Unsupervised Uncertainty Estimation Using Spatiotemporal Cues in Video Saliency Detection

Tariq Alshawi, Zhiling Long, and Ghassan AlRegib

Abstract

In this paper, we address the problem of quantifying the reliability of computational saliency for videos, which can be used to improve saliency-based video processing algorithms and enable more reliable performance and objective risk assessment of saliency-based video processing applications. Our approach to quantify such reliability is two fold. First, we explore spatial correlations in both the saliency map and the eye-fixation map. Then, we learn the spatiotemporal correlations that define a reliable saliency map. We first study spatiotemporal eye-fixation data from the public CRCNS dataset and investigate a common feature in human visual attention, which dictates a correlation in saliency between a pixel and its direct neighbors. Based on the study, we then develop an algorithm that estimates a pixel-wise uncertainty map that reflects our supposed confidence in the associated computational saliency map by relating a pixel's saliency to the saliency of its direct neighbors. To estimate such uncertainties, we measure the divergence of a pixel, in a saliency map, from its local neighborhood. Additionally, we propose a systematic procedure to evaluate uncertainty estimation performance by explicitly computing uncertainty ground truth as a function of a given saliency map and eye fixations of human subjects. In our experiments, we explore multiple definitions of locality and neighborhoods in spatiotemporal video signals. In addition, we examine the relationship between the parameters of

Tariq Alshawi is with the Electrical Engineering Department, College of Engineering, King Saud Univeristy, Riyadh, Saudi Arabia, and also with the Center for Signal and Information Processing (CSIP), School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, 30332 USA (e-mail: talshawi@gatech.edu).

Zhiling Long and Ghassan AlRegib are with the Center for Signal and Information Processing (CSIP), School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, 30332 USA (e-mail: zhiling.long@gatech.edu; alregib@gatech.edu).

Manuscript received ; revised

our proposed algorithm and the content of the videos. The proposed algorithm is unsupervised, making it more suitable for generalization to most natural videos. Also, it is computationally efficient and flexible for customization to specific video content. Experiments using three publicly available video datasets show that the proposed algorithm outperforms state-of-the-art uncertainty estimation methods with improvement in accuracy up to 63% and offers efficiency and flexibility that make it more useful in practical situations.

Index Terms

Unsupervised estimation, saliency detection, uncertainty, video signal processing, visual attention, video saliency learning.

I. Introduction

Human visual attention modeling and understanding can be a key contributor to the advancement in computational analysis of big visual data which might offer similar computational efficiency to that of the human vision system (HVS). Algorithms for object detection and recognition [1], scene understanding [2], video compression [3], and multimedia summarization [4] can be designed to exploit human visual attention mechanisms, and potentially, produce faster and more perceptually satisfying results. Driven by the fast responsiveness of HVS to low-level visual features, bottom-up spatiotemporal saliency detection has been crafted to identify perceptually unique objects in videos, and in turn predict the likelihood that a human would focus on these objects as opposed to the rest of the scene [5].

The majority of existing research efforts focus on computational saliency models [6] [7] [8] [9] [10] [11], however, less attention has been given to quantifying the reliability of the generated saliency maps [12] [13] [14]. The validity of such maps is crucial for integrating visual attention in various image and video processing applications. It is a common practice to consider the validity of a saliency detection model, at every pixel, to be directly related to its average performance on image and video datasets. In other words, a saliency detection model is, first, evaluated using typical visual stimuli datasets with eye tracking data such as CRCNS [15], MSRA [16], MIT [17], and SAVAM [3]. Then, algorithms that detect salient regions effectively, according to a predefined ground truth in the dataset, are assumed to perform well when used in various applications. However, such saliency detectors might fail to produce reliable results in certain contexts or situations, despite their superior performance in other contexts. Thus, it is important to consider the reliability of a saliency map given the context of the image or video

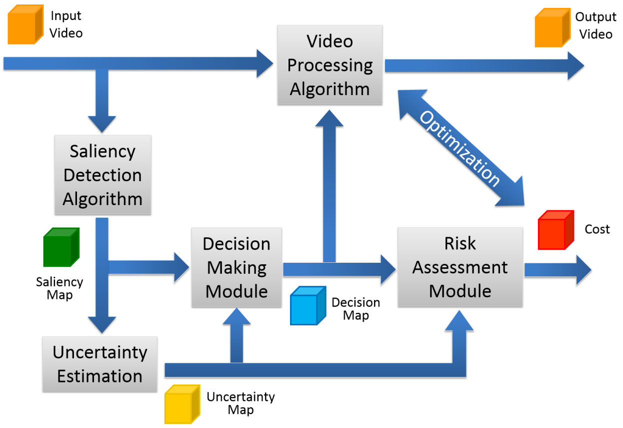

Fig. 1: Uncertainty-based framework for improving saliency-enabled video processing algorithms

at hand. Additionally, explicitly quantifying the reliability of a saliency map enables effective decision making and risk assessment in applications that exploit visual attention mechanisms. We recently proposed an uncertainty-based saliency-enabled framework [14], shown in Fig.1, for image and video processing applications that incorporates the reliability (uncertainty) of the computed saliency maps to enable a systematic decision making process and facilitate risk assessment that depends on the application at hand such as object detection and recognition [1], scene understanding [2], video compression [3], and multimedia summarization [4].

Recently, there has been some research on uncertainty specific to image and video processing applications. In [18], authors proposed using an active learning algorithm based on one-versus-one (OVO) strategy support vector machine (SVM) to solve multi-class image classification. The results of OVO SVM are combined according to a cost function that maximizes the diversity of the chosen set of examples and minimizes the uncertainty of the classification of this set. The uncertainty in this work is estimated using the difference in number of votes between the highest votes class and the second highest class. As the difference in number of votes increases, it is more likely that the highest votes class is the true representative class, so the uncertainty is lower. In the context of medical image registration, Saygili et al. [19] proposed a confidence measure that reflects the accuracy of the registration process of a pair of images. The proposed measure relates the confidence of the registration process at each pixel to the global minima and the steepness of a predefined cost function. The registration at a given pixel is expected to be more reliable if the associated cost function produces a global minima at that location and the cost function in its local region is very steep. In the context of stereo vision and depth estimation, numerous confidence measures have been proposed in literature [20]. Typically, these measures associate the confidence of pixel's match with the shape of the matching cost function, e.g. sum of absolute differences, around that pixel. Haeusler et al. [21] proposed applying random decision forest framework on a large set of diverse stereo confidence measures to improve the performance of stereo solvers.

