UNSUPERVISED UNCERTAINTY ANALYSIS FOR VIDEO SALIENCY DETECTION

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

This paper presents a new unsupervised uncertainty estimation method for video saliency detection using spatial cues of the saliency map. The algorithm exploits the relationship between a pixel and its spatial neighbours in saliency maps to estimate the uncertainty of the saliency detected at the pixel location. Unlike supervised methods that fits uncertainty model to available training data, the proposed algorithm is based on very simple observation of the eye fixation map, which is largely influenced by human visual attention mechanisms. Thus, the proposed method is data independent. The performance of the proposed algorithm is evaluated using the challenging CRCNS video dataset and quantified using Receiver Operating Characteristics (ROC). The results are promising and could lead to robust uncertainty estimation using eye-fixation neighbourhood modeling.

Index Terms— unsupervised, uncertainty analysis, video, saliency detection, attention framework, spatial correlation

I. INTRODUCTION

Computational visual saliency detection methods attempt to predict interesting regions or objects in a given scene that potentially attract human attention. In bottom-up spatio-temporal saliency detection approaches, various low-level change indicators such as intensity, color, and motion are used as features to assess the amount of discrepancy between a pixel and its neighbours and hence infer its saliency. The output of these algorithms, typically called saliency maps, can be used as pre-processing stage to improve the performance as well as the efficiency of various image and video processing applications including compression, segmentation, and classification.

The performance of saliency detection algorithms is typically evaluated by measuring detection results using images and videos datasets. These datasets such as CRCNS [1], MSRA [2], MIT [3], and SAVAM [4] typically contain different types of saliency groundtruth, e.g., eye-fixation records, bounding box for salient objects, and precise salient object segmentation. The variations in the definition of the groundtruth make the comparison between different algorithms applied on different datasets a very difficult task and in many times it results in inconsistent conclusions. It is common to have one algorithm outperforming another in one dataset but the situation may be reversed when tested using a different dataset, which can be attributed to variation in video sequences and dataset bias. For example, saliency detection algorithms that are conservative tend to achieve better scores in eye-fixation datasets compared to bounding box datasets due to overwhelming negative samples in the former compared to the latter. Additionally, conclusions regarding the performance of different algorithms tested using the same dataset may not be conclusive, because an algorithm can be designed to nicely fit a dataset regardless of the underling physical phenomenon. Factors such as groundtruth generation and evaluation methodology play a big role in the final performance score. Moreover, there are some important questions such as, how much increment in a performance metric is considered significant? or how big a gap between two algorithms needs to be in order to draw conclusions on their performance? These questions cannot be answered without a reliable measure of uncertainty. To address such important issues that are essential to the saliency detection problem, a thorough research and evaluation about the uncertainty inherent of the detection results has to be conducted.

The authors in [5] proposed a supervised method to estimate the uncertainty associated with saliency detection of a video pixel given two simple features calculated using the spatial neighbourhood of a target pixel. The two features are the distance from the center of mass of the saliency map and the connectedness of the target pixel. The coordinates of the center of mass of saliency map $[x_c, y_c]$ is first calculated before using the groundtruth saliency map. Then, the Euclidean distance, d, is used in the data-fitted probability of the pixel being salient given its distance from the center. i.e., $p(s|d)$. Similarly, the connectedness feature, c, is calculated by counting the number of salient neighbours and calculating $p(s|c)$. Finally, the uncertainty U of each pixel is calculated using the binary entropy of the likelihood estimates, $p(s|d)$ and $p(s|c)$. In this paper, we propose investigating spatial cues available within immediate neighbourhood of a pixel in a given frame for uncertainty estimation. More specifically, we propose constructing an uncertainty estimate by calculating pixel's deviation from the spatial cues of its direct neighborhood, which adapts

to scene changes and captures global variation of saliency in area surrounding the pixel. This uncertainty estimation can be used to improve performance evaluation of saliency detection algorithms as well as provide means to compare and select output decision of ensemble of saliency algorithms by choosing the decision with least uncertainty. Additionally, we propose an uncertainty-based framework that can be used to improve saliency-based video processing applications by making more reliable decisions on the saliency map given the estimated uncertainty. The proposed estimation method and framework are not tied to any specific saliency detection method and can be applied to a variety of settings.

