UNSUPERVISED ESTIMATION OF UNCERTAINTY FOR VIDEO SALIENCY DETECTION USING TEMPORAL CUES

Tariq Alshawi, Zhiling Long, and Ghassan AlRegib

Center for Signal and Information Processing (CSIP)
School of Electrical and Computer Engineering
Georgia Institute of Technology, Atlanta, GA 30332-0250, USA
{talshawi, zhiling.long, alregib}@gatech.edu

ABSTRACT

Video saliency detection is typically performed by combining temporal saliency and spatial saliency, which are detected separately. Among various available techniques, uncertainty-based combination is a unique and promising approach. In this paper, we study the uncertainty of each point within a map generated from video saliency detection. We develop an adaptive temporal correlation-based method for unsupervised uncertainty estimation. To evaluate the performance of our algorithm, we propose a systematic evaluation scheme that involves both the creation of ground truth uncertainty data and the comparison of uncertainty estimation results against the ground truth. Our experiments on the public CRCNS database show that the unsupervised uncertainty analysis algorithm is very promising.

Index Terms— unsupervised, uncertainty analysis, video, saliency detection, performance evaluation

I. INTRODUCTION

Bottom-up spatio-temporal saliency detection identifies perceptually important objects or regions in video sequences, which tend to attract attentions from a human observer [1]. Such detection may help video processing (e.g., coding, classification, etc.) to be performed more effectively as well as more efficiently [2]. In the past decade, saliency detection for visual signals such as images and videos has drawn more and more attention from the computer vision community.

As we know, the human visual system (HVS) may respond to temporal changes and spatial changes in a scene in different ways. Based on specific situations, it may be biased towards one type of changes over another. As a typical example, people tend to be attracted more to a moving object than to a static background it moves against. As such, when videos are concerned, it is a common practice to evaluate temporal saliency and spatial saliency separately, and then combine them together to yield an overall saliency detection.

Various techniques are available for combining temporal and spatial saliency maps [3]. Simple methods such as averaging or multiplication of the two are widely adopted. At the same time, more sophisticated methods such as the maximum skewness fusion [4] and the motion priority fusion [5] have also been developed. All these popular techniques attempt to evaluate the importance of saliency types, but do not address the reliability of such saliency values. This may cause problems when some of the saliency values are erroneous, which is actually very common with most saliency detection algorithms when applied to real life video signals.

In a recent work [6], a new fusion scheme denoted as uncertainty weighting was proposed, which addressed this reliability issue. This approach evaluates the uncertainty associated with each saliency value in both types of maps, and then weighs them accordingly in the fusion. To find the uncertainty for a given point, first, its spatial location is compared with respect to the "center of saliency." The probability of this point being salient is computed using an equation obtained from examining a large database of images with human labeled salient objects. From the probability, an entropy value is calculated. Similarly, the point is also examined pertaining to its spatial "connectedness" to the neighboring salient objects or regions. Another entropy value is calculated by utilizing a probability equation obtained using the same training samples. The two entropies are then added to yield the uncertainty estimate.

Although this uncertainty weighting method was verified effective by much enhanced saliency detection results, we believe more work needs to be done as far as uncertainty analysis for saliency detection results is concerned. First, the reported work depends on equations derived from a human labeled image database, thus being supervised. In reality, there is no guarantee that such equations are applicable to any videos. Thus, an unsupervised method will be more desirable. Second, the effectiveness of the uncertainty analysis was not evaluated directly against the ground truth, or the "true uncertainty," but indirectly through the performance improvement of the saliency detection. Conclusions drawn from such indirect evaluation may not be reliable as they could change when different applications are examined. Therefore, the contribution of this paper is twofold. First, we develop an unsupervised method for uncertainty estimation. As far as we know, this is the first proposed unsupervised technique to tackle the problem. Unlike the supervised method in [6], which explores the spatial relationships,

978-1-4799-7591-4/15/$31.00 ©2015 IEEE

our method examines temporal relationships in an adaptive manner. In addition, we develop a systematic evaluation scheme that handles both the generation of ground truth, and the comparison of the estimated results against the ground truth. To our best knowledge, there is no work reported in the literature concerning such direct evaluation. In this paper, our study is conducted on fused saliency maps that have combined both temporal and spatial saliency. We believe the developed methods are applicable to both types of saliency.

