UNDERSTANDING SPATIAL CORRELATION IN EYE-FIXATION MAPS FOR VISUAL ATTENTION IN VIDEOS

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

In this paper, we present an analysis of recorded eye-fixation data from human subjects viewing video sequences. The purpose is to better understand visual attention for videos. Utilizing the eye-fixation data provided in the CRCNS (Collaborative Research in Computational Neuroscience) dataset, this paper focuses on the relation between the saliency of a pixel and that of its direct neighbors, without making any assumption about the structure of the eye-fixation maps. By employing some basic concepts from information theory, the analysis shows substantial correlation between the saliency of a pixel and the saliency of its neighborhood. The analysis also provides insights into the structure and dynamics of the eye-fixation maps, which can be very useful in understanding video saliency and its applications.

Index Terms— saliency detection, video, multimedia understanding, spatial correlation, computational perception

1. INTRODUCTION

Human visual attention modeling and understanding has been shown to be effective in analyzing big visual data as well as in improving the computation efficiency of visual data processing. Numerous applications have been proposed and currently investigated, such as object detection and recognition [2], scene understanding [3], and multimedia summarization [4].

To understand the visual attention mechanism, research usually relies on eye-tracking data analysis to formulate eye fixation maps. Such maps capture the focus of human subjects watching the 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 for unsupervised methods. However, there has been limited research in the video processing community on analyzing these eye-fixation maps separately from saliency models. By studying the eye-fixation maps, we hope to better understand the spatial correlation in video scenes, and henceforth to better understand visual attention mechanisms.

The authors in [5] analyzed eye-fixation data of images given location and time sequence of human subjects gaze, using the Eigen value decomposition of the correlation matrix constructed based on eye fixation data of different subjects. Their work shows that the first Eigen vector is responsible for roughly 21% of the data, and it correlates well with salient locations in the images dataset. In [6], the authors found 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 [7] containing five categories over a total of 409 images.

In this paper, we analyze eye-fixation maps associated with spatiotemporal visual cues from the CRCNS [8] dataset. In particular, we focus on spatial correlation in saliency. We investigate the possibility of predicting a pixel's saliency, as indicated by eye fixations, given the average saliency of its neighborhood. Since spatial correlation in the context of scene understanding refers to the relative locations of objects in the space, for videos it can be examined as the correlation between a pixel and its immediate spatiotemporal neighbors, as we will present in this paper. The contribution of this paper is two fold. First, we analyze the eye-fixation maps for videos, which has not been done in the literature, to our best knowledge. Second, contrary to other studies that foucs on general statstics of the eye-fixation map as a whole, we focus on the dynamics of a pixel with respect to its neighborhood in the eye-fixation map. Our research provides an alternative quantitative approach to describing human attention, which can be very useful for various saliency-based applications.

2. MODELS

2.1. Overall Map

Given an eye-fixation map F of size $M \times N$ and depth K frames, first of all, we consider every map pixel $x[m, n, k] \in F$ to be an instance of a discrete integer random variable X

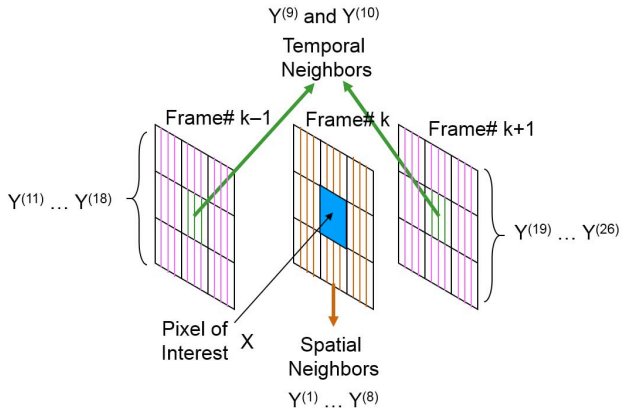


Fig. 1. Illustration of neighborhood pixels grouping. Spatial neighbors $Y^{(1)}...Y^{(8)}$ of pixel X are hashed in orange color, temporal neighbors $Y^{(9)}$ and $Y^{(10)}$ of X are shown in green color, and the rest of spatiotemporal neighbors $Y^{(11)}...Y^{(26)}$ are shown in purple.

