Video Saliency Incorporating Spatiotemporal Cues and Uncertainty Weighting

Yuming Fang, Zhou Wang, *Fellow, IEEE*, Weisi Lin, *Senior Member, IEEE*, and Zhijun Fang, *Senior Member, IEEE*

***Abstract*— We propose a novel algorithm to detect visual saliency from video signals by combining 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. Second, the spatial and temporal saliency maps are merged into one using a spatiotemporally adaptive entropy-based uncertainty weighting approach. The spatial uncertainty weighing incorporates the characteristics of proximity and continuity of spatial saliency, while the temporal uncertainty weighting takes into account the variations of background motion and local contrast. Experimental results show that the proposed spatiotemporal uncertainty weighting algorithm significantly outperforms state-of-the-art video saliency detection models.**

***Index Terms*— Visual attention, video saliency, spatiotemporal saliency detection, uncertainty weighting.**

I. Introduction

VISUAL attention is an important characteristic of the Human Visual System (HVS). It is a cognitive process of selecting the relevant regions while acquiring the most significant visual information from the visual scenes. Research on visual attention dated back to 1890 [1]. In general, the amount of information captured by human eyes is much more than what the central nervous system can process. When human eyes gaze at a natural scene, it is impossible to recognize all the components and their relationship in the scene immediately [2], [3]. Selective attention filters out redundant visual information, such that the visual systems are attracted to the salient regions that contain the most important visual information in the scene [4]. There are two basic visual attention mechanisms: bottom-up and top-down. Bottom-up mechanism is data-driven and task-independent [5], while top-down mechanism is voluntary and dependent on viewing tasks and semantic information [6]–[8].

In the past decades, existing studies have explored selective visual attention mechanisms in the fields of biology, psychology, and computer vision [3]–[20]. Recently, it has also attracted a great deal of attention in the field of multimedia communications because of its potential applications in the evaluation and improvement of quality-of-experience (QoE) in visual communication systems [51]–[53]. According to the Feature Integration Theory (FIT) developed by Treisman *et al.* [9] in the 1980s, the early selective attention mechanism leads some image regions to be salient because of certain features (color, intensity, orientation, motion, etc.) that differentiate them from their surrounding regions [9]. Koch *et al.*'s visual attention model [10] 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 models of saliency detection for images [11]–[20], [45]–[48]. One of the earliest saliency detection models was proposed by Itti *et al* [11]. This model calculates saliency map based on multi-scale center-surround feature contrast from intensity, color and orientation. Based on Itti's model, Harel *et al.* designed a graph-based saliency detection model by using a better measure of dissimilarity [12]. Different from [11], the model adopts graph theory to form saliency maps from low-level features [12]. Gao *et al.* computed the center-surround discriminant for saliency detection [16]. The saliency values of image pixels are obtained by the power of a Gabor-like feature set to discriminate the center-surround visual appearance. In [17], Yan *et al.* described a saliency detection model based on sparse coding. Bruce *et al.* used the principle of information maximization to build a saliency detection model [14], which yields saliency maps for images based on Shannon's self-information measure. Hou *et al.* measured saliency based on a new concept of Spectral Residual [13], where the saliency map is obtained based on log spectra representation of images, which is calculated by the Fourier Transform (FT). Several other studies also adopted the information from frequency domain for saliency detection [18], [47], [48]. Liu *et al.* used a machine learning technique to learn the features of salient objects for images [15]. They first calculate features of multi-scale contrast, center-surround histogram and color spatial distribution from images. Then a Conditional Random Field (CRF) is determined

Manuscript received November 21, 2013; revised April 3, 2014 and June 13, 2014; accepted June 24, 2014. Date of publication July 16, 2014; date of current version July 25, 2014. This work was supported by the Research Foundation from Department of Education of Jiangxi Province (No. GJJ14347), Research Foundation from Jiangxi Provincial Department of Science and Technology (No. 20142BAB217011), and the National Science Foundation of China. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Stefan Winkler.

Y. Fang and Z. Fang are with the School of Information Technology, Jiangxi University of Finance and Economics, Nanchang 330032, China (e-mail: fa0001ng@e.ntu.edu.sg; zjfang@gmail.com).

Z. Wang is with the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON N2L3G1, Canada (e-mail: z.wang@ece.uwaterloo.ca).

W. Lin is with the School of Computer Engineering, Nanyang Technological University, Singapore 639798 (e-mail: wslin@ntu.edu.sg).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2014.2336549

for these features to detect salient objects in images [15]. Goferman *et al.* proposed a context-aware saliency detection model by introducing more context information in the saliency map [38]. Garcia-Diaz *et al.* designed a saliency detection model based on a hierarchical definition of optical variability [45]. Recently, Riche *et al.* used the multi-scale spatial rarity for saliency detection [46].

Compared with saliency detection in still images, video saliency detection is a much more challenging problem due to the complication in the detection and utilization of temporal and motion information. So far, only a limited number of algorithms are proposed for spatiotemporal saliency detection for video signals [18], [19], [23]–[28]. Early methods of salient motion detection attempted to identify moving foreground objects as salient regions. In [23], Wildes proposed a measurement of salient motion for surveillance applications. The spatiotemporal gradient filters are used to evaluate the extent to which local regions in space-time are dominated by a single coherent salient motion [23]. Wixson designed a salient motion detection model by defining salient motion as motion resulting from typical surveillance targets as opposed to other distracting motions [24]. In that study, the motion of image pixels over time is calculated to estimate the moving distance of these image pixels. The pixel saliency is measured by the distance over which the image pixel has traveled with a consistent direction [24].

More recent methods of video saliency detection try to combine spatial and temporal information [18], [19], [25]–[28]. Itti *et al.* utilized a Bayesian model to detect surprising events as important information attracting human attention, where surprise is measured by the difference between posterior and prior beliefs for the observer [19]. Ma *et al.* integrated top-down mechanism into classical bottom-up saliency detection models for video summarization [28], where the top-down information includes semantic cues such as face and speech. Zhai *et al.* linearly combined spatial and temporal saliency maps to obtain the final saliency map for video frames [25]. In that study, the spatial saliency map is computed based on the color histogram of video frames, while the temporal saliency map is calculated by the planar motion between images (estimated by applying RANSAC on point correspondences in the scene) [25]. Le Meur *et al.* extended their saliency detection model for images [29] by adding temporal saliency information into the framework [26]. The spatiotemporal saliency map for video frames is calculated based on the feature maps from achromatic, chromatic and temporal information [26]. Mahadevan *et al.* extended the discriminant saliency detection approach in [16] by incorporating motion-based perceptual grouping and a discriminant formulation of center-surround saliency to create spatiotemporal saliency maps [27]. Guo *et al.* combined quaternion intensity, color and motion features and employed the phase spectrum of Quaternion Fourier Transform to calculate spatiotemporal saliency for video frames [18]. The spatiotemporal saliency map is computed as the Inverse Fourier Transform (IFT) on a constant amplitude and the original phase spectrum. Seo *et al.* introduced the notion of self-resemblance to measure visual saliency from video signals [32]. Butko *et al.* proposed a real-time spatiotemporal saliency detection model for robot cameras [49].

A key issue in video saliency evaluation is how to incorporate 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 and Simoncelli regarding human visual speed perception [33], 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 motion information content in a perceptually meaningful way and use it as a predictor of temporal saliency.

