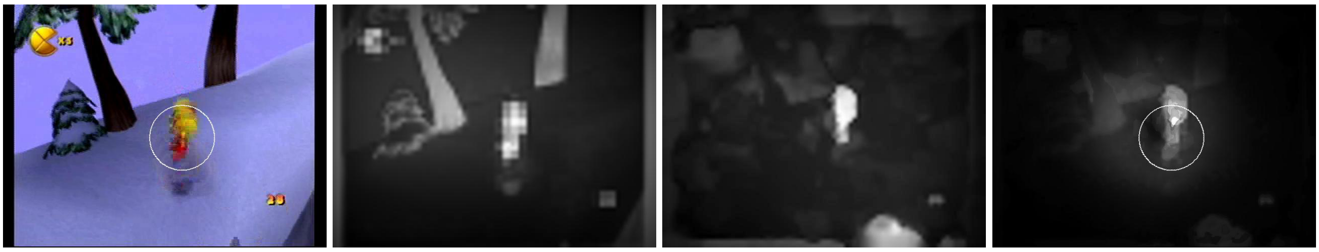

Fig. 4. Sample spatial, temporal and overall saliency maps. Column 1: original video frame with human fixation point marked with a circle; Column 2 - 4: spatial, temporal, and overall saliency maps, respectively.

where R_N represents the set of direct neighboring pixels near the current pixel, excluding itself. Based on the image database [13], we carried out statistics on the likelihood of a pixel being salient as a function of connectedness c, and the results are shown in Fig. 3. It can be observed that the more a pixel is connected to salient pixels, the more likely it is also a salient pixel. This relationship can also be summarized using an empirical function given by

$$p(s|c) = 1 - \exp\left[-\left(\frac{c}{\beta_2}\right)^{\gamma_2}\right] \tag{11}$$

where $p(s|c)$ represents the likelihood of a pixel being salient given its connectedness c to other salient pixels. β_2 and γ_2 are fitting parameters and are found to be $\beta_2 = 4.7262$ and $\gamma_2 = 5.2531$, respectively. The fitting function is shown in Fig. 3. Similarly, we can quantify the uncertainty using the entropy of the likelihood:

$$U^c = H_b(p(s|c)) \tag{12}$$

Finally, we can calculate the total uncertainty for each pixel in the image as

$$U = U^d + U^c \tag{13}$$

Applying such uncertainty computation to both spatial and temporal saliency maps computed in Sections 2.1 and 2.2, we obtain two uncertainty maps of each video frame, denoted as U_s and U_t, respectively.

2.4. Spatiotemporal Saliency Computation

The last step in creating an overall spatiotemporal saliency map is to combine the spatial and temporal saliency maps computed in Sections 2.1 and 2.2, respectively, which are also associated with different levels of uncertainty based on the computation in Section 2.3. Naturally, the saliency measure with lower uncertainty should be given larger weight. This leads to an uncertainty weighted fusion rule given by

$$S = \frac{U_t\, S_s + U_s\, S_t}{U_s + U_t} \tag{14}$$

Since both spatial and temporal uncertainty maps change over space and time, this fusion rule is spatiotemporally adaptive, which differentiate it from existing methods where fixed weighting is used to fuse spatial and temporal saliency maps. Figure 4 provides a sample video frame, together with its spatial, temporal and overall saliency maps. It can be observed that both spatial and temporal saliency maps are effective at identifying potential salient objects, and the fused overall saliency map successfully predicts the actual location of visual fixation.

3. EXPERIMENTAL EVALUATION

We use a publicly available video database [3] to evaluate the performance of the proposed model. The database contains 50 video clips totaling over 25 minutes with a variety of video content. The ground truth is obtained from human saccade data of 8 subjects recorded by an eye tracker. The performance of spatiotemporal saliency detection models is evaluated by comparing the response values at saccade and random locations in the saliency map [3]. Generally, an effective saliency detection model would have high response at saccadic locations and no response at most random locations. Here, the saliency distributions at saccadic and random locations are calculated with 10 bins of saliency values over the saliency map, as shown in Fig. 5. Kullback-Leibler (KL) distance is used to measure the similarity between these two distributions

$$KL(H,R) = \frac{1}{2}\left(\sum_n h_n \log\frac{h_n}{r_n} + \sum_n r_n \log\frac{r_n}{h_n}\right) \tag{15}$$

where H and R are saliency distributions at human saccadic locations and random locations with probability density functions h_n and r_n, respectively; n is the index of the saliency value bin ($n \in \{1, 2, 3..., 10\}$). The saliency detection model with larger KL distance can better discriminate human saccadic locations from random locations, and thus has better performance [3]. In addition, we use Receiver Operating Characteristics (ROC) curve [14] for performance evaluation. The saliency distributions at human saccadic locations and random locations are used as the test set and the discrimination set, respectively. The area under the ROC curve (AUC)