with an unknown distribution, that is:

$$X : \Omega \rightarrow E, \tag{1}$$

where Ω is the set of all possible outcomes that describes the eye-fixation events during the experiments, E is the observed set, and $x[m, n, k] \in \{0, 1, 2, ..., L\}$ numerically represents these events using $L + 1$ symbols. For such a eye-fixation map F, we compute the Shannon entropy of X as follows:

$$H(X) = E[I(X)] = -\sum_{i=0}^{L} P(x_i) \log_2 P(x_i), \tag{2}$$

where $E[\cdot]$ is the operator for expectation, $I(X)$ is the self-information of X, and $P(x_i)$ is the probability mass function for X.

2.2. All Neighbors

For each map pixel, it is of interest to examine the relationship between the pixel and its neighbors. There are altogether 26 direct spatiotemporal neighbors (i.e., 9 pixels from frame $k-1$, 9 pixels from frame $k+1$, and 8 pixels from the current frame k), as shown in Fig.1. To distinguish these neighbors from the center pixel, we label them as $Y^{(j)}$, where $j \in \{1, 2, ..., 26\}$. We compute the conditional entropy of a center pixel X given the average of its direct neighbors as follows:

$$H(X|Z) = \sum_{\forall x_i, z_j} P(x_i, z_j) \log_2 \frac{P(z_i)}{P(x_i, z_j)} \tag{3}$$

where $Z = f\left(Y^{(1)}, \ldots, Y^{(26)}\right)$ is the arithmetic mean of the 26 direct neighbors, and $P(x_i, z_j)$ is the joint probability mass function for X (the center pixel) and Z (the mean of its direct neighbors). As a basic property of conditional entropy, the following relationship always holds:

$$0 \leq H(X|Z) \leq H(X). \tag{4}$$

Here, if Z completely determines X, then $H(X|Z) = 0$; or, if X is completely independent of Z, then $H(X|Z) = H(X)$, which means knowing Z does not reduce the uncertainty about X.

2.3. Spatial Neighbors

In addition to analyzing the correlation between a pixel and its neighbors in the general sense, we investigate the effect of the video content on such correlation. To do this, we need to extend the model introduced above. Now we consider all pixels at location $[m, n]$ in the eye-fixation map across all K frames as instances of a random variable $X[m, n]$. Similarly, we calculate the arithmetic mean, denoted as $Q[m, n]$, of the eight direct spatial neighbors $X[m + i, n + j]$, where i and $j \in \{1, 0, -1\}$, as shown in Fig.1. To quantify their correlation, we compute mutual information between $X[m, n]$ and $Q[m, n]$ as follows:

$$I(X[m,n]; Q[m,n]) = \sum_{\forall x_i, q_j} P(x_i, q_j) \log_2 \frac{P(x_i, q_j)}{P(x_i)P(q_j)}, \tag{5}$$

where $P(x_i)$ is the probability mass function of random variable $X[m, n]$, $P(q_j)$ is the probability mass function of $Q[m, n]$, the arithmetic mean of the spatial neighbors, and $P(x_i, q_j)$ is the joint probability mass function of $X[m, n]$ and $Q[m, n]$.

2.4. Temporal Neighbors

Similar to the correlation with spatial neighbors, we analyze the correlation with temporal neighbors. For this purpose, we modify the model again as follows. First, we consider each pixel in frame k of an eye-fixation map, $F(k)$, as an instance of a random variable X_k. Then, we obtain another random variable W_k, which represents a pixel-wise arithmetic mean of adjacent frames $F(k + D)$. Here, $D \in \{\pm 1, \pm 2, \pm 3, \ldots\}$. Similar to Sec.2.3, we quantify the eye-fixation correlation with temporal neighbors by computing the mutual information between X_k and W_k as follows:

$$I(X_k; W_k) = \sum_{\forall x_i, w_j} P(x_i, w_j) \log_2 \frac{P(x_i, w_j)}{P(x_i)P(w_j)}, \tag{6}$$

where $P(x_i)$ is the probability mass function of X_k, $P(w_j)$ is the probability mass function of W_k, and $P(x_i, w_j)$ is the associated joint probability mass function.