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 that often use simple combination rules such as linear combination with fixed weights, we associate each saliency map with a entropy-based uncertainty map and merge the saliency maps adaptively based on the local uncertainty measures. Our uncertainty calculation is motivated by the principles of proximity and continuity of the Gestalt theory [39], [40], and also the impacts of background motion and local contrast based on the psychovisual study in [33]. The law of proximity states that elements which are close to each other tend to be perceived as forming a group, while the law of continuity indicates that elements which are connected with each other tend to be perceived into a group together. These two principles can be applied to the visual saliency as follows: 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 is more likely to be a salient location. We use these two principles to calculate the uncertainty for the spatial saliency. The psychovisual experiments in the study [33] suggest that the accuracy of motion perception varies with the speed of motion and local contrast. We adopt the mathematical models behind to determine the uncertainty of temporal saliency. The final spatiotemporal saliency is calculated by fusing spatial and temporal saliency based on their uncertainties. Partial preliminary results of the proposed spatiotemporal saliency detection model have been published in [60]. Experimental results show that the proposed spatiotemporal saliency detection model outperforms other existing ones on a public large-scale database.

II. 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

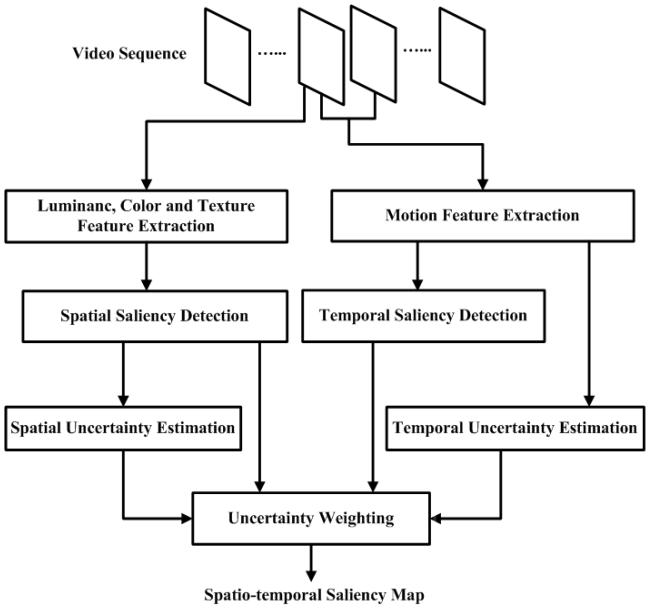

Fig. 1. The framework of the proposed model.

approach, resulting in the final spatiotemporal saliency map. The principle behind the framework is flexible. For example, other features, including high-level cognitive features, can be easily integrated into the framework.

A. Spatial Saliency Evaluation

Our previous research work have demonstrated that DCT coefficients can represent statistical features of image patches effectively for saliency detection in still images [20]. The spatial saliency detection used in this study basically follows the method introduced in [20] (with modification) and is briefly described here. Given a video frame, we first convert all image pixels into the YCbCr color space rather than RGB 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.

In this study, 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:

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

where $k \epsilon \{L, C_1, C_2, T\}$, σ_s 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_s determines the size of the neighborhood and thus the locality of the feature contrast measure. Different from the study [20], here the texture difference between image patches is calculated by the Euclidean distance rather than Hausdorff distance for the low computational complexity.

Finally, the feature maps are normalized to $[0, 1]$ and the overall spatial saliency map $S^{(s)}$ of the video frame is calculated as the average of the feature maps [20]:

$$S^{(s)} = \frac{1}{K} \sum_{k \in \{L, C_1, C_2, T\}} N(S^k) \tag{2}$$

where N is the normalization operator and K is the number of features ($K = 4$). For simplicity, we have dropped the spatial index i from both sides of the equation.

B. Temporal Saliency Evaluation

Existing studies have demonstrated that object motion is often highly correlated with visual attention [21], [22]. Our temporal saliency evaluation algorithm starts with optical flow based motion estimation [30], 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 [55], [56]. Generally, an object of strong motion with respect to the background would be a surprising event to the HVS [21], [22], [33]. If we consider the HVS as an efficient information extractor, it would pay more attention to such a event [21], [22]. Therefore, visual attention of motion can be measured 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 [33]. The results have been employed in the field of perceptual video quality assessment [34], but have not been exploited in the context of visual saliency estimation. According to the results in [33], the "perceptual" prior distribution of motion speed can be well fitted with a power-law function:

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

where v is the motion speed; and 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 the motion speed-based temporal saliency using its self-information 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 [34].

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 calculation in Eq. (1), here we evaluate the relative speed v_i of the i-th patch as

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

where D_{ij}^v is the length of the vector difference between the mean absolute motion vectors of patches i and j. Eq. (5) calculates the relative motion in a more localized fashion within a large neighboring region rather than comparing with the global background motion. By imposing a weighting factor

based on the distance between the current patch and the neighboring patches, we obtain the flexibility to put more emphasis on the regions closer to the current patch. We find this method simple and robust, and it provides more useful features and flexibilities than using global background motion.

The temporal saliency $S^{(t)}$ is evaluated at all patches in a video frame, and is then normalized to the range of [0,1], resulting in a temporal saliency map of the frame.

C. Spatial 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 spatial uncertainty evaluation method is conceptually rooted in the Gestalt theory [39], [40], in which the law of proximity indicates that elements which are close to each other tend to be perceived as forming a group, while the law of continuity states that elements which are connected with each other tend to be perceived into a group together. These two principles of Gestalt Law may be extended to visual saliency as follows: 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 is more likely to be a salient location. These properties are also in line with our empirical statistics of an image database created by Achanta *et al.* [35], which includes 1000 images and their corresponding binary ground truth maps from human subjects. Note that our goal is to create a scalar-valued prediction of the degree of visual saliency at each spatiotemporal location, but our spatial uncertainty model is built upon a binary mask of segmented main objects [35] that attract visual attention. We make this seemingly suboptimal choice for the following reasons.

Firstly, the desirable ground truth training data would be saliency maps with gradual saliency values obtained from visual saliency recording. Unfortunately, it is infeasible to obtain such desired ground truth maps, because eye tracking experiments can only obtain limited numbers of fixation points. For still images, the recorded fixation points (even when all recorded points from multiple observes are collapsed) are sparsely distributed over space. For video, the situation is even worse, where we often end up with only 0 or 1 fixation point in each video frame. Therefore, dense ground truth saliency map that directly reflects the probability of a pixel being salient is unrealistic to obtain in reality. In some existing image saliency databases, smooth saliency maps are provided, but these smooth maps are artificially created by Gaussian filtering the sparse fixation maps. It would be mistaken to derive our saliency likelihood models from such artificially smoothed maps. For example, the connectedness model would be sensitive to and largely determined by the size of the Gaussian filter, which is often picked arbitrarily. In other words, the artificial smoothing process (rather than the true degree of visual saliency) would dominate the modelling process.

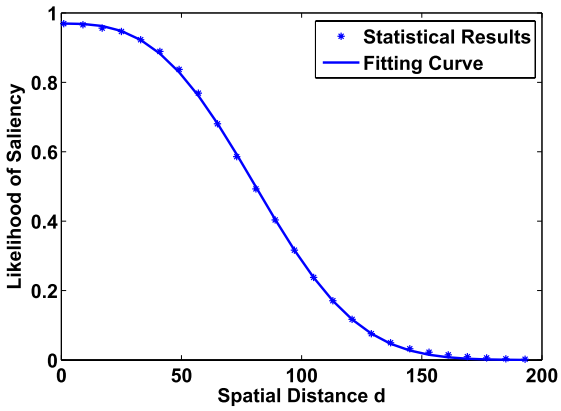

Fig. 2. Likelihood of saliency as a function of spatial distance from saliency center.