II. PROPOSED FRAMEWORK

In an attempt to mimic the advanced processing capability of the human vision system (HVS), saliency detection has been incorporated into various image and video processing algorithms for improved performance. The diversified applications include but are not limited to compression [6], segmentation [7], object recognition [8], tracking [9], and quality assessment [10]. However, there has been no explicit design of a saliency-based video processing framework, to the best of our knowledge. Most of the proposed methods do not evaluate the validity of saliency maps generated online, but rather design or choose a saliency detection algorithm that exhibits good performance in evaluation datasets, and then hope for the best when the algorithm goes online. Therefore, we propose a unified framework for enhancing video processing algorithms using saliency that is neither application- nor algorithm-specific and can be reliable in real world scenarios. The proposed uncertainty-based framework is depicted in Fig.1. It evaluates the saliency map and produces associated uncertainty map that describes the level of confidence in the generated saliency map. By reliably estimating uncertainty, we can expand the framework to include a systematic decision-making procedure that makes application-specific decisions. Additionally, having a separate module for decision making helps clarify which assumptions are application-specific and which ones are saliency related. Knowledge about the application space can influence the design of this module without making drastic changes to the whole framework. Similarly, the availability of uncertainty estimations allows for risk assessment that can be used to guide the optimization of video processing algorithms.

III. UNSUPERVISED UNCERTAINTY ESTIMATION

To accomplish an unsupervised uncertainty estimation, our assumption is that both temporal correlation and spatial correlation exist in visual saliency. We showed in [11] that pixel-wise temporal consistency can help estimate the uncertainty of pixel's saliency given its divergence from the local mean. We note similar observation, as did others [5], regarding the spatial domain. Typically, saliency maps consist of a number of concentrated salient pixels that correspond to a region or an object, and rarely exist a lone salient pixel. Consequently, if computationally detected saliency does not satisfy this basic assumption, we consider it unreliable, or

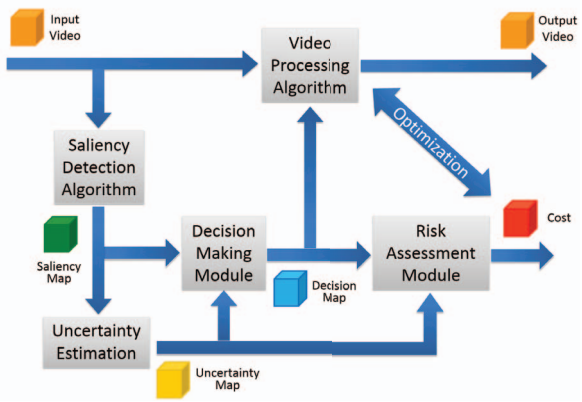

Fig. 1. Uncertainty-based framework for improving saliency-enabled video processing algorithms.

uncertain. In other words, a substantial variation between a pixel saliency and its neighbourhood saliency should lower our trust with that particular pixel in saliency map. Taking into account natural changes in scenery, we propose using adaptive mean deviation of saliency in spatial 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 saliency frames $S_k[m,n]$ using $W_k^{L\times L}[m,n]$, a 2-D average filter of size $L \times L$ as detailed in Eq.(3), which can be reduced to Eq.(4). Here, m=1, 2, ..., M, n=1, 2, ..., N, and k=1, 2, ..., K.

$$\boldsymbol{S} = [S_1[m,n] \quad S_2[m,n] \quad \dots \quad S_K[m,n]] \tag{1}$$

$$\boldsymbol{U} = [U_1[m,n] \quad U_2[m,n] \quad \dots \quad U_K[m,n]] \tag{2}$$

$$U_k[m,n] = |S_k[m,n] - W_k^{L\times L}[m,n] * S_k[m,n]| \tag{3}$$

$$U_k[m,n] = |S_k[m,n] * Kernel^{L\times L}[m,n]| \tag{4}$$

Given an appropriate size L, $Kernel^{L\times L}[m,n]$ adapts to the changes in the scene and, to some extent, approximates the overall trend of pixel saliency changes over the saliency map. 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.