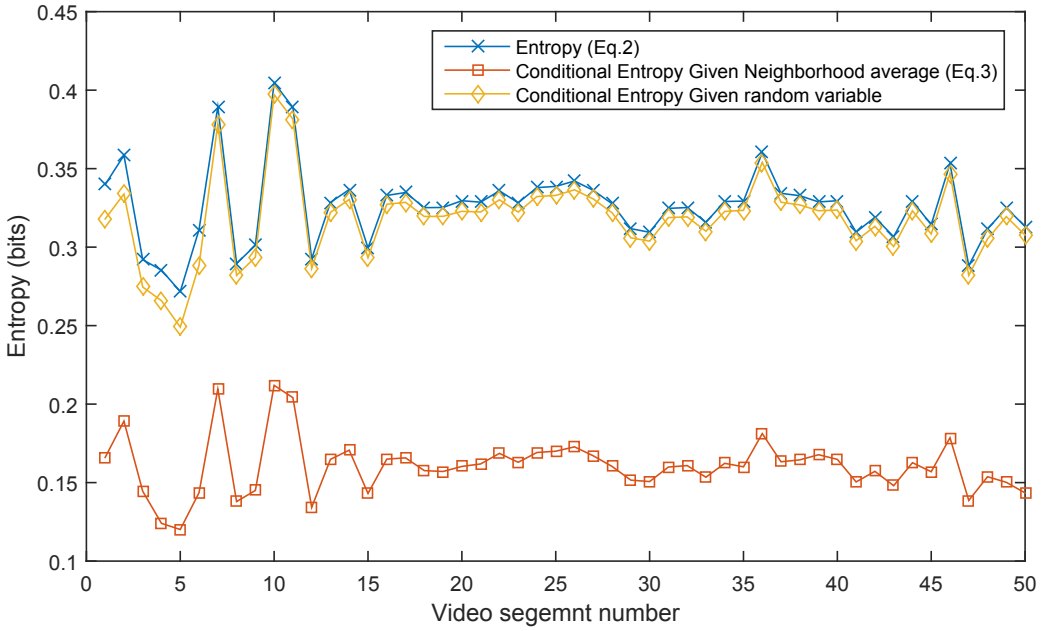


Fig. 2. Entropy reduction over all video segments.

3. EXPERIMENTS

3.1. Preparing Eye-fixation Data

Our study in this paper uses eye-fixation maps from the public CRCNS database [8]. The database includes 50 videos, with a resolution of 480×640, durations ranging between 5 and 90 seconds, and a frame rate of 30fps. The videos 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. The eye tracking data were collected from eight subjects using an ISCAN RK-464 eye-tracker at 240 Hz sampling rate, which was calibrated every five clips using 9-point calibration. The stimuli were displayed on 22" CRT monitor at $80cm$ viewing distance with mean screen luminance of 30 cd/m^2. Eye tracking data are provided for each human subject separately in a string of eye gaze coordinates, which span 0 to 639 in the horizontal direction and 0 to 479 in the vertical direction with location (0,0) being at the top left corner of the monitor. Labels are available for each eye-gaze sample, e.g., fixation, saccade, and during blink, just to name a few.

For a given video sequence, we prepare an eye fixation map from the eye-tracking data in the following manner:

1. Construct a frame of size 480×640 and initialize it to zeros.
2. Collect all eye-gaze samples corresponding to a given video frame. We only select samples that are fixation or smooth pursuit. Saccade or loss-of-tracking samples are not included.
3. For every sample obtained in step two, we set the pixel value at the corresponding location to one. If two samples coincide in spatial location, we set the pixel value equal to the number of samples pointing to that location.
4. Following the procedure above, we process the eye-tracking data frame by frame, and finally construct an eye-fixation map with the same size and number of frames as the video sequence.