In the context of saliency detection, there has been very limited work to address the problem of quantifying uncertainty. Directly applying uncertainty and confidence measures proposed for other image and video processing applications might not take into consideration characteristics of human visual attention mechanisms which are crucial for saliency detection. It is important to note here that the term uncertainty has been used in saliency detection research and visual attention modeling to describe a phenomenon that steers attention. The authors in [22] argue that attention can be understood as inferring the level of uncertainty during perception. Other papers such as [23], [24], [25], [26] have proposed saliency attention models that are based on entropy and information theory measures that quantify the level of uncertainty in visual stimuli. However, in the context of our work, uncertainty estimation is mainly concerned with quantifying the reliability of saliency maps. In other words, we are interested in quantifying the confidence in saliency maps during decision making process rather than the entropy caused by uncertainty during perception. Relevant to this kind of uncertainty, the authors in [12] proposed a supervised method to estimate the uncertainty associated with detected saliency of a video pixel. The method uses binary entropy function to measure uncertainty according to the probability of a pixel being salient given the distance of the target pixel from the center of mass $p(s|d)$, and connectedness of the target pixel $p(s|c)$. The coordinates of the center of mass of saliency map $[x_c, y_c]$ are first calculated using the ground truth map. Then, the Euclidean distance, d, is calculated for each pixel in the computed saliency map. Similarly, the connectedness feature, c, is calculated by counting the number of salient neighbors. The probability densities $p(s|d)$ and $p(s|c)$ are fitted using salient object segmentation ground truth from images dataset by Achanta et al. [27]. Despite the fact that the proposed algorithm has been reported to yield enhanced saliency detection results, we believe there are four fundamental issues that are overlooked. First, the modeling of the probability densities $p(s|d)$ and $p(s|c)$, being supervised and based on ground truth from images dataset, may not be generally applicable to videos. Second, the uncertainty estimation is based on individual frames of saliency map, thus, losing cues about uncertainty in temporal axis. Third, the method does not offer any degrees of freedom in customizing the estimation process to video content, despite the diverse nature of videos in real-world applications. Finally, an indirect evaluation is performed by showing that the uncertainty-based fusion of spatial and temporal saliency maps is enhanced over other fusion methods. Thus, a direct application-independent performance evaluation methodology is missing.

To understand the visual attention mechanism, research usually relies on eye-tracking data analysis to form eye fixation maps. Such maps capture the focus of human subjects watching test videos and potentially correlate well with their visual attention. These maps are often used as the ground truth for saliency in learning-based methods, or as feature space in unsupervised methods. Nevertheless, there has been limited research analyzing the structure of these eye-fixation maps separately from visual stimuli. By studying the eye-fixation maps, we expect to better understand the spatial correlation in video scenes, and henceforth to better understand visual attention mechanisms. The authors in [28] analyzed eye-fixation data of images given location and time sequence of human subjects gaze, using spectral decomposition of the correlation matrix constructed based on eye fixation data of different subjects. Their work shows that the first eigenvector is responsible for roughly 21% of the data, and it correlates well with salient locations in the images dataset. In [29], the authors found that it is possible to decode the stimulus category by analyzing statistics (location, duration, orientation, and slope histograms) of fixations and saccades. They used a subset of the NUSEF dataset [30] containing five categories over a total of 409 images.

In this paper, we address the problem of quantifying the uncertainty of detected saliency maps for videos. First, we study spatiotemporal eye-fixation data from the public CRCNS dataset and demonstrate that typically there is high-correlation in saliency between a pixel and its direct neighbors. Then, we propose estimating a pixel-wise uncertainty map that reflects our confidence in the computational saliency map by relating a pixel's value to the values of its direct neighbors in a computationally efficient way. The novelty of this method is that it is unsupervised and independent from the dataset used for testing, which makes it more suitable for generalization. Also, the method exploits information from both spatial and temporal domain, thus, it is uniquely suitable for videos. Moreover, the flexibility of the algorithm parameters allows for customization to specific video content. Additionally, we propose a systematic procedure to evaluate uncertainty estimation performance by explicitly computing uncertainty ground truth in terms of a given saliency map and eye fixations of human subjects watching the associated video segment.

II. Analysis of the Eye-Fixation Data

To motivate the proposed uncertainty estimation method, we present in this section analysis of recorded eye-fixation maps provided in CRCNS as performed in our preliminary study in [31]. In particular, the analysis quantifies the predictability of a pixel in the eye-fixation map given the knowledge of its spatial context. By modeling the pixels of eye-fixation maps and the average of their neighborhood as random variables, we infer the correlation between the eye-fixation map pixels and their immediate $3 \times 3 \times 3$-neighborhood by computing entropy of the eye-fixation pixels versus the entropy of the eye-fixation pixels conditioned on the average of their neighbors. Using the basic properties of entropy, if the neighborhood average completely determines the eye-fixation pixel value then the conditional entropy is equal to zero. Otherwise, the conditional entropy can be any value between zero and a maximum equal to the entropy of the eye-fixation map pixels depending on the correlation between the two quantities (i.e., the pixel value and that of its neighbors). Additionally, to verify the statistical significance of this correlation we compute the entropy of the eye-fixation pixel entropy conditioned on uniformly-distributed random variable.

As reported in [31] and shown in Fig. 2, in most cases, there is roughly a 50% reduction in entropy when conditioned on neighboring pixels average $H(X|Z)$ compared with the eye-fixation pixels entropy $H(X)$. Notably, the results shown in Fig. 2 are normalized to the highest conditional entropy in the dataset. We can observe from Fig. 2 that the entropy reduction is consistent across the dataset regardless of the video content. Also, to avoid confusing this reduction with a computational limitation error, we compute the entropy of the eye-fixation map given the value of a uniformly distributed random variable $H(X|n)$. As seen in Fig.2, the gap between the two conditioned entropy is quite significant indicating the existence of a structure in the eye-fixation maps that can be exploited. Entropy reduction is consistent regardless of probability distribution skewness, as well. This can be shown by redistributing a portion of the probability density from the zero symbol, which dominates eye-fixation maps and explains its low entropy value, to the rest of probability set. This correlation between a pixel value and the average of its neighborhood can be exploited to obtain a rough estimate of a pixel uncertainty in the saliency map given its direct neighborhood, which we detail in the next section.

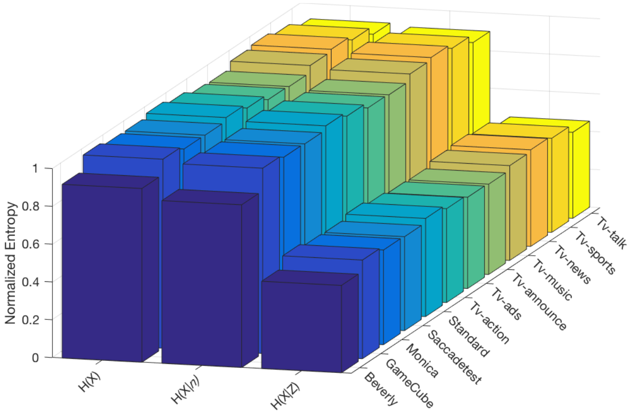

Fig. 2: Entropy reduction across all videos in CRCNS dataset. Results reported here are computed using *Scale 1* saliency map of size 12×16.

III. Unsupervised Uncertainty Estimation using local spatiotemporal neighborhood Cues

As discussed in the previous section, pixels in eye-fixation maps are correlated and such dependency can be exploited to identify unlikely occurrences in the computational saliency maps. Basically, we assume that visual saliency is consistent and changes in saliency values happen gradually. Thus, sudden changes in saliency value should lower our trust in that particular spatiotemporal event. Thus, saliency map pixels that are significantly different from their neighborhood are most likely uncertain and should be examined more carefully.