IV. 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 [12]. The saliency maps were generated for spatially downsampled video frames with the frame size being 12×16..

To objectively evaluate the performance of the uncertainty estimation algorithm, ideally we need to compare the estimated uncertainty against the ground truth, or the true uncertainty. However, such true uncertainty data is not readily available. To tackle this problem, we use the performance evaluation methodology recently proposed in [11] and is illustrated in Fig. 2). In this process, we generate the true uncertainty data and evaluate the estimated uncertainty using ROC analysis.

The details of the evaluation method are presented as follows. First, we change the fixation data into a binary map $\hat{F}_k^{tr}[i,j]$, where i=$1,2,...,M'$, j=$1,2,...,N'$, and k=$1,2,...,K$, with M', N', and K are the height, the width, and the total number of frames, respectively. $\hat{F}_k^{tr}[i,j]$=1 when $[i,j,k]$ corresponds to a location of a recorded eye fixation. Second, we resize the binary fixation map $\hat{F}_k^{tr}[i,j]$ to the same size as the saliency maps $S_k[m,n]$ from a saliency detection algorithm, where m=$1,2,...,M$, n=$1,2,...,N$, with M, N being the respective height and width of the saliency maps. This resizing is necessary because many saliency detection techniques work on downsampled video frames for computational efficiency. However, for the binary map $\hat{F}_k^{tr}[i,j]$, the resizing is not exactly a downsampling procedure. Denoted as $F_k^{tr}[i,j]$, the resized binary fixation map is obtained as

$$F_k^{tr}[m,n] = \bigoplus_{(i,j)\in\Phi} \hat{F}_k^{tr}[i,j], \tag{5}$$

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 $S_k[m,n]$. 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 $S_k[m,n]$ is normalized, we calculate the true uncertainty as

$$U_k^{tr}[m,n] = \big|S_k[m,n] - F_k^{tr}[m,n]\big|. \tag{6}$$

Obviously, $U_k^{tr}[m,n]$ captures the distance between the saliency estimate and the fixation truth. Thus, it can serve as a measure of the associated uncertainty. Please note that although the fixation maps $\hat{F}_k^{tr}[i,j]$ and $F_k^{tr}[m,n]$ are binary, the derived true uncertainty data $U_k^{tr}[m,n]$ are continuous values because the saliency detection results $S_k[m,n]$ are continuous.

With the true uncertainty data available, we can use detection theory-based scheme for the performance evaluation.

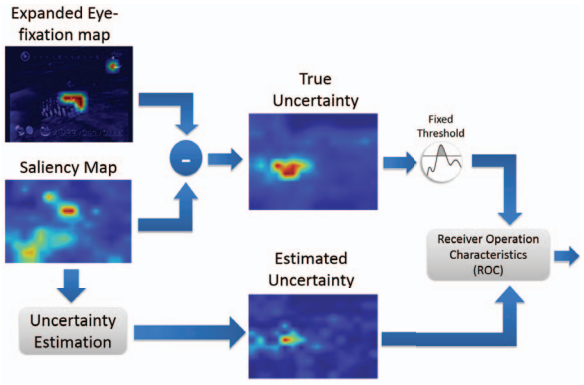

Fig. 2. Evaluation methodology proposed in [11].

The scheme generates a receiver operating characteristics (ROC) curve and uses the area under the curve (AUC) as the metric [13]. Since our true uncertainty data $U_k^{tr}[m,n]$ are real values, they need to be converted to binary data, denoted as $U_k^{trb}[m,n]$, 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 $U_k[m,n]$, obtained using our algorithm presented in Sec. III, are also thresholded by T_2 into a binary form, $U_k^b[m,n]$, and compared against $U_k^{trb}[m,n]$. 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.