Additionally, we construct the eye-fixation maps at various scales by reducing the original map size, which are useful in practical applications when video frames are often processed at reduced frame sizes. For an original map $F(m,n,k)$, the size-reduced map of scale s, $F^{(s)}(m,n,k)$, is formed as

$$F^{(s)}(m,n,k) = \sum_{\forall i,j \in R_s} F(i,j,k) \tag{7}$$

where R_s is the window that contains pixels of $F(m,n,k)$ corresponding to pixel (m,n,k) in $F^{(s)}(m,n,k)$.

3.2. Results and Discussions

3.2.1. Correlation with All Neighbors

In the following experiments, we set R_s to 40×40 window size. This helps reduce the computation time and still gener-

ates results similar to those obtained using the original map size. First, we evaluate the correlation between a map pixel and its direct spatiotemporal neighbors as detailed in Sec. 2. We plot the entropy values computed for each of the 50 video sequences in the dataset in Fig.2. As shown in the figure, the entropy of the eye-fixation drops when the spatiotemporal neighborhood average is considered, which is the red curve in the plot. To have a basis for comparison, we also calculate an entropy conditioned on a uniformly-distributed random variable and show the results as the yellow curve in the same figure.

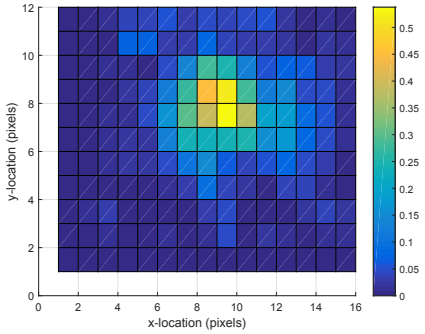
The reduction in the entropy values in many cases reaches 50% in the red curve. Such drop value is significant when compared to reduction in entropy due to conditioning on the uniformly-distributed random variable, indicating a strong and meaningful correlation. Another important observation from Fig.2 is that the entropy reduction is consistent across all videos in the dataset. The average entropy reduction is 0.0815 bits with variance 3.2416×10^{-05}.

It is worth noting that the entropy of the eye-fixation maps is generally low due to the sparsity of such maps. In fact, most of the map pixels have a value that equals to zero. However, since the probability mass is concentrated in a single symbol, the skewness of the probability mass function does not affect statistical test outcome. It merely moves the entropy (and the conditional entropy in turn) up or down. The three video segments with the highest entropy values are gamecube02, gamecube06, and gamecube13. These videos have relatively longer duration and engaging content, which potentially enable more cognitive processes to take place, thus contributing to the higher entropy.

3.2.2. *Correlation with Spatial Neighbors*

Second, we study the correlation between a map pixel in a given spatial location and its direct spatial neighborhood. The purpose is to evaluate the impact of the famous center-bias phenomenon [9] on the correlation. It should be noted here that the researchers collecting the database have every video segment preceded by a blinking cross in the middle of the screen, exactly at [239,319]. The blinking lasts for 1 sec before the video is shown. Consequently, *lack of knowledge* center-bias is present in almost all videos. However, it contributes mostly to the viewing of the first few frames, thus having not much effect on the overall correlation. The majority of the videos in this dataset have another center-bias factor that significantly influences the end results, i.e., the *photography* center-bias. This bias is due to the tendency of photographers to place the object(s) of interest at the center of the video frames. Even though many image datasets for visual attention research have taken this into account, it is difficult to do the same for videos. This photography center-bias is particularly obvious in the gamecube video segments, with a sample frame shown in Fig.3.(a), since the in-game camera

(a) gamecube06 sample frame taken at 01m:57s:076'



(b) gamecube06 mutual information given spatial location.

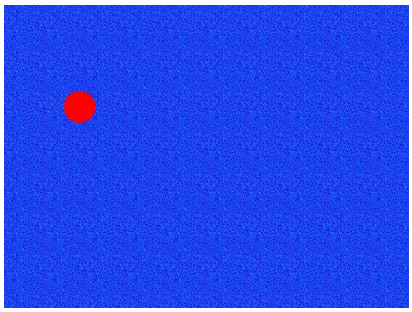
Fig. 3. gamecube06 sample frame alone with mutual information given spatial location, involving only spatial neighbors

system is designed to place the game character(s) in the center of the video.