Secondly, the Gestalt theory we employ to guide our modelling of spatial uncertainty only suggests us to consider the relationship between the locations of salient pixels, without taking into account the degree of saliency. As a result, to apply the Gestalt theory, we have to rely on a binary map that classifies each pixel to be either salient or not, as is the case of the database given in [35]. This gives us a clean and easy method to build the uncertainty model, but meanwhile only provides a rough estimate of the uncertainty in estimating human visual saliency.

Finally, despite of the limitations, our approach finds a shortcut capturing the most important ingredients of the intuitive ideas that could help us constrain the likelihood of saliency. To the best of our knowledge, there was no previous work that uses a similar approach to create uncertainty maps to help with saliency detection. Moreover, our experiments in III demonstrate significant effect of such an approach in improving saliency prediction.

For images in the database [35], the salient pixels in the ground truth maps are the pixels with saliency value one, while the saliency values of the other pixels are zero. Specifically, given an image and its ground truth saliency map S, we first compute the expected center location of its saliency map by:

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

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

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 a pixel being a salient pixel as a function of d. Specifically, for a given distance value d, the likelihood is estimated by counting the percentage of saliency pixels in all pixels of all the binary maps that have distance d from

their corresponding saliency centers. The statistical results are shown in Fig. 2. As expected, with the increase of d from the saliency center, the likelihood decreases, conforming with the law of proximity. 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 [35]. 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 (the property of continuity). For each pixel, we calculate its connectedness as

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

where R_N represents the set of direct neighboring pixels near the current pixel, excluding itself. Based on the image database [35], 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, based on the image database [35]. 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, assuming independence between proximity and connectedness, we calculate the total uncertainty for each pixel in the spatial saliency as

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

Applying such uncertainty computation to the spatial saliency map generated in Section II-A, we obtain the spatial uncertainty map of each video frame.

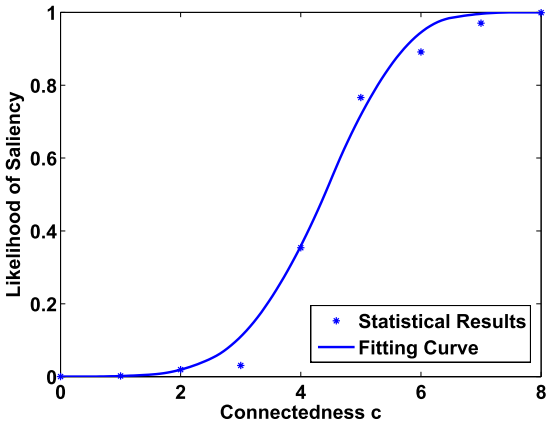

Fig. 3. Likelihood of saliency as a function of connectedness.

D. Temporal Uncertainty Evaluation

When the background motion in a video sequence is very large (most likely caused by camera motion), the visual system cannot identify the motion of objects in the scene as accurately as those in static background [33]. In addition, the accuracy may decrease with the increase of local contrast [33], [57]–[59]. Interestingly, these intuitive ideas about perceptual uncertainty find direct justifications from the psychovisual studies in [33], where the likelihood function of perceived speeds given the stimulus speed v_s is modeled as

$$p(v_m|v_s) = \frac{1}{\sqrt{2\pi}\sigma v_m} \exp\left[\frac{-(\log v_m - \log v_s)^2}{2\sigma^2}\right] \tag{14}$$

where v_s and v_m are the speed of the true stimulus motion and its perceptual measurement, respectively; the width parameter σ in the log-normal distribution determines the level of perceptual uncertainty. Furthermore, the experimental results in [33] show that in the logarithmic speed domain, σ is roughly constant for any stimulus speed v_s and inversely dependent on the stimulus contrast q. This can be expressed as

$$\sigma = \lambda / q^{\gamma} \tag{15}$$

where λ and γ are both positive constants.

Since the uncertainty of speed is associated with the background speed [33], [34], assuming the stimulus speed v_s is the speed of the background motion v_g, we can quantify the perceptual uncertainty using the entropy of the likelihood:

$$\begin{aligned} U^{(t)} &= -\int_{-\infty}^{\infty} p(v_m|v_g)\log p(v_m|v_g)dv_m \\ &= \log v_g - \gamma \log q + \delta \end{aligned} \tag{16}$$

where $\delta = \frac{1}{2} + \frac{1}{2}\log(2\pi) + \log\lambda$ is a constant. The parameters are set according to the experiments from the study [34]. As expected, this uncertainty measure for temporal saliency $U^{(t)}$ increases with background motion and decreases with the stimulus contrast.

E. Spatio Temporal Uncertainty Weighting (STUW)

The last step in creating an overall spatiotemporal saliency map is to combine the spatial and temporal saliency maps

Fig. 4. Sample 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.

computed in Sections II-A and II-B, respectively, which are also associated with different levels of uncertainty based on the computation in Sections II-C and II-D. Naturally, the saliency measure with lower uncertainty should be given larger weight. This leads to an Uncertainty Weighted (UW) fusion rule given by

$$S = \frac{U^{(t)} S^{(s)} + U^{(s)} S^{(t)}}{U^{(s)} + U^{(t)}} \tag{17}$$

which is also the final step of our SpatioTemporal Uncertainty Weighting (STUW) algorithm.

Since both spatial and temporal uncertainty maps change over space and time, this fusion rule is spatiotemporally adaptive, which differentiates it from existing methods where fixed weighting schemes are used to combine spatial and temporal saliency maps [19], [25], [26], [28]. Fig. 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 locations of visual fixations.

III. Experimental Evaluation

We use two experiments based on a publicly available video database [19] to evaluate the performance of the proposed STUW algorithm. The first experiment shows the effect of the fusion method by uncertainty weighting, and the second experiment compares the performance of the proposed STUW algorithm against existing ones.

A. Evaluation Methodology

The test database [19], [50] contains 50 video clips totaling over 25 minutes with a variety of video content such as sports video, video games, outdoor video in daytime and nighttime, television broadcast, etc. The ground truth of visual fixations is obtained from the fixation points from 8 subjects recorded by an eye tracker. Some samples of video frames and their corresponding ground truth are shown in Figs. 7 and 10.

We use similar assessment methods as the study in [19] to evaluate the performance of the proposed STUW method. The data is collected from all video frames with fixation points in all sequences. The performance of spatiotemporal saliency detection models is evaluated by comparing the response values at fixation and random locations in the saliency map [19]. Generally, an effective saliency detection model would have high response at fixation locations and no response at most random locations. Here, the saliency distributions at fixation and random locations are calculated with 10 bins of saliency values over the saliency map, as shown in Figs. 5 and 8, in which the x-axis represents the saliency value bins from 0 to 1 with 0.1 interval, while the y-axis represents the number of human fixation locations or random locations with different saliency value bins. The saliency distributions are obtained as the histogram of saliency values at fixation locations or randomly chosen locations from all video frames with fixation locations. Kullback-Leibler (KL) distance is used to measure the similarity between these two distributions:

$$KL(H, R) = \frac{1}{2}\left(\sum_n h_n \log \frac{h_n}{r_n} + \sum_n r_n \log \frac{r_n}{h_n}\right) \tag{18}$$

where H and R are saliency distributions at human fixation locations and random locations with probability density functions h_n and r_n, respectively; and n is the index of the saliency value bin ($n \in \{1, 2, 3 \ldots, 10\}$). The saliency detection model with larger KL distance can better discriminate human fixation locations from random locations, and thus has better performance [19].