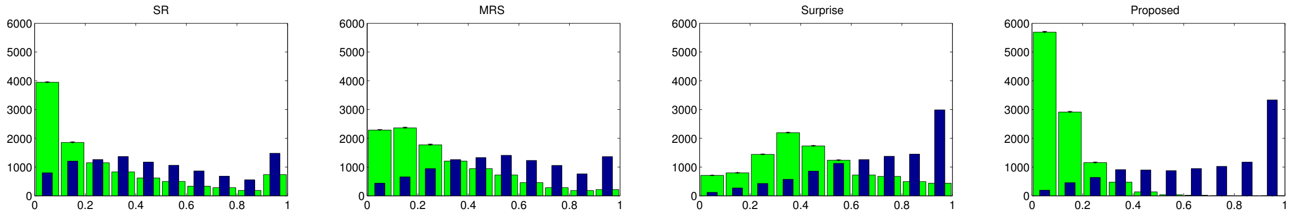

Fig. 5. Saliency distributions at human saccade locations (narrow blue bars) and random locations (wide green bars) from different spatiotemporal saliency models. The x- and y-axis represent the predicted saliency values from different models and histograms of the corresponding salient values, respectively.

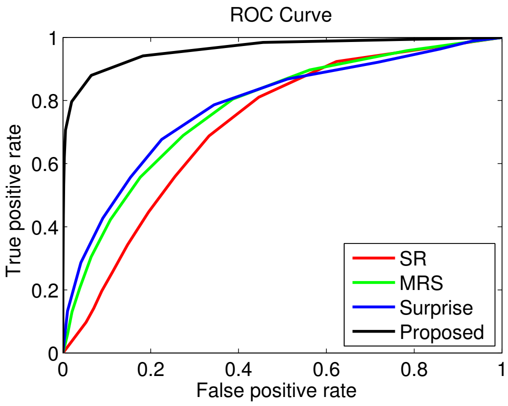

Fig. 6. ROC comparison of video saliency models.

Table 1. KL distance and AUC Comparisons of spatiotemporal saliency models.

Models	SR [15]	MRS [8]	Surprise [3]	Proposed
KL Dist.	0.391	0.529	0.593	2.584
AUC	0.722	0.771	0.782	0.951

provides an overall evaluation. A better video saliency detection model is expected to have a larger AUC value.

In addition to the proposed algorithm, three state-of-the-art spatiotemporal saliency models are under comparison, which include self-resemblance-based model (SR) [15], surprise-based model (Surprise) [3], and phase-based model (MRS) [8]. The source code of all three models are available at their public websites. The saliency distributions of all models are shown in Fig. 5, where we can see that the difference between the saliency distributions at saccadic and random locations computed from the proposed model is much larger than those from the other models. This suggests that the proposed method can better discriminate saccadic from random locations. This is confirmed by the ROC curves given in Fig. 6, where the ROC curve of the proposed model appears to be much higher, especially when the false positive rate is low. Furthermore, the KL distance and AUC values provided in Table 1 quantify the significant improvement of the proposed algorithm over state-of-the-art.

Figure 7 provides several visual examples to demonstrate the superior performance of the proposed model. All saliency models give useful predictions of visual fixation, but the SR, Surprise and MRS models fail to clearly distinguish the fixated object from many other objects in the background. By contrast, the proposed model predicts visual fixations with much higher accuracy.

4. CONCLUSION

We propose a novel video saliency model where the major contributions are in the use of a psychological model of human visual speed perception to quantify temporal saliency and the incorporation of an uncertainty-based adaptive weighting approach in the fusion of spatial and temporal saliency maps. These have led to the superior performance of the proposed method against state-of-the-art approaches. The general framework of the proposed method can be extended in many ways. For example, the uncertainty measure can be generalized to account for the ambiguity in motion and relative speed estimations. Top-down mechanisms and semantic cues may also be employed to improve the spatial and temporal saliency or the uncertainty measurement.

5. REFERENCES

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

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

[3] L. Itti, and P. Baldi. Bayesian Surprise Attracts Human Attention. *NIPS*, 2006.

[4] Y. Ma, X. Hua, L. Lu, and H. Zhang. A Generic framework of user attention model and its application in video summarization *IEEE T-MM*, 7(5), 2005.