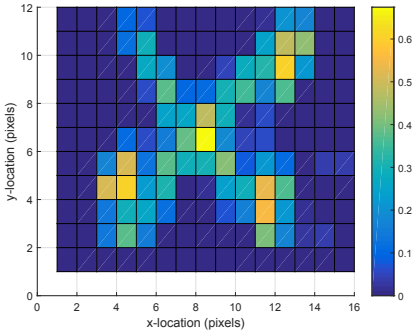
As described in Sec.2.3, we compute the mutual information between a pixel at a given location and its direct spatial neighbors. The results are shown in Fig.3.(b). We can clearly notice a high correlation in the case of gamecube06, as exists in virtually every gamecube video, which can be attributed to the photography center-bias. It is surprising to notice that even though textual information is located at the corners of the screen, it does not attract eye-fixation for prolonged periods due the relatively low information content they convey.

On the other hand, videos that lack photography center-bias exhibit totally different behaviour. For example, the saccadetest video, with a sample frame shown in Fig.4.(a), consists of a red dot against a blue textured background. The dot is static and changes places for the first half of the clip, then it moves diagonally with a consistent speed for the rest of the video. As shown in Fig.4.(b), the correlation is highest when there is a smooth pursuit following the red dot, due to high sampling rate and low spatial displacement in the object of interest. Similar trends can also be observed in video segments such as beverly06, beverly07, and beverly08.

Additionally, interesting trends can be observed when there are multiple salient objects present in the scene, such as in the tv-news03 video, a sample frame of which shown



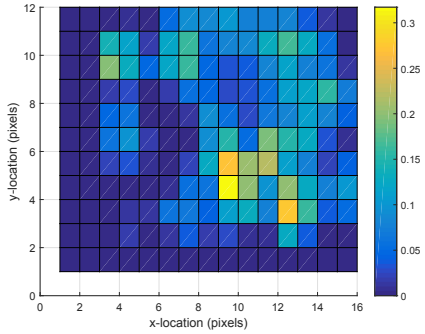
(a) saccadetest sample frame taken at 00m:07s:606'.



(b) saccadetest mutual information given spatial location.

Fig. 4. saccadetest sample frame alone with mutual information given spatial location

(a) tv-news03 sample frame taken at 06m:51s:026'.



(b) tv-news03 mutual information given spatial location.

Fig. 5. tv-news03 sample frame alone with mutual information given spatial location

in Fig.5.(a). The space-localized mutual information map, shown in Fig.5.(b), exhibits two centers of attention. One corresponds to the most semantically informative object in the scene, i.e., the news anchor's face. The other is the textual messages in the lower banner. Since the human subjects spend considerable periods of time looking at these two locations, the correlation is significantly higher than other locations in the eye-fixation map. The example results, especially those from the latter two without obvious center-bias, demonstrate that the higher correlation areas match very well with the human attention.

3.2.3. Correlation with Temporal Neighbors

Finally, we study the correlation over time as described in Sec.2.4. We begin by computing mutual information between a given frame $F(k)$ and the average of its temporal neighbors $F(k+D)$ and $F(k-D)$ with different values of distance D. As shown in Fig.6, mutual information between a frame and its direct neighbors (i.e., $D = 1$) is significant compared to the information shared with distant frames. For all videos, roughly 50% of information is shared between adjacent neighboring frames (recall that the average information content in an eye-fixation map is about 0.3 bits as shown in Fig.2). Another important realization is that this trend is seen in every category in the dataset, regardless of the content. However, the rate of change in some categories are small, such as saccadetest, and some are large, such as tv-ads and tv-sport. These differences can be explained by the level of complexity of these videos. Simple videos (e.g., saccadetest) have very few stimuli. Thus, human subjects tend to fixate on a single target, which results in more correlated frames in the eye-fixation map. On the other hand, complex videos (e.g., tv-ads and tv-sports) contain much more stimuli, requiring more efforts from the human subjects to examine and comprehend the scene.