However, the size of local neighborhoods crucially depends on the video content. For example, fast action videos would most likely have a small group of contiguous correlated pixels, in the saliency map, around location of the main scene actor. In contrast, a slow changing scene gives the viewers more freedom to explore different parts of the video frame, thus, the corresponding eye fixation map would have a larger group of pixels that are correlated. Therefore, it is important to include uncertainty cues from the appropriate scales in order to more reliably capture context-based events.

In most video saliency detection algorithms, the processing of video frames usually consumes significant computation time. Hence, a common practice is to resize the input video frames to several sizes and define saliency maps generated in terms of the frame scale. It is worth noting that saliency maps generated from size-reduced video frames differ from saliency maps downsampled from saliency maps of higher scale. In the first case, video details lost in the downsampling process are not included in the downsampled saliency map, while in the second case, downsampled saliency maps still maintain such details. Generally, uncertainty estimation should take advantage of saliency maps of multiple scales to enhance the estimation performance. One way to approximate the contribution to uncertainty estimation from different scales is to generate a multi-scale uncertainty map that is a weighted combination of uncertainty generated from different scales. In this paper, we focus our study on how to estimate uncertainty from a single scale.

Formally, given a saliency map $\boldsymbol{S}^{(d)}$ of scale d and size $M \times N$ and of depth K frames, we seek to estimate an uncertainty map $\boldsymbol{U}^{(d)}$ of the same scale, size and depth as $\boldsymbol{S}^{(d)}$ that is roughly approximated by saliency value divergence from spatiotemporal local neighborhood mean. The estimation is efficiently computed by processing the map $\boldsymbol{S}^{(d)}$ according to Eq.(1):

$$\boldsymbol{U}^{(d)} = \gamma \big| \alpha \boldsymbol{S}^{(d)} * W^{L_1 \times L_2 \times L_3} \big|, \tag{1}$$

where $\big|.\big|$ is the operation to find the absolute value and $d = 1, 2, ...D$ is the scale label, $L_1 \times L_2 \times L_3$ is the size of the spatiotemporal kernel $W^{L_1 \times L_2 \times L_3}$, α is a scaling factor for the saliency map to fix its range to be [0,1], and γ is a scaling factor for the uncertainty map to ensure the output range is [0,1]. In this paper, we use a simple averaging kernel defined as follows

$$W^{L_1 \times L_2 \times L_3} = \begin{cases} \frac{R-1}{R} & \text{at the center} \\ -\frac{1}{R}, & otherwise, \end{cases} \tag{2}$$

where R = $L_1 \times L_2 \times L_3$. The design of $W^{L_1 \times L_2 \times L_3}$ can be viewed as the difference between saliency value and a moving average window of size $L_1 \times L_2 \times L_3$. $W^{L_1 \times L_2 \times L_3}$, with appropriate size, can follow the changes in the scene and, to some extent, approximates the common trend of pixel saliency change over time.

In order to systematically analyze spatiotemporal uncertainty estimation, we study the contribution of spatial neighbors separate from temporal neighbors which might lead to a better understanding of spatial context in saliency maps. Thus, we introduce in the following subsections two special cases of the proposed algorithm: uncertainty estimation from temporal cues and uncertainty estimation from spatial cues. Relying only on temporal neighbors, the proposed algorithm estimates the uncertainty of a pixel in frame k by studying its correlation with its neighbors in the same location across all K frames, as we have proposed in [13]. By dividing the saliency map into temporal neighborhoods, we can treat each pixel location as separate 1-D signal that can be processed using a simple 1-D filter of length L_t to calculate pixel-neighborhood divergence. Similarly, we can divide the saliency map into spatial neighborhoods that span $L_{s_1} \times L_{s_2}$ pixels in a single frame, as we have reported in [14].

A. Uncertainty Estimation from Temporal Cues

For a given saliency map $\boldsymbol{S}$ of size $M \times N$ and of depth K frames, we decompse the map into 1-D signals as follows

$$\boldsymbol{S} = \begin{bmatrix} s[1,1] & s[1,2] & \dots & s[1,n] & \dots & s[1,N] \\ s[2,1] & s[2,2] & \dots & s[2,n] & \dots & s[2,N] \\ \vdots & \vdots & \dots & \vdots & \dots & \vdots \\ s[m,1] & s[m,2] & \dots & s[m,n] & \dots & s[2,N] \\ \vdots & \vdots & \dots & \vdots & \dots & \vdots \\ s[M,1] & s[M,2] & \dots & s[M,n] & \dots & s[M,N] \end{bmatrix}, \tag{3}$$

where m = $1, 2, ..., M$, n = $1, 2, ..., N$, are the spatial coordinates of the saliency map.

We seek to construct an uncertainty map $\boldsymbol{U}$ of the same size and depth as $\boldsymbol{S}$ by iteratively processing 1-D signals $\boldsymbol{s}$ located at saliency map pixel $[m, n]$ according to

$$U[m,n] = \gamma \big| \alpha S[m,n] * W^{L_t} \big|, \tag{4}$$

where m = $1, 2, ..., M$, n = $1, 2, ..., N$, are the spatial coordinates of both the saliency map and uncertainty map, α and γ are scaling factors, and W^{L_t} is the temporal filter of length L_t, defined by

$$W^{L_t} = [\frac{-1}{L_t} ... \frac{-1}{L_t}, \frac{L_t - 1}{L_t}, \frac{-1}{L_t} ... \frac{-1}{L_t}], \tag{5}$$

B. Uncertainty Estimation from Spatial Cues

Similar to the temporal neighborhood case, given a saliency map $\boldsymbol{S}$ (Eq. (6)) of size $M \times N$ and of depth K frames, we construct an uncertainty map $\boldsymbol{U}$ (Eq. (7)) of the same size and depth as $\boldsymbol{S}$ by iteratively processing saliency frames S_k using a 2-D averaging kernel $W^{L_{s_1} \times L_{s_2}}$ (Eq.(8)) of size $L_{s_1} \times L_{s_2}$.

$$\boldsymbol{S} = \begin{bmatrix} S_1 & S_2 & \dots & S_K \end{bmatrix}, \tag{6}$$

$$\boldsymbol{U} = \Big[U_1 \quad U_2 \quad \dots \quad U_K\Big], \tag{7}$$

$$U_k = \gamma\big|\alpha S_k * W^{L_{s_1} \times L_{s_2}}\big|, \tag{8}$$

where k = $1, 2, ..., K$ is the frame index, $W^{L_{s_1} \times L_{s_2}}$ is a spatial filter similar to averaging kernel W^{L_t}, symmetrical around its center and has a size of $L_{s_1} \times L_{s_2}$, α and γ are scaling factors.

IV. Methods for Ground Truth Generation and 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 is not readily available.