We tested different values for threshold Th_1 and kernel size $L = 3$. 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, and follow similar trends suggesting that the underling phenomenon is similar regardless of exact threshold value. 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 Th_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 separately. 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. Closely examining the ROC curves behaviour indicate rather interesting patterns regarding the category contents. The three highest categories, in terms of AUC values, are saccadetest (showing a dot moving against a light textured background), TV-action (showing a person sliding through a designed course), and Beverly (showing different scenes of people playing and running in parks), all of which are semantically non-complex and require little to no effort to comprehend the scene. On the

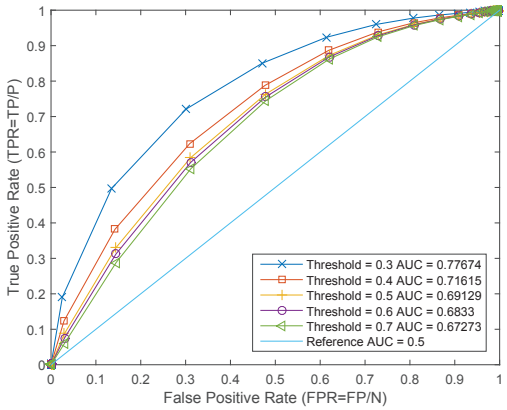

Fig. 3. ROC curves for the whole dataset using different T_1 values.

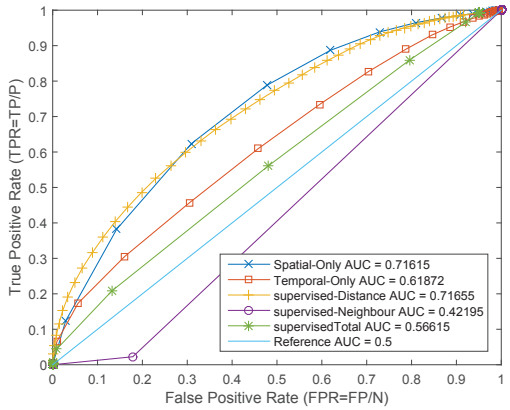

Fig. 5. ROC curves for supervised estimation versus the proposed unsupervised method, T_1=0.4 and L=3.

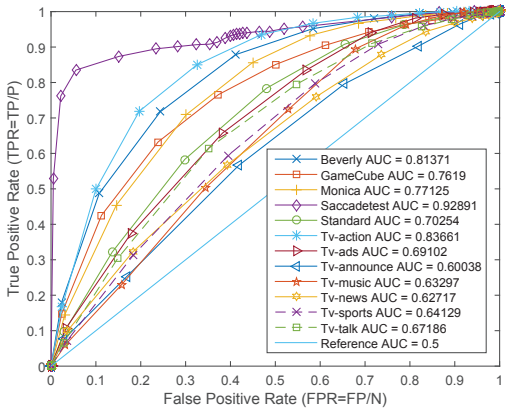

Fig. 4. ROC curves for each category in the dataset, T_1=0.4 and L=3.

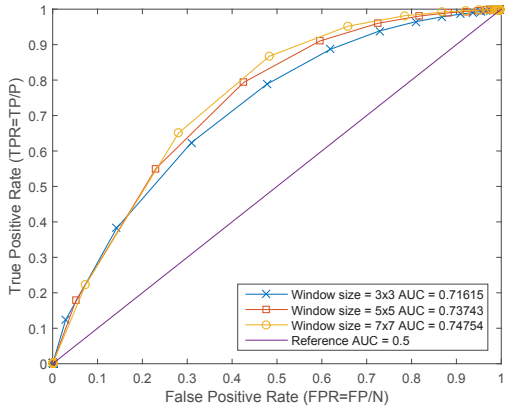

Fig. 6. ROC curves for different window sizes, T_1=0.4.

other hand, categories that are more complex semantically are not performing as well, especially the ones with overly textual information such as TV-news, TV-announcement, and TV-music.