Besides KL distance, Receiver Operating Characteristics (ROC) curve [36] is also adopted for performance evaluation. The ROC curve is a graphical plot of *the True Positive Rate (TPR) VS. the False Positive Rate (FPR)* for a binary classifier with varying discrimination thresholds, as shown in Figs. 6 and 9. The saliency distributions at human fixations and random locations are used as the test and discrimination set, respectively. For each threshold, the *TPR* is calculated as the percentage of the number of human fixation locations with salient values larger than this threshold over the total number of human fixation locations; the *FPR* is computed as the percentage of the number of random locations with salient values larger than this threshold over the total number of random locations. The area under the ROC curve (AUC) provides an overall evaluation. A better video saliency detection model is expected to have a larger AUC value.

Additionally, NSS (Normalized Scanpath Saliency) [42] is also adopted for performance evaluation in this study. As indicated in [41], ROC depends on the ordering of the fixations and does not capture the metric amplitude differences. Thus, we use the NSS metric to conduct a more comprehensive performance evaluation of the proposed model. The NSS is defined as the response value at human fixation locations in

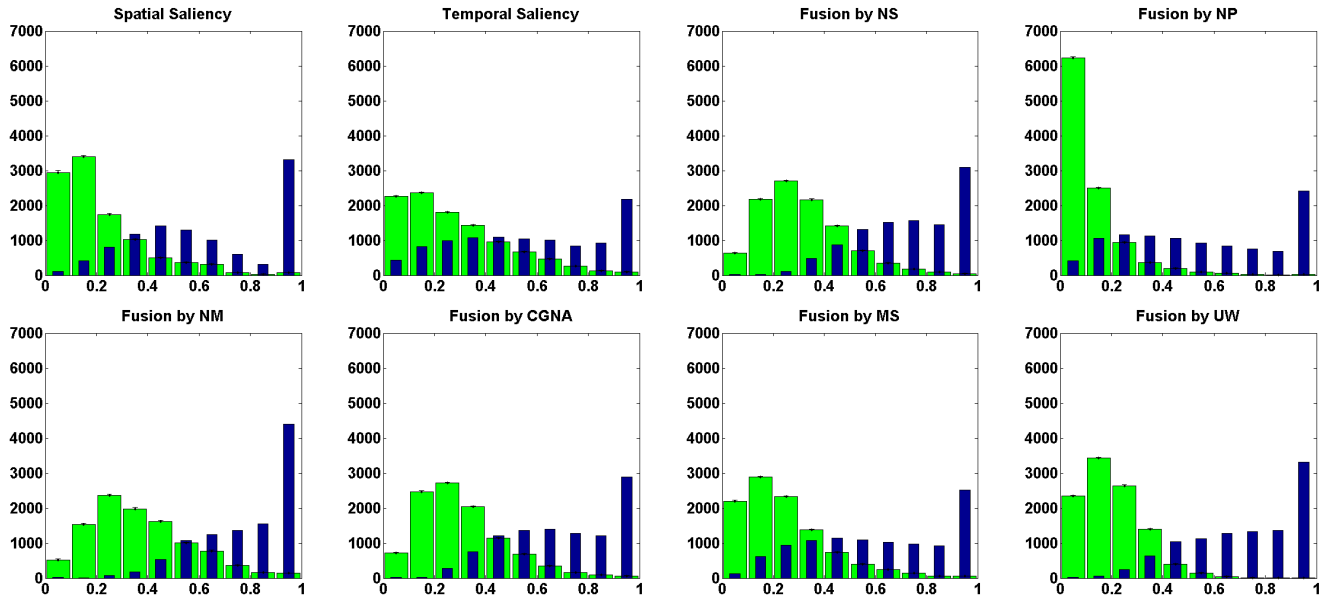

Fig. 5. The saliency distributions at human fixation locations (narrow blue bars) and random locations (wide green bars) from spatial saliency, temporal saliency, and spatiotemporal saliency by different fusion methods.

the normalized saliency map with zero mean and unit standard deviation. A larger NSS value implies better predication performance of the saliency detection model.

B. Experiment 1

In this experiment, we compare the performance of the spatial, temporal, and spatiotemporal saliency maps to demonstrate the significance of the uncertainty weighting based fusion algorithm. Furthermore, we compare the performance of the proposed UW fusion method with those from other existing fusion methods.

Commonly used fusion methods to combine spatial and temporal saliency maps include [37]: (1) *Normalized and Sum (NS)*: a simple fusion method that normalizes the spatial and temporal saliency maps to the same dynamic range and then sums them to obtain the final spatiotemporal saliency map; (2) *Normalized and Maximum (NM)*: the fusion method that normalizes the spatial and temporal saliency maps to the same dynamic range and then uses the maximum value as the final saliency value at each location; (3) *Normalized and Product (NP)*: the fusion method that normalizes the spatial and temporal saliency maps to the same dynamic range and then multiplies the maps to produce the spatiotemporal saliency map; (4) *Contents-based Global Nonlinear Amplification (CGNA)* [43]: the fusion method that consists of globally promoting feature maps with a small number of strong peaks of activity, while globally suppressing feature maps eliciting peak responses in the visual scene; (5) *Maximum and Skewness (MS)* [44]: the fusion method that adopts the maximum value of spatial saliency and skewness of temporal saliency to combine the spatial and temporal saliency. The detailed description of NS, NM and NP can be found in [37], while the detailed introduction of CGNA and MS can be found in [43] and [44], respectively.

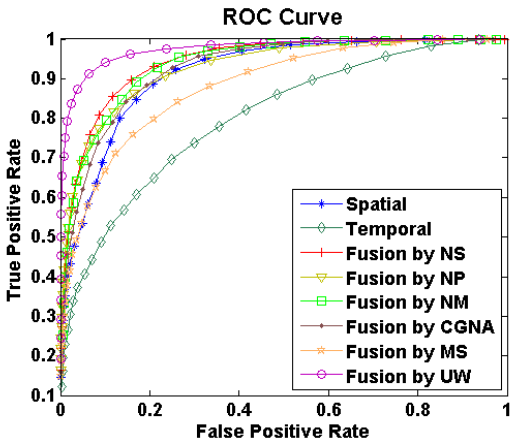

Fig. 6. ROC comparison of spatial saliency, temporal saliency, and spatiotemporal saliency by different fusion methods.

We compare the proposed UW fusion method with aforementioned existing fusion methods (*NS*, *NM*, *NP*, *CGNA*, and *MS*). Fig. 5 shows the saliency distributions at human fixations and random locations for the spatial saliency map, temporal saliency map, and spatiotemporal saliency maps created by different fusion methods. It can be seen that the difference between saliency distributions at human fixation and random locations from spatial saliency is larger than that from the temporal saliency. The difference between saliency distributions of human fixation and random locations from the proposed fusion method is larger, compared with the other fusion methods (*NS*, *NP*, *NM*, *CGNA* and *MS*). These conclusions are further confirmed by Fig. 6, which shows the ROC curves of spatial saliency, temporal saliency, and spatiotemporal saliency from different fusion methods.