II. PROPOSED METHODS

II-A. Unsupervised Uncertainty Estimation

To accomplish an unsupervised uncertainty estimation, our assumption is that temporal correlation exists with the visual saliency. As we observe, changes in saliency for a region or an object are generally gradual. Consequently, if computationally detected saliency does not satisfy this basic assumption, we consider it unreliable, or uncertain. In other words, sudden changes in saliency should lower our trust with that particular spatio-temporal event. Taking into account natural changes in scenery, we propose using adaptive mean deviation of saliency in time domain as an estimate of the associated uncertainty. Specifically, given a saliency map $\boldsymbol{S}$ (Eq. (1)) of size $M \times N$ and of depth K frames, we can construct an uncertainty map $\boldsymbol{U}$ (Eq. (2)) of the same size and depth as $\boldsymbol{S}$ by iteratively processing 1-D signals $S_{mn}[k]$ located at saliency map pixel (m, n) using $W_{mn}^{(L)}[k]$, a moving average obtained by applying a window of length L as detailed in Eq. (3) and (4). Here, m=$1, 2, ..., M$, n=$1, 2, ..., N$, and k=$1, 2, ..., K$.

$$\boldsymbol{S} = \begin{bmatrix} S_{11}[k] & S_{12}[k] & \dots & S_{1N}[k] \\ S_{21}[k] & S_{22}[k] & \dots & S_{2N}[k] \\ \vdots & \vdots & \ddots & \vdots \\ S_{M1}[k] & S_{M2}[k] & \dots & S_{MN}[k] \end{bmatrix} \tag{1}$$

$$\boldsymbol{U} = \begin{bmatrix} U_{11}[k] & U_{12}[k] & \dots & U_{1N}[k] \\ U_{21}[k] & U_{22}[k] & \dots & U_{2N}[k] \\ \vdots & \vdots & \ddots & \vdots \\ U_{M1}[k] & U_{M2}[k] & \dots & U_{MN}[k] \end{bmatrix} \tag{2}$$

$$U_{mn}[k] = |S_{mn}[k] - W_{mn}^{(L)}[k]| \tag{3}$$

$$W_{mn}^{(L)}[k] = \frac{1}{L} \sum_{i=k-\frac{L}{2}}^{k+\frac{L}{2}} S_{mn}[i] \tag{4}$$

Given an appropriate length L, the adaptive average $W_{mn}^{(L)}[k]$ follows the changes in the scene and, to some extent, approximates the common trend of pixel saliency change over time. This method can capture sudden changes in saliency in two different ways. As the peak of saliency value diverges from the adaptive average, so does the estimated uncertainty. Alternatively, as the peak gets narrower (i.e., the rate of change increases), the peak contribution to the adaptive average decreases (thus the deviation from average increases). An illustration is given in Fig. 1. In case 1, the peak value changes while peak width is relatively constant, and one can see larger deviation from the average (dashed line) is associated with larger peak values. Even though, larger peak value will shift the average higher, the damping effect of window averaging will lower the contribution of the peak value, thus capturing this abnormality. In the second case, wider peaks associate with lower deviation values since the adaptive average will catch up to the pixel value changes and appropriately capture the scene change. Obviously, changes captured by the proposed method are relative to the range of the saliency value itself. To ensure consistency across the whole map, the estimated uncertainty is normalized by the local maximum value, therefore fixing the range of the generated uncertainty map to be $[0, 1]$. It is worth noting that the simplicity of our proposed method is rather an advantage, that enables computationally efficient implementation of the uncertainty measure while providing reasonably good results (as demonstrated in the experiments).

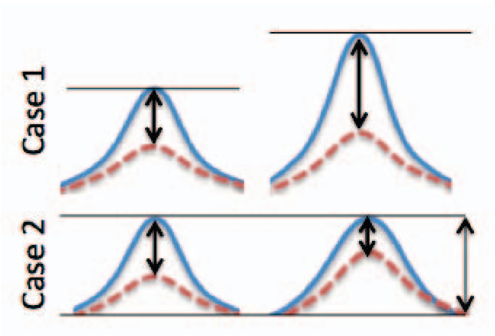
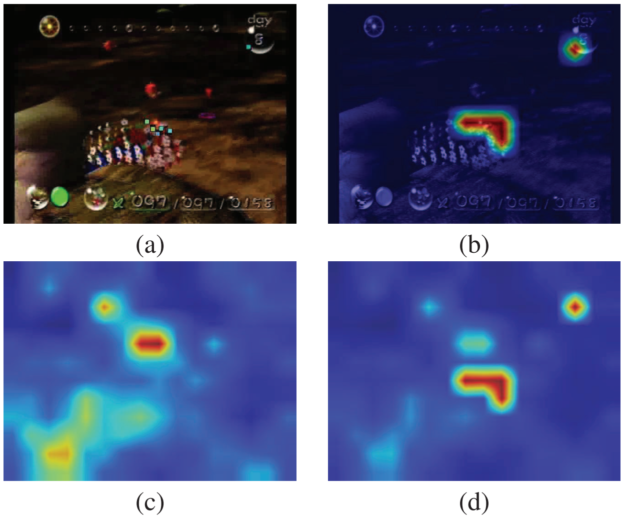


Fig. 1. Illustration of how the proposed uncertainty measure captures abnormality.