The overall performance of our proposed method is compared with uncertainty estimation from temporal cues [11] and supervised estimation proposed in [5] in Fig.5, where both temporal-only and spatial-only curves use $T_1 = 0.4$. As observed, our method outperforms both temporal-based uncertainty and the supervised method (denoted as supervisedTotal in the figure). Also, to further understand the supervised algorithm, we include the performance of each component of the supervised algorithm (supervised-Distance and supervised-Neighbour), separately. We notice that a major contribution of the supervised algorithm performance is made by the distance from center of mass, which is closely related to our assumption and hence performs similarly.

Additionally, we were interested in studying the effect of different supporting region for the processing $Kernel^{LXL}[m,n]$, specifically, the effect of the size parameter L. As shown in Fig.6, the average performance of the algorithm slightly improves as the size of the kernel increases, which can be attributed to the kernel being able to capture more context of the scene hence improving its estimation. All these results show that our spatial-based uncertainty estimation works well, thus validating our basic assumption that spatial deviation from the local average is a viable uncertainty indicator.

V. CONCLUSIONS

In this paper, we presented an algorithm to estimate uncertainty using spatial cues in visual saliency maps for videos. It relies on the assumption that correlation exists in saliency perceived by HVS and a lone salient pixel rarely exists. The method is inspired by HVS properties, which makes it independent of datasets. In addition, the method is formulated in a computationally efficient way, which makes it attractive for practical applications. We also proposed an uncertainty-based framework that is generic and can be

used for any saliency-enabled image and video processing application. The experiments on the public CRCNS database indicated that our algorithm is useful and promising. Experiments suggest that the algorithm is performing well when the test videos are semantically non-complex, and more challenging videos might require more advanced features than the ones proposed in this paper.

VI. 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] Tie Liu, Zejian Yuan, Jian Sun, Jingdong Wang, Nanning Zheng, Xiaoou Tang, and Heung-Yeung Shum, "Learning to detect a salient object," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 33, no. 2, pp. 353–367, Feb 2011.

[3] Tilke Judd, Krista Ehinger, Frédo Durand, and Antonio Torralba, "Learning to predict where humans look," in *Computer Vision, 2009 IEEE 12th International Conference on*, Sept 2009, pp. 2106–2113.

[4] Yury Gitman, Mikhail Erofeev, Dmitriy Vatolin, Bolshakov Andrey, and Fedorov Alexey, "Semiautomatic visual-attention modeling and its application to video compression," in *Image Processing (ICIP), 2014 IEEE International Conference on*, Oct 2014, pp. 1105–1109.

[5] 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.

[6] Chenlei Guo and Liming Zhang, "A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression," *Image Processing, IEEE Transactions on*, vol. 19, no. 1, pp. 185–198, Jan 2010.

[7] Wenguan Wang, Jianbing Shen, and F. Porikli, "Saliency-aware geodesic video object segmentation," in *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*, June 2015, pp. 3395–3402.

[8] Zhixiang Ren, Shenghua Gao, Liang-Tien Chia, and I.W.-H. Tsang, "Region-based saliency detection and its application in object recognition," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 24, no. 5, pp. 769–779, May 2014.

[9] V. Mahadevan and N. Vasconcelos, "Biologically inspired object tracking using center-surround saliency mechanisms," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 35, no. 3, pp. 541–554, March 2013.

[10] W. Zhang, A. Borji, Z. Wang, P. Le Callet, and H. Liu, "The application of visual saliency models in objective image quality assessment: A statistical evaluation," *Neural Networks and Learning Systems, IEEE Transactions on*, vol. PP, no. 99, pp. 1–1, 2015.

[11] Tariq Alshawi, Zhiling Long, and Ghassan AlRegib, "Unsupervised uncertainty estimation in saliency detection for videos using temporal cues," in *IEEE Global Conf. on Signal and Information Processing (GlobalSIP), Orlando, Florida, Dec. 14-16*. SPIE, 2015.

[12] 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.

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