Table I provides the detailed KL distance, AUC and NSS values of spatial saliency, temporal saliency, and

TABLE I
KL DISTANCE, AUC, AND NSS COMPARISONS OF SPATIAL SALIENCY, TEMPORAL SALIENCY, AND SPATIOTEMPORAL SALIENCY BY DIFFERENT FUSION METHODS

Models	Spatial	Temporal	NS	NP	NM	CGNA	MS	UW
KL Dist.	1.575	0.694	2.169	1.918	2.037	1.82	1.242	3.327
AUC	0.912	0.799	0.939	0.921	0.933	0.928	0.882	0.975
NSS	1.8109	1.3155	1.6919	1.9966	1.6436	1.7147	1.7322	2.3852

Fig. 7. Visual comparison of spatial, temporal, and spatiotemporal saliency. Row 1: video frames with human fixation point marked with circles; Rows 2 - 4: spatial, temporal, and spatiotemporal saliency maps, respectively.

spatiotemporal saliency by different fusion methods. It appears that spatial saliency alone can predict more accurate fixations than temporal saliency. The KL distance and AUC values from some spatiotemporal saliency maps (*NS*, *NP*, *NM*, *CGNA* and the proposed one) are higher than those of the spatial or temporal saliency map only, while the KL distance and AUC values from spatiotemporal saliency maps by *MS* are lower than those from spatial saliency only. From the NSS results, we can see that the spatiotemporal saliency maps from *NP* and the proposed *UW* can obtain better performance than the spatial saliency. Among the results from different fusion methods, the proposed UW fusion method obtain the highest KL distance, AUC and NSS values. Thus, the proposed fusion method by uncertainty weighting achieves better performance than other existing fusion methods.

Fig. 7 provides sample saliency maps produced by the proposed method. From the first (or second) column of this figure, the saliency value at human fixation location in the spatial saliency map is not highest in this saliency map. Although the saliency value at human fixation location in the temporal saliency map might be the highest, there are other locations with high saliency values in the temporal saliency map. On the contrary, the saliency value at human fixation location is always highest in the spatiotemporal saliency map. Other comparison samples also demonstrate this advantage of the spatiotemporal saliency by uncertainty weighting. These comparison samples demonstrate that the spatiotemporal saliency map predicts human fixations more accurately compared with the spatial or temporal saliency maps alone.

As suggested by Table I, the performance of the spatial saliency is closer to that of the spatiotemporal saliency compared with that of the temporal saliency. This is not surprising because many regions of interests in a video sequence are also of interests in individual frames. On the other hand, salient motion from the temporal saliency map can enhance certain salient regions in the spatial saliency map and helps in determining the final spatiotemporal saliency map. Therefore, both

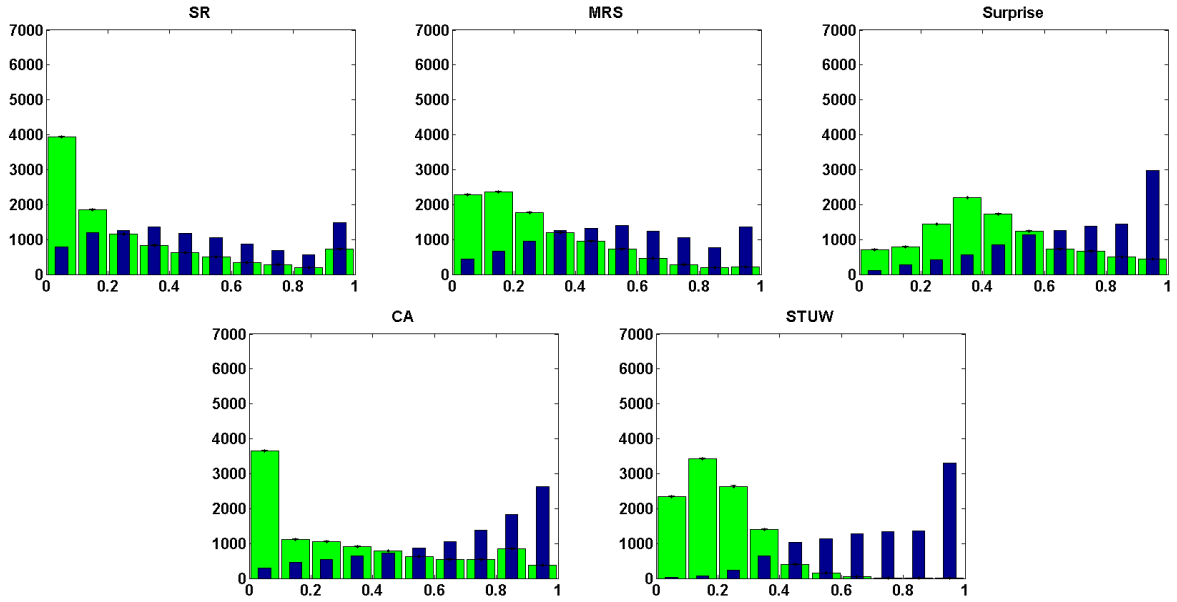

Fig. 8. Saliency distributions at human fixation locations (narrow blue bars) and random locations (wide green bars) from different spatiotemporal saliency models. The x- and y-axis represent the predicted saliency values from different models and histograms of the corresponding salient values, respectively.

spatial and temporal saliency contribute to the overall spatiotemporal saliency, as quantitatively indicated in Table I.

C. Experiment 2

In this experiment, we compare the proposed STUW algorithm with existing ones. In addition to STUW, three state-of-the-art spatiotemporal saliency models are under comparison, which include self-resemblance-based model (SR) [32], surprise-based model (Surprise) [19], and phase-based model (MRS) [18]. In addition, we also include a recent context-aware saliency detection model for still images (CA) [38]. The source code of all four models is available at their public websites. The experimental results from different existing models are shown in Figs. 8–10 and Table II.

From Fig. 8, it can be seen that the difference between the saliency distributions at fixations and random locations computed from the proposed STUW algorithm is much larger than those from the other models. This suggests that the proposed STUW algorithm can better discriminate human fixations from random locations. This is confirmed by the ROC curves given in Fig. 9, where the ROC curve of the proposed STUW algorithm appears to be much higher, especially when the *FPR* is low. Furthermore, the KL distance, AUC, and NSS values provided in Table II quantify the significant improvement of the proposed STUW algorithm over state-of-the-art.

The results in Table II also shows that the recent image saliency detection model CA [38] performs similarly or even better than the other spatiotemporal saliency detection models (SR, Surprise and MRS models). This is not overly surprising, given the good performance of the spatial-only saliency detection method reported in Section III-B.

As reported in [54], most human fixation data recorded by head-mounted eye tracking systems have strong center bias

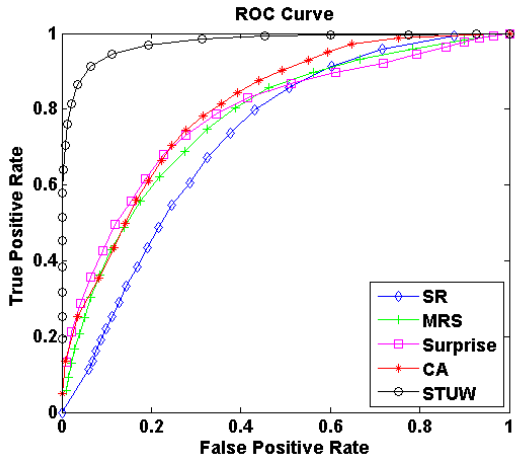

Fig. 9. ROC comparison of spatiotemporal saliency models.