II-B. Performance Evaluation

To objectively evaluate the performance of an uncertainty estimation algorithm, ideally we need to compare the estimated uncertainty against the ground truth, or the true uncertainty. However, such true uncertainty data are not readily available. This likely being the reason, to our best knowledge, no work has been reported that performs such kind of objective assessment as far as the application of saliency detection is concerned.

Available databases for saliency detection research usually come with ground truth data recording eye fixations of human subjects viewing the images or videos. Based on the eye fixation data, we propose the following method to generate the true uncertainty data. First, we change the fixation data into a binary map $\hat{F}_{ij}^{tr}[k]$, where i=$1, 2, ..., M'$, j=$1, 2, ..., N'$, and k=$1, 2, ..., K$, with M', N', and K being the height, width, and the total number of frames, respectively. $\hat{F}_{ij}^{tr}[k]$=1 when (i, j, k) corresponds to a location of eye fixations. Second, we resize the binary fixation map $\hat{F}_{ij}^{tr}[k]$ to the same size as the saliency maps $S_{mn}[k]$ from a saliency detection algorithm, where m=$1, 2, ..., M$, n=$1, 2, ..., N$, with M, N being the respective height and



(a) (b) (c) (d)

Fig. 2. Examples illustrating how true uncertainty data are obtained. (a) Original video frame with eye fixation superimposed (small color squares in the center and top-right corner); (b) Resized eye fixation map superimposed on the original frame; (c) Saliency detection results; (d) True uncertainty. We note that the color display is only for a better illustration, which involves some interpolation causing the binary resized fixation map not appearing exactly binary.

width of the saliency maps. This resizing is necessary because many saliency detection techniques work on down-sampled video frames for computational efficiency. However, for the binary map $\hat{F}_{ij}^{tr}[k]$, the resizing is not exactly a downsampling procedure. Denoted as $F_{mn}^{tr}[k]$, the resized binary fixation map is obtained as

$$F_{mn}^{tr}[k] = \bigoplus_{(i,j)\in\Phi} \hat{F}_{ij}^{tr}[k], \tag{5}$$

where $\bigoplus$ refers to the operation of logical OR, and Φ is the set of coordinates that will be mapped to point (m, n) by the downsampling associated with $S_{mn}[k]$. Here, we apply logical OR so that we will not lose the sparse "1"s in the original fixation truth data. Finally, assuming that the saliency map $S_{mn}[k]$ is normalized, we calculate the true uncertainty as

$$U_{mn}^{tr}[k] = \left|S_{mn}[k] - F_{mn}^{tr}[k]\right|. \tag{6}$$

Obviously, $U_{mn}^{tr}[k]$ shows how far away each saliency estimate is from the fixation truth. Thus, it can serve well as a measure of the associated uncertainty. Fig. 2 illustrates this procedure with some examples. Please note that although the fixation maps $\hat{F}_{ij}^{tr}[k]$ and $F_{mn}^{tr}[k]$ are binary, the derived true uncertainty data $U_{mn}^{tr}[k]$ are continuous values because the saliency detection results $S_{mn}[k]$ are continuous.

With the true uncertainty data available, we propose a detection theory-based scheme for the performance evaluation. The scheme generates a receiver operating characteristics (ROC) curve and uses the area under the curve (AUC) as the metric [7]. Since our true uncertainty data $U_{mn}^{tr}[k]$ are real values, they need to be converted to binary data, denoted as $U_{mn}^{trb}[k]$, as the ROC curve is intended for binary classifiers. This conversion is conducted by applying a certain threshold T_1. To generate the ROC curve, the uncertainty estimates $U_{mn}[k]$, obtained using our algorithm presented in Sec. II-A, are also thresholded by T_2 into a binary form, $U_{mn}^{b}[k]$, and compared against $U_{mn}^{trb}[k]$. Thus, both the true detection rate (TDR) and the false positive rate (FPR) are obtained. When we change the value of T_2, sweeping through its whole range, pairs of TDR and FPR are obtained to yield an ROC curve plotted as TDR vs. FPR. Then the AUC is easily computed. AUC is ranged from 0 to 1, with a greater value indicating better performance.

