An Integrated Approach To Visual Attention Modelling Using Spatial-Temporal Saliency And Objectness

Jean-Baptiste Weibel[1], Hui Li Tan[2], Shijian Lu[2]
weibel@acin.tuwien.ac.at, {hltan, slu}@i2r.a-star.edu.sg
[1] Image & Pervasive Access Lab (IPAL), Singapore
[2]Institute for Infocomm Research (I^{2}R), A*STAR, Singapore

***Abstract*—Visual attention modelling is an important research topic with a wide range of applications in visual tracking, perceptual quality assessment, re-targeting, video summarization, etc. In this paper, we propose a visual attention model that captures both bottom-up spatial-temporal saliency and top-down objectness. Leveraging on co-occurrence histograms, the proposed model captures a number of low-level cues including contrast, gradient, as well as, magnitude and gradient of optical flow. Additionally, the proposed model incorporates mid-level objectness cue which helps to boost the modelling performance greatly. The proposed model obtained superior AUC-ROCs when evaluated over the ASCMN dataset and the UCF Sports Action dataset.**

I. Introduction

The human visual system has a remarkable ability to automatically attend to regions-of-interest in a visual environment. This ability enables efficient allocation of limited perceptual and cognitive resources on salient objects in cluttered visual scenes, and is an evolutionary adaptation to allow organisms to quickly detect predators, preys, or mates in their visual environments. Visual attention comprises a bottom-up process and a top-down process. The bottom-up process operates rapidly and uses low-level features to find salient regions. On the other hand, the top-down process takes time and uses a priori knowledge about the scene or task-oriented knowledge to inhibit or enhance the bottom-up process.

Significant progress has been made in attention modelling for still images. On the other hand, attention modelling for videos remains lesser explored. Studies have shown that motion attracts attention and the importance of temporal cues for videos [1]. Therefore, unlike the spatial saliency models which only consider spatial cues, e.g., SUN [2], SIG [3] and CCH [4] [5] models, spatial-temporal saliency models consider both spatial and temporal cues. The SEO model [6] is a spatial-temporal model to compute bottom-up saliency maps, by using local steering kernels and kernel density estimation. The model measures the likeness of a voxel from its neighbouring voxels, by computing the centre-surround differences between the voxel and its neighbouring voxels. Unlike the SEO model which considers local rarity, the Mancas model [7] computes bottom-up saliency maps, by using features derived from optical flow with a global motion rarity quantification process. Similarly, the Culibrk model [8] measures global motion rarity. After using a multi-scale background subtraction approach to detect salient motion regions in a video frame, a modified Z-score is applied to detect the motion outlier regions in the video frame. Existing spatial-temporal models consider either local or global motion rarity. Our proposed saliency model considers both local and global motion rarity.

While the bottom-up factors that influence attention are better understood, the top-down factors that influence attention remain lesser understood. Detection proposal methods [9] generate detection windows with corresponding objectness scores, where each objectness score measures how likely the detection window contains an object (of any category). Mainly used as precursors to object detection, detection proposal methods have been used to speed up object detection and improve object detection quality. Since psychological studies have shown the importance of higher-level stimulus such as objects in attentional guidance, some recent works explore attention modelling by referencing the object proposals. For instance, Chang et al. fuse the saliency and objectness maps by optimizing an energy function [10]. Judd et al. treat each map as a feature and then use supervised learning to build a saliency classifier [11]. Similarly, Goferman et al. use maximum operation to combine the bottom-up and top-down results [12]. As objectness has been shown to be useful as a top-down cue, we also adopt such top-down objectness cues in our attention modelling.

In this paper, we proposed a visual attention model that incorporates both bottom-up spatial-temporal saliency and top-down objectness. The bottom-up cues are incorporated by means of the co-occurrence histogram spatial-temporal saliency maps while the top-down cues are incorporated by means of an objectness map. This paper is organized as follows. First, our proposed method is presented in Section II. Subsequently, our experimental results are shown in Section III. Finally, the conclusion is discussed in Section IV.