TABLE II
KL DISTANCE, AUC AND NSS COMPARISONS OF SPATIOTEMPORAL SALIENCY MODELS

Models	SR [32]	MRS [18]	Surprise [19]	CA [38]	STUW
KL Dist.	0.391	0.529	0.593	0.76	3.327
AUC	0.722	0.771	0.782	0.802	0.975
NSS	1.6642	1.4246	1.5537	1.4672	2.3852

due to the following two factors: the setup of the experiments (such as subjects being centered at the center of the display screen), and the fact that human photographers tend to center objects of interest. To eliminate the center bias during performance evaluation, we further conduct comparisons by using similar methods as in [32] and [54]. Specifically, we first calculate the saliency distribution at the human eye fixation

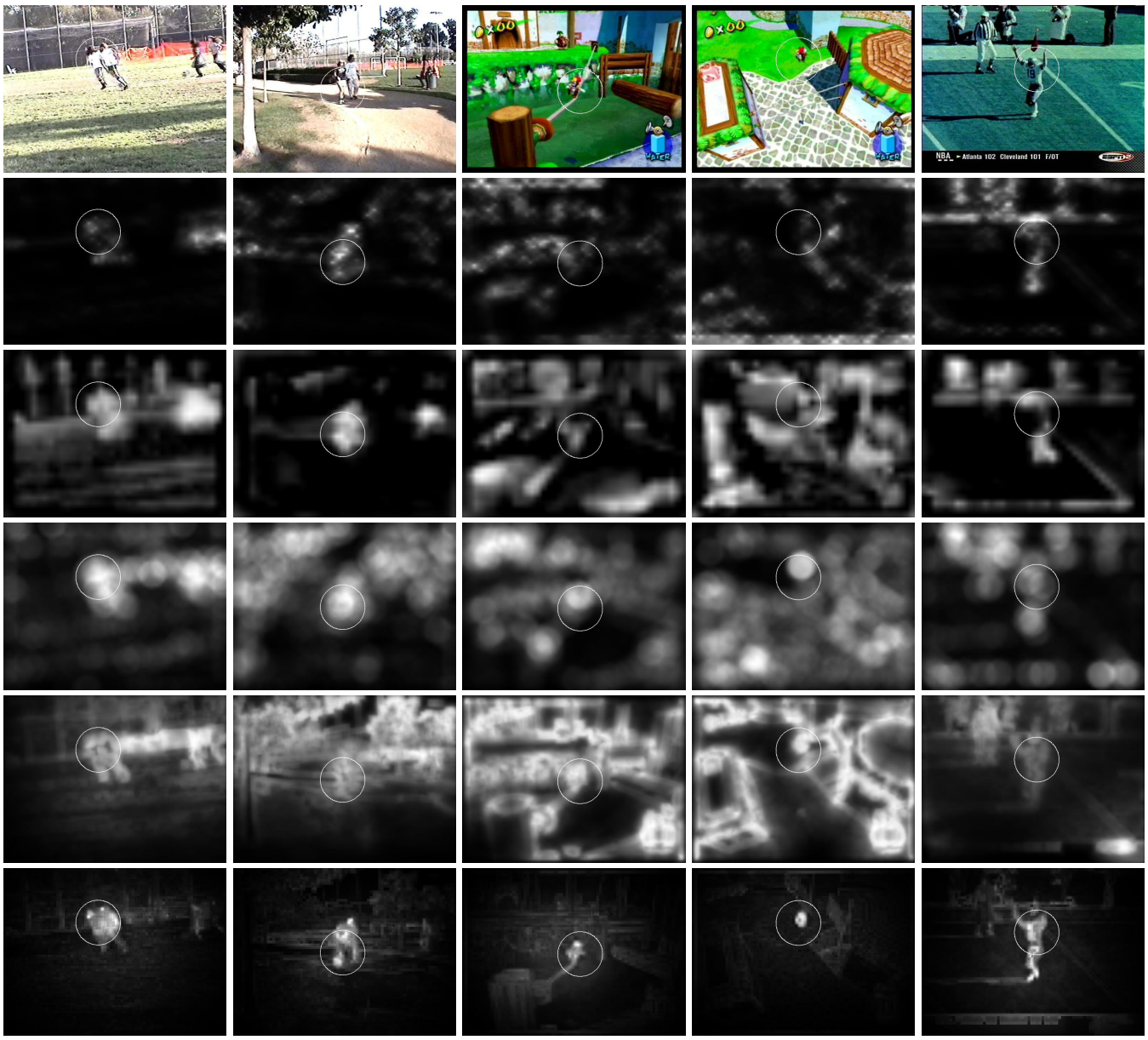

Fig. 10. Visual comparison of saliency models. Row 1: video frames with human fixation point marked with circles; Rows 2 - 6: saliency maps by SR [32], Surprise [19], MRS [18], CA [38] and STUW, respectively.

TABLE III
KL DISTANCE AND AUC COMPARISONS OF SPATIOTEMPORAL SALIENCY MODELS WITH SHUFFLED VIDEO FRAMES

Models	SR [32]	MRS [18]	Surprise [19]	CA [38]	STUW
KL Dist.	0.047	0.068	0.07	0.06	0.788
AUC	0.575	0.600	0.595	0.591	0.6056

locations at the video frames. The saliency distribution from randomly chosen locations is computed by calculating the saliency values at the same human fixation locations but of randomly chosen video frames from the test set. The experimental results from different models are shown in Table III, where we can see that the overall performance of the proposed model is again significantly better than existing models of saliency prediction.

Fig. 10 provides several visual examples to demonstrate the superior performance of the STUW algorithm. All saliency models give useful predictions of visual fixations, but the SR, Surprise and MRS models fail to clearly distinguish the fixated object from many other objects in the background. There are other locations with equal or higher saliency values besides the human fixation locations in the saliency maps given by SR, Surprise, MRS and CA models. By contrast, the saliency values within human fixation locations are always the highest in the corresponding spatiotemporal saliency maps created by the STUW algorithm, as shown in the last row of Fig. 10.

The computational complexity of the current implementation of the proposed method is much higher than the existing

methods, mainly due to the motion estimation process (optical flow algorithm) involved. Further investigation is desired to reduce the complexity.

IV. Conclusion

In this study, we have proposed a novel spatiotemporal saliency detection model that has two major contributions. One is the use of a psychological model of human visual speed perception to quantify temporal saliency; the other is the incorporation of an uncertainty-based adaptive weighting approach in the fusion of spatial and temporal saliency maps. The spatial uncertainty calculation is motivated by Gestalt laws of proximity and continuity, while the temporal uncertainty is determined by motion speed and local contrast features. Experimental results demonstrate that the proposed STUW algorithm achieves superior performance against state-of-the-art approaches. The general framework of the proposed method may be extended in many ways. For example, top-down mechanisms and semantic cues may be employed to improve the spatial and temporal saliency or the uncertainty measurement.

References

[1] W. James, *The Principles of Psychology*. Cambridge, MA, USA: Harvard Univ. Press, 1890.

[2] W. Schneider and R. M. Shiffrin, "Controlled and automatic human information processing: I. Detection, search, and attention," *Psychol. Rev.*, vol. 84, no. 1, pp. 1–66, 1977.

[3] K. Koch *et al.*, "How much the eye tells the brain," *Current Biol.*, vol. 16, no. 14, pp. 1428–1434, 2006.

[4] H. Pashler, *The Psychology of Attention*. Cambridge, MA, USA: MIT Press, 1997.