1) Computing True Uncertainty: Available databases for saliency detection research usually contain ground truth data recording eye fixations of human subjects viewing the images or videos. Based on the eye fixation data, as we proposed in [13], the following method is used to generate the true uncertainty data. Fig. 3 illustrates this procedure with some examples while the block diagram is shown in Fig.4. First, we compile the fixation data from all subjects in CRCNS dataset into a single map $\hat{\boldsymbol{F}}^{tr}$ of size M', N', and K being the height, width, and the total number of frames, respectively. We add 1 to $\hat{F}^{tr}[i, j, k]$ for every eye fixation that corresponds to pixel location $[i, j, k]$. Second, we resize the fixation map $\hat{\boldsymbol{F}}^{tr}$ to M, N and K; the respective height, width, and depth of the saliency map $\boldsymbol{S}$ from a saliency detection algorithm. This resizing is necessary because many saliency detection techniques work on downsampled video frames for computational efficiency. However, for the binary map $\hat{\boldsymbol{F}}^{tr}$, the resizing is not exactly a downsampling procedure.Denoted as $\boldsymbol{F}^{tr}$, the resized binary fixation map is obtained as follows

$$F^{tr}[m, n, k] = \sum_{\forall (i,j) \in \Phi[m,n,k]} \hat{F}^{tr}[i, j, k], \tag{9}$$

where $\Phi[m, n, k]$ is an indexing function that points to the set of pixels in $\hat{F}^{tr}$ that corresponds to pixel $[m, n, k]$ in F^{tr} map. Here, we use the sum of eye-fixation points from all subjects so that salient locations agreed upon by majority of subjects have the highest saliency, but at the same time sparse “1”s in the original fixation truth data are not lost. Finally, assuming that the saliency map $\boldsymbol{S}$ is normalized, we normalize $\boldsymbol{F}^{tr}$ and calculate the true uncertainty as

$$\boldsymbol{U}^{tr} = \big|\boldsymbol{S} - \boldsymbol{F}^{tr}\big|. \tag{10}$$

Obviously, $\boldsymbol{U}^{tr}$ shows how far each saliency estimate is from the recorded fixations. Thus, it can serve as a measure of the estimation uncertainty. Even though the individual eye-fixation

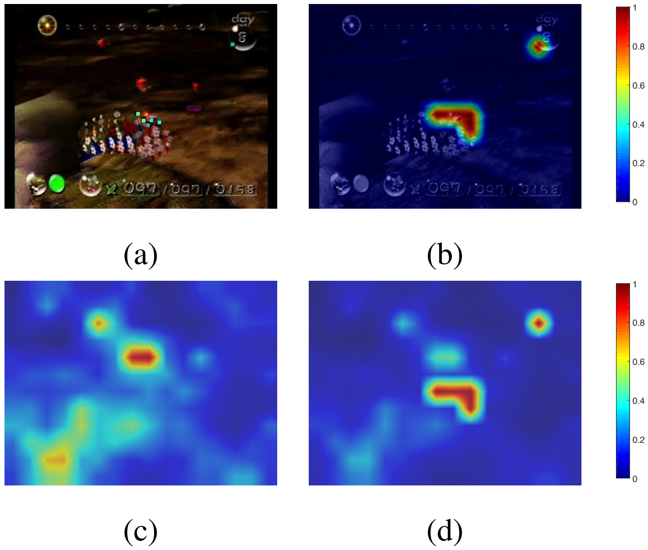

Fig. 3: Examples illustrating true uncertainty data. (a) Original video frame with eye fixation superimposed (small color squares in the center and top-right corner); (b) Resized eye fixation map superimposed on the original frame; (c) Saliency detection results; (d) True uncertainty. We note that the color display is only for a better illustration, which involves some interpolation causing the discrete resized fixation map to appear continuous.

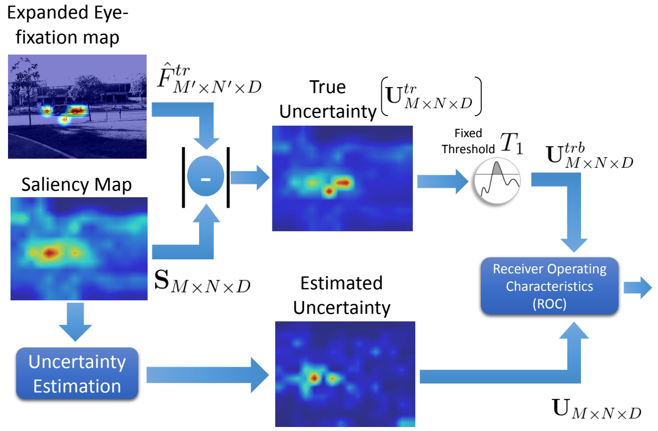
Expanded Eye-fixation map
$\hat{F}^{tr}_{M'\times N'\times D}$
True Uncertainty
$\mathbf{U}^{tr}_{M\times N\times D}$
-
Fixed Threshold T_1
$\mathbf{U}^{trb}_{M\times N\times D}$
Saliency Map
$\mathbf{S}_{M\times N\times D}$
Estimated Uncertainty
Receiver Operating Characteristics (ROC)
Uncertainty Estimation
$\mathbf{U}_{M\times N\times D}$

Fig. 4: Evaluation methodology [13].

data is binary, the aggregated fixation maps $\hat{\boldsymbol{F}}^{tr}$, $\boldsymbol{F}^{tr}$, the derived true uncertainty data $\boldsymbol{U}^{tr}$, and the saliency detection results $\boldsymbol{S}$ are continuous values.

2) Performance Measurement: With the true uncertainty data available, we use a detection theory-based scheme for the performance evaluation [13]. The scheme generates an ROC curve and uses AUC as the performance metric [32]. Since our true uncertainty data $\boldsymbol{U}^{tr}$ is continuous, it needs to be converted to binary data, denoted as $\boldsymbol{U}^{trb}$, as the ROC curve is intended for binary classifiers. This conversion is conducted by applying a threshold T_1. To generate the ROC curve, the uncertainty estimates $\boldsymbol{U}$ are also thresholded by T_2 into a binary form, $\boldsymbol{U}^b$, and compared against $\boldsymbol{U}^{trb}$. 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 ranges between 0 and 1, with a greater value indicating better performance, and 0.5 indicating a performance equivalent to random classifier.

V. Experiments

We conducted three sets of experiments to study several aspects of the proposed algorithm. In the first set, we compare the relative performance based on the neighborhood selection. We evaluate and compare the performance of the proposed algorithm using:

- Spatiotemporal neighborhood as described in III, labeled Spatiotemporal Uncertainty (*STU*)
- Temporal neighborhood as described in III-A, labeled Temporal Uncertainty (*TU*) [13]
- Spatial neighborhood as described in III-B, labeled Spatial Uncertainty (*SU*) [14]
- Naive fusion of Spatial and Temporal Uncertainty (*SU+TU*), a pixel-wise addition of *TU* and *SU* maps
- Entropy-based Uncertainty (*EU*) [12]
- Local variance of spatiotemporal neighborhood, labeled *Baseline*

The performance of these algorithms is quantified in terms of Area-Under-the-Curve (AUC) values of their corresponding Receiver-Operating-Characteristic (ROC) curves. We, also, show effects of saliency map scale as well as kernel size on the proposed algorithm's performance. Details on data and experiments procedure are provided in the dataset section and the performance evaluation methodology section, respectively. The second set of experiments are designed to show performance of the proposed uncertainty estimation algorithm given different categories of videos. Also, we show the distinct effects of kernel size on the proposed algorithm performance given radically different video contents. The third set of experiments verifies the performance of the proposed algorithms using additional datasets and saliency detection models.