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

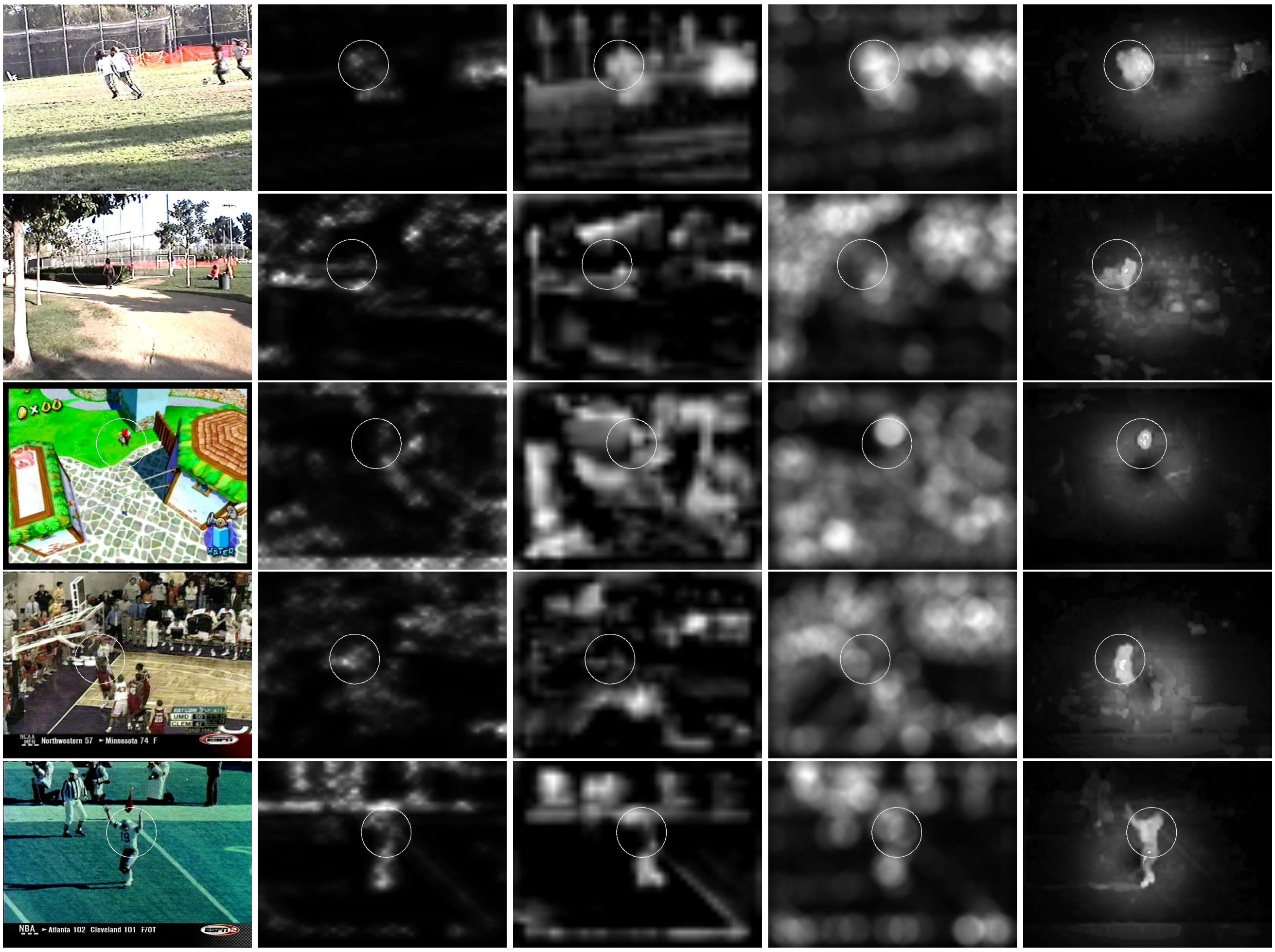

Fig. 7. Visual comparison of saliency models. Column 1: video frames with human fixation point marked with circles; Columns 2 - 5: saliency maps by SR [15], Surprise [3], MRS [8] and the proposed models, respectively.

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

[7] V. Mahadevan and N. Vasconcelos. Spatiotemporal saliency in dynamic scenes. *IEEE T-PAMI*, 32(1), 2010.

[8] C. Guo and L. Zhang. A novel multi-resolution spatiotemporal saliency detection model and its applications in image and video compression. *IEEE T-IP*, 19(1), 185-198, 2010.

[9] A. A. Stocker and E. P. Simoncelli. Noise characteristics and prior expectations in human visual speed perception. *Nature Neuroscience*, 9, 578 - 585, 2006.

[10] Y. Fang, Z. Chen, W. Lin, C.-W. Lin. Saliency detection in the compressed domain for adaptive image retargeting. *IEEE T-IP*, 21(9), 3888-3901, 2012.

[11] D. Sun, S. Roth, and M. J. Black. Secrets of optical flow estimation and their principles. *IEEE CVPR*, 2010.

[12] Z. Wang, and Q. Li. Video quality assessment using a statistical model of human visual speed perception. *Journal of the Optical Society of America A*, 24(12), B61-B69, 2007.

[13] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk. Frequency-tuned salient region detection. *IEEE ICCV*, 2009.

[14] K. Fukunaga. *Introduction to Statistical Pattern Recognition*. Academic Press, San Diego, CA, 1990, Second Edition.

[15] H. J. Seo, and P. Milanfar. Static and Space-time Visual Saliency Detection by Self-Resemblance. *The Journal of Vision*, 9(12):15, 1-27, 2009.