[5] H. E. Egeth and S. Yantis, "Visual attention: Control, representation, and time course," *Annu. Rev. Psychol.*, 48, pp. 269–297, Feb. 1997.

[6] Z. Lu, W. Lin, X. Yang, E. Ong, and S. Yao, "Modeling visual attention's modulatory aftereffects on visual sensitivity and quality evaluation," *IEEE Trans. Image Process.*, vol. 14, no. 11, pp. 1928–1942, Nov. 2005.

[7] A. Torralba, A. Oliva, M. S. Castelhano, and J. M. Henderson, "Contextual guidance of eye movements and attention in real-world scenes: The role of global features in object search," *Psychol. Rev.*, vol. 113, no. 4, pp. 766–786, 2006.

[8] C. Kanan, M. Tong, L. Zhang, and G. Cottrell, "SUN: Top-down saliency using natural statistics," *Vis. Cognit.*, vol. 17, no. 6, pp. 979–1003, 2009.

[9] A. Treisman and G. Gelade, "A feature-integration theory of attention," *Cognit. Psychol.*, vol. 12, no. 1, pp. 97–136, 1980.

[10] C. Koch and S. Ullman, "Shifts in selective visual attention: Towards the underlying neural circuitry," *Human Neurobiol.*, vol. 4, no. 4, pp. 219–227, 1985.

[11] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 11, pp. 1254–1259, Nov. 1998.

[12] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," in *Advances in Neural Information Processing Systems*. Cambridge, MA, USA: MIT Press, 2006.

[13] X. Hou and L. Zhang, "Saliency detection: A spectral residual approach," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2007, pp. 1–8.

[14] N. D. Bruce and J. K. Tsotsos, "Saliency based on information maximization," in *Advances in Neural Information Processing Systems*, vol. 18. Cambridge, MA, USA: MIT Press, 2006, pp. 155–162.

[15] T. Liu, J. Sun, N. Zheng, X. Tang, and H. Y. Shum, "Learning to detect a salient object," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2007, pp. 1–8.

[16] D. Gao and N. Vasconcelos, "Bottom-up saliency is a discriminant process," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2007, pp. 1–6.

[17] J. Yan, M. Zhu, H. Liu, and Y. Liu, "Visual saliency detection via sparsity pursuit," *IEEE Signal Process. Lett.*, vol. 17, no. 8, pp. 739–742, Aug. 2010.

[18] C. Guo and L. Zhang, "A novel multi-resolution spatiotemporal saliency detection model and its applications in image and video compression," *IEEE Trans. Image Process.*, vol. 19, no. 1, pp. 185–198, Jan. 2010.

[19] L. Itti and P. Baldi, "Bayesian surprise attracts human attention," in *Advances in Neural Information Processing Systems*. Cambridge, MA, USA: MIT Press, 2006.

[20] Y. Fang, Z. Chen, W. Lin, and C.-W. Lin, "Saliency detection in the compressed domain for adaptive image retargeting," *IEEE Trans. Image Process.*, vol. 21, no. 9, pp. 3888–3901, Sep. 2012.

[21] R. A. Abrams and S. E. Christ, "Motion onset captures attention," *Psychol. Sci.*, vol. 14, no. 5, pp. 427–432, 2003.

[22] S. Yantis and J. Jonides, "Abrupt visual onsets and selective attention: Evidence from visual search," *J. Experim. Psychol., Human Perception Perform.*, vol. 10, no. 5, pp. 601–621, 1984.

[23] R. Wildes, "A measure of motion salience for surveillance applications," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 1998, pp. 183–187.

[24] L. E. Wixson, "Detecting salient motion by accumulating directionally-consistent flow," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 774–780, Aug. 2000.

[25] Y. Zhai and M. Shah, "Visual attention detection in video sequences using spatiotemporal cues," in *Proc. 14th Annu. ACM Int. Conf. Multimedia*, 2006, pp. 815–824.

[26] O. Le Meur, P. Le Callet, and D. Barba, "Predicting visual fixations on video based on low-level visual features," *Vis. Res.*, vol. 47, no. 19, pp. 2483–2498, 2007.

[27] V. Mahadevan and N. Vasconcelos, "Spatiotemporal saliency in dynamic scenes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 1, pp. 171–177, Jan. 2010.

[28] Y. Ma, X. Hua, L. Lu, and H. Zhang, "A generic framework of user attention model and its application in video summarization," *IEEE Trans. Multimedia*, vol. 7, no. 5, pp. 907–919, Oct. 2005.

[29] O. Le Meur, P. Le Callet, D. Barba, and D. Thoreau, "A coherent computational approach to model bottom-up visual attention," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 5, pp. 802–817, May 2006.

[30] D. Sun, S. Roth, and M. J. Black, "Secrets of optical flow estimation and their principles," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2010, pp. 2432–2439.

[31] K. Zhang and J. Kittler, "Global motion estimation and robust regression for video coding," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 1998, pp. 2589–2592.

[32] H. J. Seo and P. Milanfar, "Static and space-time visual saliency detection by self-resemblance," *J. Vis.*, vol. 9, no. 12, p. 15, 2009.

[33] A. A. Stocker and E. P. Simoncelli, "Noise characteristics and prior expectations in human visual speed perception," *Nature Neurosci.*, vol. 9, no. 4, pp. 578–585, 2006.

[34] Z. Wang and Q. Li, "Video quality assessment using a statistical model of human visual speed perception," *J. Opt. Soc. Amer. A*, vol. 24, no. 12, pp. B61–B69, 2007.

[35] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, "Frequency-tuned salient region detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2009, pp. 1597–1604.

[36] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, 2nd ed. San Diego, CA, USA: Academic, 1990.

[37] C. Chamaret, J. C. Chevet, and O. Le Meur, "Spatio-temporal combination of saliency maps and eye-tracking assessment of different strategies," in *Proc. 17th IEEE Int. Conf. Image Process.*, Sep. 2010, pp. 1077–1080.

[38] S. Goferman, L. Zelnik-Manor, and A. Tal, "Context-aware saliency detection," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2010, pp. 2376–2383.

[39] J. C. Banerjee, "Gestalt theory of perception," in *Encyclopaedic Dictionary of Psychological Terms*. New Delhi, India: M.D. Publications Pvt. Ltd., 1994, pp. 107–109.

[40] H. Stevenson, *Emergence: The Gestalt Approach to Change*. Novelty, OH, USA: Cleveland Consulting Group, Inc., Apr. 2012.

[41] Q. Zhao and C. Koch, "Learning a saliency map using fixated locations in natural scenes," *J. Vis.*, vol. 11, no. 3, pp. 1–15, 2011.

[42] R. Peters, A. Iyer, L. Itti, and C. Koch, "Components of bottom-up gaze allocation in natural images," *Vis. Res.*, vol. 45, no. 18, pp. 2397–2416, 2005.

[43] L. Itti and C. Koch, "Comparison of feature combination strategies for saliency-based visual attention systems," *Proc. SPIE*, vol. 3644, pp. 473–482, Jan. 1999.

[44] A. Rahman, D. Houzet, D. Pellerin, S. Marat, and N. Guyader, "Parallel implementation of a spatio-temporal visual saliency model," *J. Real-Time Process.*, vol. 6, no. 1, pp. 3–14, 2011.