A. Datasets

We tested the proposed unsupervised uncertainty estimation algorithm using three publicly available databases: CRCNS [15], DIEM [33], and AVD [34]. The CRCNS [15] 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 diverse with a total of 12 categories ranging from street scenes to video games and from TV sports to TV news. In many cases the videos contain variations of lighting conditions, severe camera movements, and high motion blur effects. Eye fixation data are provided with each video, recorded for a group of eight human subjects watching the videos under task-free view condition. The DIEM [33] database includes 85 videos, with varying resolutions and duration up to 130 seconds with 30 frames per second. The videos content are mainly limited to TV and film content including film trailers, music videos, and advertisement. The eye fixation data are collected from 250 participants under task-free view conditions. The AVD [34] database includes 148 videos, with varying resolutions and mean duration of 22 seconds with 30 frames per second. The video contents are limited to moving objects, landscape, and faces. The eye fixations data are collected from 176 observers. The AVD dataset contains two sets of videos of the same visual content but one with audio and the other without. According to their findings on the effect of audio on the attention of the participants, we only select the videos without associated audio.

For our experiments, we generated saliency maps for the videos using a recent algorithm based on 3D FFT local spectra (3DFFT) [35]. However, for validation, we also share the results from two additional saliency models: STSR [36] and PQFT [37], which are shown at the end of this section. Unless stated otherwise, saliency maps used in all experiments are generated using 3DFFT. In most of our experiments, the saliency maps are reduced in size to three different scales. *Scale 1* is of size 12×16; a downscale of frames original size 480×640, where every 40×40 region in the original frame corresponds to a single pixel in *Scale 1*. Similarly, *Scale 2* saliency maps are 24×32, where every pixel is equivalent to 20×20 region of pixels in the original sized frame, and *Scale 3* saliency maps are 48×64, where every pixel is equivalent to 10×10 regions.

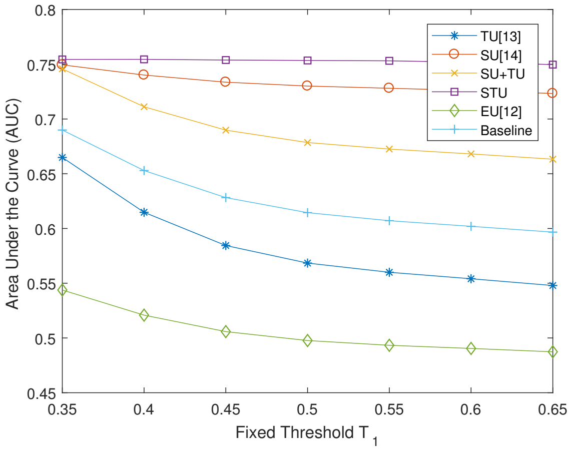

Fig. 5: Examples illustrating that relative uncertainty estimation performance is independent of fixed threshold T_1 applied to true uncertainty. Results reported here were generated using *Scale 1* maps with averaging kernel of length 5 for TU, of size 5×5 for SU, and $5 \times 5 \times 5$ for STU.

B. Results and Discussions

The performance evaluation procedure described earlier utilizes a fixed threshold T_1 to transform the continuous valued true uncertainty $\boldsymbol{U}^{tr}$ to binary ground truth. First, we examine the impact of changing the value of T_1. The algorithms under consideration are: Temporal Uncertainty (*TU*), Spatial Uncertainty (*SU*), Fused Spatial and Temporal Uncertainty (*SU+TU*), Spatiotemporal Uncertainty (*STU*), Spatiotemporal local variance (*Baseline*) computed on the same neighborhood as STU, and Entropy-based Uncertainty (*EU*). Fig.5 shows the performance of these algorithms in terms of AUC versus T_1. As shown in Fig.5, T_1 directly affects AUC value; as the value of T_1 increases, the AUC value of all algorithms considered here decreases. It is also interesting to point out that the gradient of AUC levels-off as T_1 reaches higher values. Although we can see that T_1 value significantly changes AUC, conclusions based on relative AUC values are consistent regardless of the value of T_1. As shown in Fig.5, STU outperforms all other algorithms while EU is performing the worst in this experiment. Please note that the reported AUC results are for *Scale 1* maps with averaging kernel of length 5 for TU, of size 5×5 for SU, and $5 \times 5 \times 5$ for STU.

1) Neighborhood Selection (domain,scale,size): As shown earlier, in addition to the threshold T_1, the neighborhood selection affects AUC value. Additionally, scale of the saliency maps and

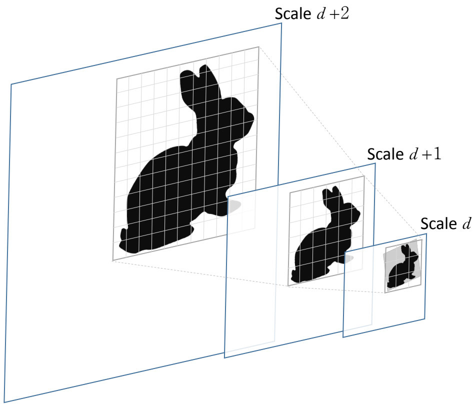

Fig. 6: Kernel size changes between scales according to support region size.

size of the processing kernels affect the performance of proposed estimation algorithm as well. In Fig. 8, we show the AUC values for the algorithms under test using different saliency map scales. The experiment is conducted using saliency maps of *scale 1*, *2* and *3* and an averaging kernel. In order to fix the kernel size relative to the support region size in the original frame, we use different kernel size for each scale, as illustrated in Fig. 6. In Fig. 8, *Scale 1* experiment uses 5×5 for SU and $5 \times 5 \times 5$ for STU. Similarly, for *Scale 2*: 11×11 for SU and $11 \times 11 \times 5$ for STU, and for *Scale 3*: 21×21 for SU and $21 \times 21 \times 5$ for STU. The length of TU kernel is fixed $L_t = 5$. We can see that the change in AUC value is relatively small, thus, shows the effectiveness of the proposed uncertainty algorithm even when saliency maps are considerably small size. This feature of the proposed estimation algorithm can be exploited to reduce the required computations, thus speeding up the estimation process without much sacrifice in terms of performance. Please note that AUC value for EU algorithm changes over different scales, due to true uncertainty $\boldsymbol{U}^{tr}$ containing more details as the scale increases. Moreover, kernel size affects the performance of the proposed algorithm as well. Fig. 7 shows the performance of the estimation algorithms under test, in terms of AUC values, when the estimation kernel size is changed. The experiment is conducted using *scale 2* saliency map and variable kernel size r (r for *TU*, $r \times r$ for *SU*, and $r \times r \times r$ for *STU*). As shown in Fig. 7, AUC of the proposed algorithm changes as the size of the kernel changes. However, the change in *TU* performance is significantly smaller than that of *SU* and *STU* because the number of pixels added into *SU* and *STU* kernels is significantly more than the number of pixels added to *TU* kernel. There