III. EXPERIMENTS

We tested our unsupervised uncertainty estimation algorithm using the public CRCNS database [1]. The database includes 50 videos, with the resolution being 480×640 and the duration ranging from 5 to 90 seconds with 30 frames per second. The videos contents are of diversified nature (a total of 12 categories), covering street scenes, TV sports, TV news, TV talks, video games, etc. and in many cases realistically characterize typical video segments including various lighting conditions, camera movement, speed blur, etc. Eye fixation data are provided with each video, recorded for a group of human subjects watching the videos under freeview condition. For our experiments, we generated saliency maps for the videos using a recent algorithm based on 3D FFT local spectra [8]. The saliency maps are spatially downsampled into 12×16 each frame. For all our experiments, the moving average window length L was empirically set to 10. When we sweep through the range of T_2, we utilized a log-scaled set of values.

We first tested different values for threshold T_1. Fig. 3 shows the resulting ROC curves for the whole dataset using T_1 ranging from 0.3 to 0.7. We observe that all ROC curves are above the reference, which represents guessing by chance. As T_1 increases, the performance drops as demonstrated by the ROC curve moving closer to the reference. However, the performance drops by a smaller margin as T_1 increases. The ROC curves for T_1=0.5, 0.6, 0.7 are very close to each other.

Given the variation of scenes and dynamics in the dataset, we also evaluate the performance of our proposed algorithm for each category. For these experiments, we set T_1=0.4. As shown in Fig. 4, all ROC curves are above the reference, indicating that the algorithm is advantageous over a random guessing. Most of the ROC curves share a similar pattern. The only exception is observed with the video of Saccadetest, which is a much simpler video showing a dot moving against a light textured background.

The overall performance of our proposed method is compared with supervised estimation proposed by [6] in Fig.5, both curves use $T_1 = 0.4$. As observed, our method slightly outperforms the supervised method. For further verification, we also provide visual examples of uncertainty estimates

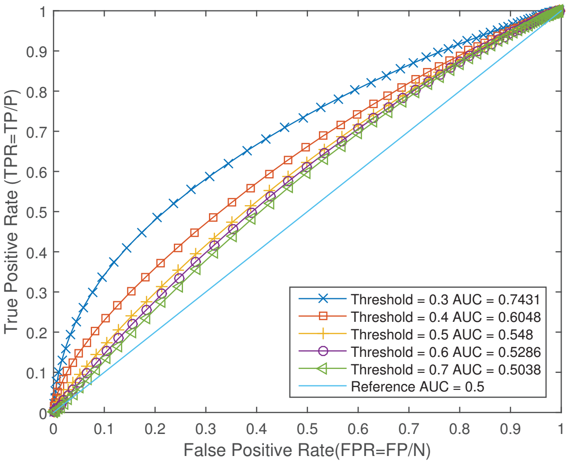


Fig. 3. ROC curves for the whole dataset using different T_1 values.

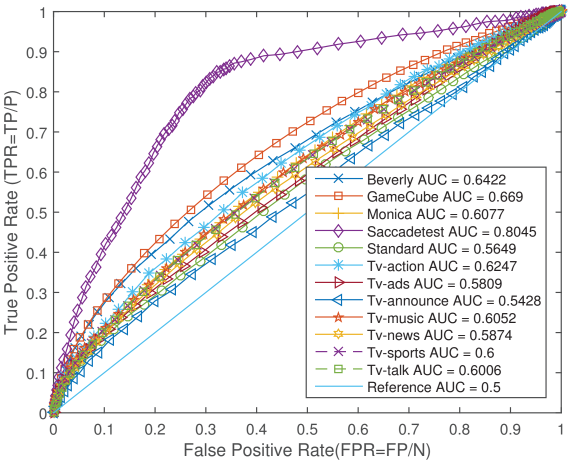


Fig. 4. ROC curves for each category in the dataset, T_1=0.4.

against true uncertainties, both in the time domain and the space domain, in Fig. 6 and 7, respectively. All these results show that our uncertainty estimation algorithm works well, thus validating our basic assumption that temporal deviation from the local average is a viable uncertainty indicator. However, for some videos categories like TV-announce and TV-news, the proposed method doesn't perform as well. We believe this is due to the video content itself being TV dialog with stationary view of the subject in the scene typically overlaid with textual information, which might go beyond basic low-level stimuli and incorporate more advanced attention mechanisms.