Moreover, we compute the correlation between a given frame $F(k)$ and the average of its N direct neighbors up to a certain frame distance. As shown in Fig.7, mutual information between a given frame and its nearest neighbors contains most of the information shared with the rest of the frames. For most categories, the nearest $5-6$ neighbors contain almost all the correlated information in the eye-fixation map. Therefore, including more frames in the neighborhood average does not necessarily add any more useful information. This trend is observed in every category in the dataset, regardless of the content. However, some categories (such as monica and gamecube) yield higher mutual information than other categories, suggesting that the video content makes some differences. For most categories, the mutual information level-off after 8-10 frames, with the exception of standard. This can be attributed to the process of averaging that may cause some mutual information in the nearest neighbors to be marginal-

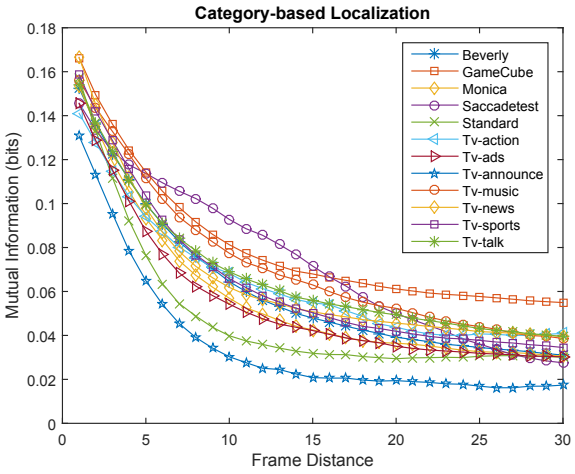


Fig. 6. Average mutual information between $F(k)$ and temporal neighbors $F(k+D)$ and $F(k-D)$, where D is the frame distance.

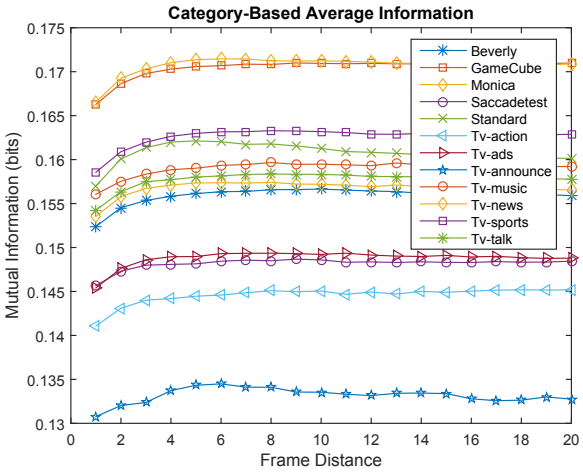


Fig. 7. Average mutual information between $F(k)$ and the average of its N temporal neighbors up to a certain frame distance.

ized as the number of frames included gets greater.

4. CONCLUSIONS

In this paper, we presented an information-theory-based analysis of recorded eye-fixation data from human subjects viewing video sequences to gain insights into visual attention mechanisms for videos. The analysis focused on the relationship between the saliency of an eye-fixation map pixel and that of its neighbors. Our experiments demonstrated that a substantial correlation between the saliency of a pixel and the saliency of its neighborhood exist. Such correlation is localized both spatially and temporally, and is significantly affected by the video's content and complexity. Our research provides an alternative quantitative approach to describing human attention. We believe such an approach is very important for many saliency applications. For example, ground truth data for saliency detection can be changed from the traditional eye-fixation data into a more descriptive format based on the correlations. The various correlations discussed in the paper can also be used as measures of the reliability of detected saliency, thus being a guide for optimizing saliency-based video processing.

5. REFERENCES

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

[2] Z. Ren, S. Gao, Chia L., and I. 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.

[3] R. Bharath, L.Z.J. 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.

[4] J. Peng and Q. Xiao-Lin, "Keyframe-based video summary using visual attention clues," *IEEE MultiMedia*, vol. 17, no. 2, pp. 64–73, 2010.

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

[6] A. Borji, H.R. Tavakoli, D.N. 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.

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

[8] 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.

[9] D. Gao, V. Mahadevan, and N. Vasconcelos, "On the plausibility of the discriminant center-surround hypothesis for visual saliency," *Journal of Vision*, vol. 8, no. 7, pp. 1–18, 2008.