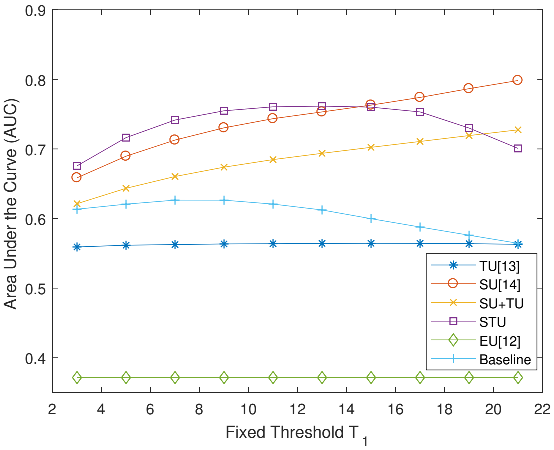

Fig. 7: AUC value is affected by the choice of the kernel size at the same scale. Results reported here use *Scale 2* saliency maps and $T_1 = 0.55$.

is, however, a slight degradation in *TU* performance as the kernel size increases (starting from $L_t = 13$ onwards), which can be attributed to including less relevant pixel in the estimation process as the kernel size increase. For kernels of sizes $3 \times 3 \times 3$ till $11 \times 11 \times 11$, it can be seen that *STU* achieves higher AUC than *SU*. However, such trend inverts starting from kernel size $13 \times 13 \times 13$ onwards. This could be explained by noting the similar trend in *TU* as the kernel size increases in time domain due to inclusion of pixels that might be less relevant. The performance degradation in *STU* (and *Baseline* as well) is more profound than *TU* because, for a kernel size of $n \times n \times n$, n^2 pixels are added to *STU* estimation process for every additional frame while only a single pixel is added for *TU* estimation. It is important here to clarify that these results are obtained for the whole dataset (50 videos). Thus, trends that are observed here are not necessarily true for every video type. We discuss in details the performance as related to the video categories in the next section.

2) Video Categories (domain,size): Given the diverse nature of scenes and dynamics in the dataset, we evaluate the performance of our proposed algorithm for each category in the dataset. For these experiments, we set $T_1 = 0.55$ and use *Scale 1* saliency maps. Table I shows AUC values for *TU* ($L_t = 5$), *SU* ($L_{s_x} = 5$), *ST+SU*, *STU* ($L_{st_x} = 5$), *EU*, and *Baseline* ($L_{st_x} = 5$), for each category, separately. As shown in Table I, AUC values for the proposed algorithm are above 0.5, indicating that the proposed algorithm is advantageous over random guessing.

Additionally, the algorithm performs better than *EU* in every category and in some by a wide margin. One interesting result is that AUC for Saccadetest video is significantly higher than other categories for all algorithms considered here. This can be attributed to its non-complex structure, which shows a disk moving against a light textured background. Notably, *STU* achieves highest performance in every category except Saccadetest. This could be attributed to its relative constant scenes in the first segment of the video.

Moreover, we explore the effect of kernel size on the estimation performance. In these experiments, we focus on *STU*, however, *TU*, *SU*, and *SU+TU* exhibit similar behavior. Fig.9 shows AUC for *STU* estimation algorithm on three video categories; saccadetest, tv-talk, and gamecube for kernel sizes: L_{st_x} = 3, 7, 11, and 15, using *Scale 2* saliency maps and T_1 = 0.55. As shown in Fig.9, as the kernel size increases, *STU* performance on saccadetest degrades indicating that the relevance saliency context in saccadetest video is strictly local and including more pixels than direct neighbors degrades uncertainty estimation performance. Indeed, the structure of saccadetest video justifies these results due to its simplicity. In contrast, gamecube video uncertainty estimation results increase as the kernel size increase. This indicates that the set of correlated saliency pixels for gamecube is larger than its direct neighbors. The large set of correlated saliency pixels in gamecube might be explained by its complex structure and the fact that these videos contain multiple salient actors in the same scene making it more difficult to capture saliency context from small local neighborhoods. On the other hand, *STU* performance in estimating uncertainty for tv-talk reaches maximum level in intermediate kernel sizes and then decreases as we increase the kernel size, indicating that the most appropriate kernel size to capture relevant saliency context is half the frame size.

3) Comparison across various datasets and saliency models: In this section, we present evaluation results for the proposed algorithm across various datasets. We compare the performance of the proposed algorithm using videos from three datasets: CRCNS [15], DIEM [33], and AVD [34]. Fig. 10 shows the the AUC values of the five uncertainty estimation methods using videos from the three datasets. In Fig. 10, *STU* performance is the highest among all datasets. In general, the trend and ranking between the uncertainty estimation methods is consistent across the three datasets.

Additionally, we present the evaluation results for the proposed algorithm across using three saliency models: 3DFFT [35], STSR [36], and PQFT [37]. Fig. 11 shows the AUC values of the five uncertainty estimation algorithms. In Fig. 11, a consistent trend and ranking between the five algorithms exist across all three saliency models, where *STU* achieves the highest AUC value.

Moreover, we evaluate the proposed algorithm, in terms of the computed uncertainty map distribution versus uncertainty ground truth maps distribution, using four distribution-based metrics; Jeffrey Divergence (JD), Jensen-Shannon divergence (JS), Histogram Intersection (HI), L2-norm. As shown in Table.II, the proposed algorithm provides the closest distribution to that of the ground truth maps across all four metrics and all datasets.

VI. Conclusion

In this paper, we discussed the problem of quantifying uncertainty for video saliency detection. To solve this problem, we presented an algorithm to estimate pixel-wise uncertainty in computational saliency maps, which relies on a common feature of human fixation. Our experiments, using CRCNS dataset, showed a reduction of roughly 50%, across all videos, in pixel entropy when conditioned on its local neighbors' average. The experiment shows that local correlation exists in saliency perceived by HVS. Thus, saliency map's pixels ought to be highly correlated with their local neighbors. Using this result, we formulated the proposed algorithm according to temporal, spatial, and spatiotemporal neighborhoods and studied the effect of neighborhood selection on the algorithm performance. Additionally, we showed that the appropriate size of local neighborhood is mainly determined by the video content and makes a significant impact on the algorithm performance. For performance evaluation, we proposed a systematic performance evaluation scheme including the generation of true uncertainty and ROC curve-based objective assessment. The proposed algorithm outperforms state-of-the-art uncertainty estimation algorithms across three different datasets: CRCNS, DIEM, and AVD. Consistent performance has been observed with different saliency models. Our algorithm is unsupervised and computationally highly efficient. Additionally, the performance of proposed algorithm could be further enhanced by using a weight-average combination of uncertainty maps from different scales depending on the video content, which we did not explore in this study. The proposed algorithm can be very useful, either as a stand-alone objective evaluation method for saliency detection algorithms, or as an effective means of quality control for saliency-based video processing applications.