IV. CONCLUSIONS

In this paper, we presented an algorithm to estimate uncertainty in visual saliency maps. It relies on the assumption that temporal correlation exists in saliency perceived

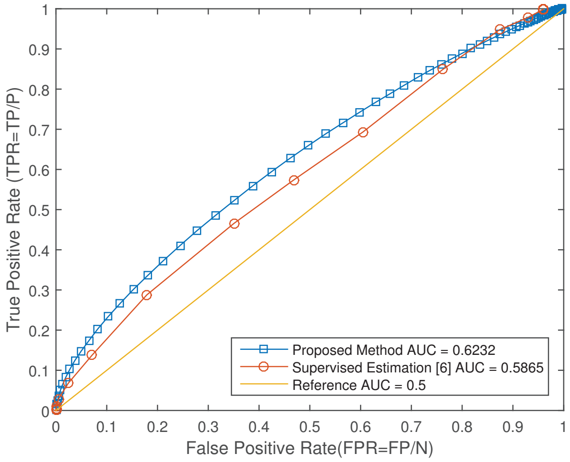


Fig. 5. ROC curves for supervised estimation vs. proposed method, T_1=0.4.

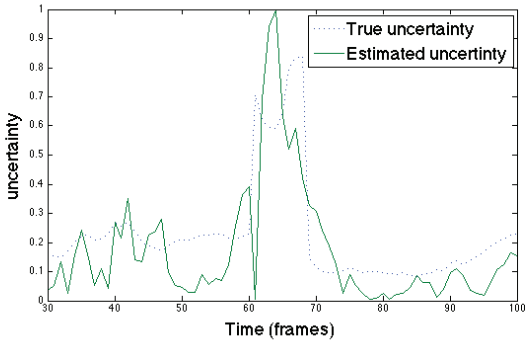


Fig. 6. Example of estimated uncertainty in comparison with true uncertainty along time, for Beverly01, pixel location $(8, 6)$, frame 30 to 100.

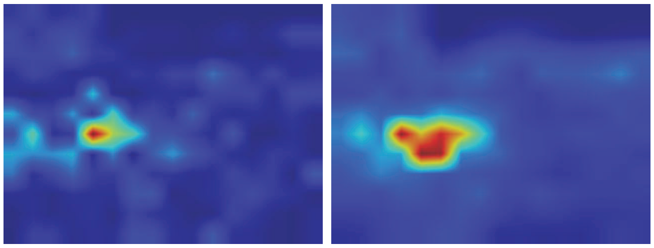


Fig. 7. Example of estimated uncertainty (left) in comparison with true uncertainty (right), for beverly05, frame 187.

by HVS. The method is unsupervised and computationally efficient, which makes it attractive for real-world applications. We also proposed a systematic performance evaluation scheme including the generation of true uncertainty and ROC curve-based objective assessment. The experiments on the public CRCNS database indicated that our algorithm is useful and promising for the intended application.

V. REFERENCES

[1] Laurent Itti, “Quantifying the contribution of low-level saliency to human eye movements in dynamic scenes,” *Visual Cognition*, vol. 12, no. 6, pp. 1093–1123, 2005.

[2] Laurent Itti, “Automatic foveation for video compression using a neurobiological model of visual attention,” *IEEE Transactions on Image Processing*, vol. 13, no. 10, pp. 1304–1318, 2004.

[3] Satya Muddamsetty, Désiré Sidibé, Alain Trémeau, and Fabrice Mériaudeau, “A performance evaluation of fusion techniques for spatio-temporal saliency detection in dynamic scenes,” in *ICIP*, 2013, pp. 1–5.

[4] Sophie Marat, Tien Ho Phuoc, Lionel Granjon, Nathalie Guyader, Denis Pellerin, and Anne Guérin-Dugué, “Modelling spatio-temporal saliency to predict gaze direction for short videos,” *International journal of computer vision*, vol. 82, no. 3, pp. 231–243, 2009.

[5] Jiang Peng and Qin Xiao-Lin, “Keyframe-based video summary using visual attention clues,” *IEEE MultiMedia*, vol. 17, no. 2, pp. 64–73, 2010.

[6] Yuming Fang, Zhou Wang, and Weisi Lin, “Video saliency incorporating spatiotemporal cues and uncertainty weighting,” in *Multimedia and Expo (ICME), 2013 IEEE International Conference on*. IEEE, 2013, pp. 1–6.

[7] Trevor Hastie, Robert Tibshirani, and Jerome Friedman, *The elements of statistical learning: data mining, inference and prediction*, Springer, 2005.

[8] Zhiling Long and Ghassan AlRegib, “Saliency detection for videos using 3D fft local spectra,” in *Human Vision and Electronic Imaging XX, SPIE Electronic Imaging*. SPIE, 2015.