[45] A. Garcia-Diaz, V. Leboran, X. R. Fdez-Vidal, and X. M. Pardo, "On the relationship between optical variability, visual saliency, and eye fixations: A computational approach," *J. Vis.*, vol. 12, no. 6, p. 17, 2012.
[46] N. Riche, M. Mancas, M. Duvinage, M. Mibulumukini, B. Gosselin, and T. Dutoit, "RARE2012: A multi-scale rarity-based saliency detection with its comparative statistical analysis," *Signal Process., Image Commun.*, vol. 28, no. 6, pp. 642–658, 2013.
[47] B. Schauerte and R. Stiefelhagen, "Quaternion-based spectral saliency detection for eye fixation prediction," in *Proc. 12th Eur. Conf. Comput. Vis. (ECCV)*, Florence, Italy, 2012, pp. 116–129.
[48] J. Li, M. D. Levine, X. An, X. Xu, and H. He, "Visual saliency based on scale-space analysis in the frequency domain," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 4, pp. 996–1010, Apr. 2013.
[49] N. J. Butko, L. Zhang, G. W. Cottrell, and J. R. Movellan, "Visual salience model for robot cameras," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2008, pp. 2398–2403.
[50] L. Itti, "Quantifying the contribution of low-level saliency to human eye movements in dynamic scenes," *Vis. Cognit.*, vol. 12, no. 6, pp. 1093–1123, 2005.
[51] U. Engelke, H. Kaprykowsky, H.-J. Zepernick, and P. Ndjiki-Nya, "Visual attention in quality assessment," *IEEE Signal Process. Mag.*, vol. 28, no. 6, pp. 50–59, Nov. 2011.
[52] J. A. Redi, H. Liu, R. Zunino, and I. Heynderickx, "Interactions of visual attention and quality perception," *Proc. SPIE*, vol. 7865, p. 78650S, Feb. 2011.
[53] O. Le Meur, A. Ninassi, P. Le Callet, and D. Barba, "Do video coding impairments disturb the visual attention deployment," *Signal Process., Image Commun.*, vol. 25, no. 8, pp. 597–609, Sep. 2010.
[54] L. Zhang, M. H. Tong, T. K. Marks, H. Shan, and G. W. Cottrell, "SUN: A Bayesian framework for saliency using natural statistics," *J. Vis.*, vol. 8, no. 7, p. 32, 2008.
[55] R. J. Snowden, "Sensitivity to relative and absolute motion," *Perception*, vol. 21, no. 5, pp. 563–568, 1992.
[56] D. A. Poggel, H. Strasburger, and M. MacKeben, "Cueing attention by relative motion in the periphery of the visual field," *Perception*, vol. 36, no. 7, pp. 955–970, 2007.
[57] P. Thompson, "Perceived rate of movement depends on contrast," *Vis. Res.*, vol. 22, no. 3, pp. 377–380, 1982.
[58] L. Stone and P. Thompson, "Human speed perception is contrast dependent," *Vis. Res.*, vol. 32, no. 8, pp. 1535–1549, 1992.
[59] E. P. Simoncelli, "Distributed analysis and representation of visual motion," M.S. thesis, MIT, Cambridge, MA, USA, Jan. 1993.
[60] Y. Fang, Z. Wang, and W. Lin, "Video saliency incorporating spatiotemporal cues and uncertainty weighting," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2013, pp. 1–6.

Yuming Fang is currently a Lecturer with the School of Information Technology, Jiangxi University of Finance and Economics, Nanchang, China. He received the Ph.D. degree in computer engineering from Nanyang Technological University, Singapore, in 2013, and the B.E. and M.S. degrees from Sichuan University, Chengdu, China, and the Beijing University of Technology, Beijing, China, respectively. From 2011 to 2012, he was a Visiting Ph.D. Student with National Tsinghua University, Hsinchu, Taiwan. In 2012, he was a Visiting Scholar with the University of Waterloo, Waterloo, ON, Canada. He was also a (Visiting) Post-Doctoral Research Fellow with the IRCCyN Laboratory, PolyTech' Nantes, and the University of Nantes, Nantes, France, University of Waterloo, and Nanyang Technological University, Singapore. His research interests include visual attention modeling, visual quality assessment, image retargeting, computer vision, and 3D image/video processing. He was a Secretary of HHME2013 (the 9th Joint Conference on Harmonious Human Machine Environment). He was also a Workshop Organizer in ICME 2014 and a Special Session Organizer in VCIP 2013 and QoMEX 2014.

Zhou Wang (S'99–M'02–SM'12–F'14) received the Ph.D. degree in electrical and computer engineering from the University of Texas at Austin, Austin, TX, USA, in 2001. He is currently an Associate Professor with the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON, Canada. His research interests include image processing, coding, and quality assessment, computational vision and pattern analysis, multimedia communications, and biomedical signal processing. He has more than 100 publications in these fields with over 20 000 citations (Google Scholar).

Dr. Wang is a member of the IEEE Multimedia Signal Processing Technical Committee (2013–2015). He served or has been serving as an Associate Editor of the IEEE TRANSACTIONS ON IMAGE PROCESSING (2009–2013), PATTERN RECOGNITION (since 2006), and the IEEE SIGNAL PROCESSING LETTERS (2006–2010), and the Guest Editor of IEEE JOURNAL OF SELECTED TOPICS IN SIGNAL PROCESSING (since 2013 and 2007–2009), *EURASIP Journal of Image and Video Processing* (2009–2010), and *Signal, Image and Video Processin* (2011–2013). He was a recipient of the 2014 NSERC E.W.R. Steacie Memorial Fellowship Award, the 2013 IEEE Signal Processing Best Magazine Paper Award, the 2009 IEEE Signal Processing Society Best Paper Award, the 2009 Ontario Early Researcher Award, and the ICIP 2008 IBM Best Student Paper Award (as senior author). He is a registered Professional Engineer in Ontario, Canada.

Weisi Lin (M'92–SM'98) received the Ph.D. degree from King's College London, London, U.K. He is an Associate Professor in Computer Engineering with Nanyang Technological University, Singapore. He was the Lab Head and an Acting Department Manager of Media Processing with the Institute for Infocomm Research, Singapore. His research areas include image processing, perceptual multimedia modeling and evaluation, and visual signal compression and communication. He has published more than 100 refereed journal papers and more than 170 conference papers, and holds seven patents.

He has been on the editorial boards of the IEEE TRANSACTIONS ON MULTIMEDIA (2011–2013), the IEEE SIGNAL PROCESSING LETTERS and *Journal of Visual Communication and Image Representation*. He currently chairs the IEEE MMTC IG on Quality-of-Experience. He was elected as an APSIPA Distinguished Lecturer (2012-2013). He is a Technical-Program Chair of the 2013 IEEE International Conference on Multimedia and Expo, the 2012 Pacific-Rim Conference on Multimedia, and the 2014 International Workshop on Quality of Multimedia Experience. He is a fellow of the Institution of Engineering Technology, and an Honorary Fellow of the Singapore Institute of Engineering Technologists.

Zhijun Fang received the Ph.D. degree from Shanghai Jiao Tong University, Shanghai, China. He is currently a Professor and the Dean with the School of Information Technology, Jiangxi University of Finance and Economics, Nanchang, China. His current research interests include image processing, video coding, and pattern recognition. He was the General Chair of HHME2013 (the 9th Joint Conference on Harmonious Human Machine Environment). He received the Gan Po Elite 555 Plan Award and the One-Hundred, One-Thousand, Ten-Thousand Talent Project Award of Jiangxi Province.