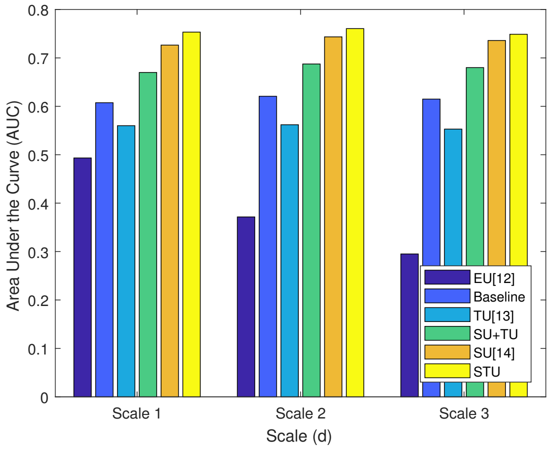

Fig. 8: The impact of scale change with constant support region (using different kernel sizes). AUC value is relatively the same when processing different scales. Results reported here use threshold T_1 = 0.55.

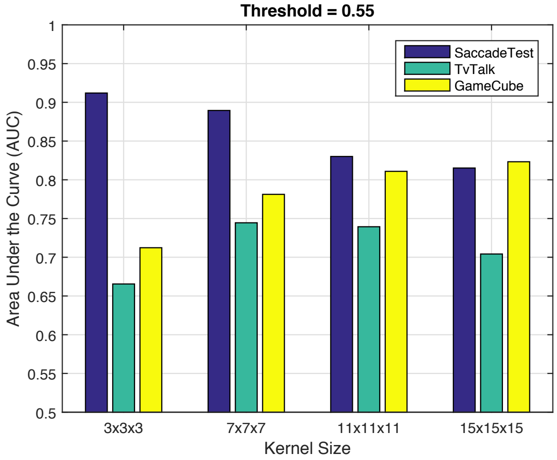

Fig. 9: Examples illustrating the effect of kernel size on the estimation performance using *STU* extracted from uncertainty maps of *Scale 2*.

References

[1] Z. Ren, S. Gao, L.-T. Chia, and I.-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.

[2] R. Bharath, L. Nicholas, and X. Cheng, "Scalable scene understanding using saliency-guided object localization," in *Control and Automation (ICCA), 2013 10th IEEE International Conference on*, June 2013, pp. 1503–1508.

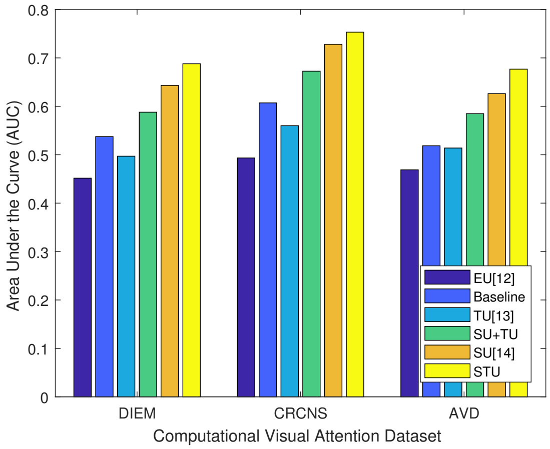

Fig. 10: The performance of the proposed algorithm across the datasets CRCNS [15], DIEM [33], and AVD [34].

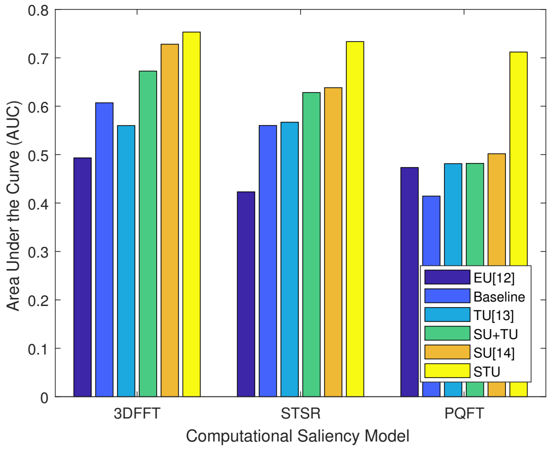

Fig. 11: The performance of the proposed algorithm using the saliency models 3DFFT [35], STSR [36], PQFT [37]

[3] Y. Gitman, M. Erofeev, D. Vatolin, B. Andrey, and F. 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.

[4] J. Peng and Q. Xiao-Lin, “Keyframe-based video summary using visual attention clues,” *IEEE MultiMedia*, no. 2, pp. 64–73, 2009.

[5] L. 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.

[6] W. Wang, J. Shen, and L. Shao, “Consistent video saliency using local gradient flow optimization and global refinement,”

TABLE I: List of AUC value for different categories using fixed threshold $T_1 = 0.55$ and *Scale 1* saliency maps. Note that the highest AUC value in each category is labeled in green and lowest AUC value in red. Also, the category with the highest AUC in the dataset is shown in **bold**

	TU [13]	*SU* [14]	*SU+TU*	*STU*	*EU* [12]	*Baseline*
beverly	0.5793	0.8088	0.7174	0.8130	0.5835	0.6915
gamecube	0.5987	0.7636	0.7155	0.7913	0.5906	0.6834
monica	0.6152	0.7801	0.7240	0.7994	0.5728	0.6506
saccadetest	**0.7722**	**0.8734**	**0.8216**	0.8587	**0.8458**	**0.8308**
standard	0.5866	0.7190	0.6609	0.7462	0.5165	0.5841
tv-action	0.7481	0.8466	0.7970	**0.8667**	0.7245	0.6491
tv-ads	0.5565	0.7248	0.6565	0.7476	0.5228	0.5360
tv-announce	0.4555	0.6679	0.5550	0.7321	0.4434	0.5818
tv-music	0.5548	0.6721	0.6236	0.7427	0.4471	0.5771
tv-news	0.5051	0.6497	0.5885	0.6947	0.4861	0.5029
tv-sports	0.5156	0.6746	0.6170	0.7172	0.5020	0.5368
tv-talk	0.5692	0.7142	0.6393	0.7364	0.5299	0.5250

TABLE II: Estimated distances using distribution-based metrics for the proposed algorithm in comparison with the state-of-the-art algorithms using *Scale 1* saliency maps computed using 3DFFT algorithm [35]. Note that the highest value in each distance metric is labeled in green and the lowest value in red. Also, the distance values for the proposed algorithm is shown in **bold**

	CRCNS				DIEM				AVD			
Algorithms	JS	JD	HI	L2	JS	JD	HI	L2	JS	JD	HI	L2
TU [13]	0.34	0.68	0.64	0.25	0.16	0.32	0.30	0.12	0.33	0.66	0.60	0.24
SU [14]	0.14	0.29	0.32	0.12	0.05	0.09	0.13	0.04	0.09	0.18	0.24	0.08
SU+TU	0.14	0.27	0.32	0.12	0.05	0.11	0.14	0.05	0.10	0.20	0.28	0.10
STU	**0.08**	**0.15**	**0.24**	**0.08**	**0.03**	**0.05**	**0.10**	**0.04**	**0.06**	**0.12**	**0.21**	**0.08**
EU [12]	0.51	1.02	0.82	0.46	0.23	0.45	0.36	0.19	0.45	0.91	0.71	0.41
Baseline	0.38	0.76	0.68	0.32	0.15	0.30	0.29	0.13	0.29	0.58	0.57	0.25

IEEE Transactions on Image Processing, vol. 24, no. 11, pp. 4185–4196, Nov 2015.

[7] H. Kim, Y. Kim, J. Y. Sim, and C. S. Kim, “Spatiotemporal saliency detection for video sequences based on random walk with restart,” *IEEE Transactions on Image Processing*, vol. 24, no. 8, pp. 2552–2564, Aug 2015.

[8] C. R. Huang, Y. J. Chang, Z. X. Yang, and Y. Y. Lin, “Video saliency map detection by dominant camera motion removal,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 24, no. 8, pp. 1336–1349, Aug 2014.

[9] D. Mahapatra, S. O. Gilani, and M. K. Saini, “Coherency based spatio-temporal saliency detection for video object segmentation,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 8, no. 3, pp. 454–462, June 2014.

[10] Z. Liu, X. Zhang, S. Luo, and O. L. Meur, “Superpixel-based spatiotemporal saliency detection,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 24, no. 9, pp. 1522–1540, Sept 2014.

[11] M. Liang and X. Hu, “Predicting eye fixations with higher-level visual features,” *IEEE Transactions on Image Processing*, vol. 24, no. 3, pp. 1178–1189, March 2015.

[12] Y. Fang, Z. Wang, W. Lin, and Z. Fang, “Video saliency incorporating spatiotemporal cues and uncertainty weighting,” *IEEE Transactions on Image Processing*, vol. 23, no. 9, pp. 3910–3921, Sept 2014.

[13] T. Alshawi, Z. Long, and G. 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*. IEEE, 2015.

[14] ——, “Unsupervised uncertainty analysis for video saliency detection,” in *the 49th Asilomar Conference on Signals, Systems and Computers, Pacific Grove, CA, Nov. 8-11*. IEEE, 2015.

[15] L. Itti, “Eye-tracking data from human volunteers watching complex video stimuli.” [Online]. Available: https://crcns.org/data-sets/eye/eye-1

[16] T. Liu, Z. Yuan, J. Sun, J. Wang, N. Zheng, X. Tang, and H.-Y. Shum, “Learning to detect a salient object,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 33, no. 2, pp. 353–367, Feb 2011.

[17] T. Judd, K. Ehinger, F. Durand, and A. Torralba, “Learning to predict where humans look,” in *Computer Vision, 2009 IEEE 12th International Conference on*, Sept 2009, pp. 2106–2113.

[18] Y. Gu, Z. Jin, and S. C. Chiu, “Active learning combining uncertainty and diversity for multi-class image classification,” *IET Computer Vision*, vol. 9, no. 3, pp. 400–407, 2015.

[19] G. Saygili, M. Staring, and E. A. Hendriks, “Confidence estimation for medical image registration based on stereo confidences,” *IEEE Transactions on Medical Imaging*, vol. 35, no. 2, pp. 539–549, Feb 2016.

[20] X. Hu and P. Mordohai, “A quantitative evaluation of confidence measures for stereo vision,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 11, pp. 2121–2133, Nov 2012.

[21] R. Haeusler, R. Nair, and D. Kondermann, “Ensemble learning for confidence measures in stereo vision,” in *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, June 2013, pp. 305–312.

[22] H. Feldman and K. Friston, “Attention, uncertainty, and free-energy,” *Frontiers in human neuroscience*, vol. 4, p. 215, 2010.

[23] N. Bruce and J. Tsotsos, “Saliency based on information maximization.”

[24] N. D. Bruce and J. K. Tsotsos, “Saliency, attention, and visual search: An information theoretic approach,” *Journal of vision*, vol. 9, no. 3, pp. 5–5, 2009.

[25] L. Wei and D. Luo, “A biologically inspired spatiotemporal saliency attention model based on entropy value,” *Optik-International Journal for Light and Electron Optics*, vol. 125, no. 21, pp. 6422–6427, 2014.

[26] W. Wang, Y. Wang, Q. Huang, and W. Gao, “Measuring visual saliency by site entropy rate,” in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 2368–2375.

[27] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, “Frequency-tuned salient region detection,” in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, June 2009, pp. 1597–1604.

[28] P. Sharma, F. Cheikh, and J. Hardeberg, “Spatio-temporal analysis of eye fixations data in images,” in *Image Processing (ICIP), 2014 IEEE International Conference on*, Oct 2014, pp. 1150–1154.

[29] A. Borji, H. Tavakoli, D. Sihite, and L. Itti, “Analysis of scores, datasets, and models in visual saliency prediction,” in *Computer Vision (ICCV), 2013 IEEE International Conference on*, Dec 2013, pp. 921–928.

[30] S. R., H. Katti, N. Sebe, M. Kankanhalli, and T. Chua, “An eye fixation database for saliency detection in images,” in *European Conference on Computer Vision (ECCV), 2010*, 2010.

[31] T. Alshawi, Z. Long, and G. AlRegib, “Understanding spatial correlation in eye-fixation maps for visual attention in videos,” in *submitted to IEEE International Conference on Multimedia and Expo (ICME 2016), Seattle, WA, Jul. 11-15*. IEEE, 2016.

[32] T. Hastie, R. Tibshirani, and J. Friedman, *The elements of statistical learning: data mining, inference and prediction*. Springer, 2005.

[33] P. K. Mital, T. J. Smith, R. Hill, and J. M. Henderson, “Clustering of gaze during dynamic scene viewing is predicted by motion,” *Cognitive Computation*, vol. 3, no. 1, pp. 5–24, march 2011.

[34] P. Marighetto, A. Coutrot, N. Riche, N. Guyader, M. Mancas, B. Gosselin, and R. Laganiere, “Audio-visual attention: Eye-tracking dataset and analysis toolbox,” in *Image Processing (ICIP), 2017 IEEE International Conference on*, September 2017, pp. 1802–1806.

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

[36] H. J. Seo and P. Milanfar, “Static and space-time visual saliency detection by self-resemblance,” *Journal of Vision*, vol. 9, no. 12, p. 15, 2009. [Online]. Available: + http://dx.doi.org/10.1167/9.12.15

[37] C. Guo, Q. Ma, and L. Zhang, “Spatio-temporal saliency detection using phase spectrum of quaternion fourier transform,” in *2008 IEEE Conference on Computer Vision and Pattern Recognition*, June 2008, pp. 1–8.

[38] T. Cover and J. Thomas, *Elements of Information Theory*. New York, NY, USA: Wiley-Interscience, 1991.