II. Proposed Method

In this section, we present a visual attention model that captures several low-level and mid-level saliency cues by using co-occurrence histograms [4] [5] and objectness.

A. Co-occurrence Histogram Saliency Model

Our proposed attention model adopts the idea of saliency modelling by using co-occurrence histograms as described in [4] [5]. The steps to extract a co-occurrence histogram saliency model are described as follows. Let I denote an input image and K denote the possible values within I. Then, the two-dimensional co-occurrence histogram, $H(m,n)$, $m,n \in K$, is constructed as

$$H(m,n) = \sum_{i'=i-z}^{i+z} \sum_{j'=j-z}^{j+z} \sum_{I(i,j)=m} \begin{cases} 1 & \text{if } I(i',j') = n, \\ 0 & \text{if otherwise.} \end{cases} \quad (1)$$

$I(i,j)$ and $I(i',j')$ denote the input image values at locations (i,j) and (i',j') respectively, and z denotes the window size around a pixel. In other words, for each pixel with an input image value m, the occurrence of neighbouring pixels with an input image value n contributes to the histogram count $H(m,n)$. A distribution function, $P(m,n)$, $m,n \in K$, is then obtained by normalizing H, i.e.,

$$P(m,n) = \frac{H(m,n)}{\sum_{m'=1}^{K} \sum_{n'=1}^{K} H(m',n')}. \quad (2)$$

Since saliency is typically negatively correlated with occurrence and co-occurrence, an inverted distribution function, $\tilde{P}(m,n)$, m, $n \in K$, is computed as

$$\begin{aligned} &\tilde{P}(m,n) \\ = \quad &\begin{cases} \frac{1}{\|P\|_0} - P(m,n) & \text{if } \frac{1}{\|P\|_0} - P(m,n) \geq 0, \\ 0 & \text{if otherwise,} \end{cases} \end{aligned} \quad (3)$$

where $\|P\|_0$ is the number of non-zero elements in P. The threshold ensures that input image value pairs that are more common than average are not considered salient.

Finally, the visual saliency map, $M(i,j)$, is constructed as

$$M(i,j) = \sum_{i'=i-z}^{i+z} \sum_{j'=j-z}^{j+z} \tilde{P}(I(i,j), I(i',j')). \quad (4)$$

Similarly, $I(i,j)$ and $I(i',j')$ denote the input image values at locations (i,j) and (i',j') respectively, and z denotes the window size around a pixel. The co-occurrence histogram saliency model captures both the occurrence and co-occurrence of the input image values in an input image.

B. Visual Attention Model

Human visual attention is attracted by a number of factors including spatial saliency as modelled by centre-surround difference, temporal saliency due to object motion, as well as, various object semantics such as faces, texts, vehicles, etc. Our proposed visual attention model aims to capture these low-level and mid-level cues by leveraging on the idea of using co-occurrence histogram saliency modelling as described in Section II-A.

Given a video frame, the attention map, S, is computed as

$$S = M_{\mathrm{C}} + M_{\mathrm{G}} + \alpha(M_{\mathrm{OM}} + M_{\mathrm{OG}}) + M_{\mathrm{O}}, \quad (5)$$

where M_{C}, M_{G}, M_{OM}, and M_{OG} are the co-occurrence histogram saliency maps obtained from contrast, gradient, magnitude of optical flow, and gradient of optical flow respectively, while M_{O} is an objectness map. An illustration of the proposed framework for a video frame is shown in Figure 1.

Saliency maps M_{C} and M_{G} are used to capture the bottom-up spatial cues. To obtain M_{C}, the RGB images $I \in \mathbb{R}^3$ are first converted to LAB images $I_c \in \mathbb{R}^3$. Then, for each channel in I_c, a saliency map is computed by using the co-occurrence histogram model. Finally, the saliency maps are fused by taking maximum operation across the color channels. To obtain M_{G}, the RGB images $I \in \mathbb{R}^3$ are first converted to gradient images $I_G \in \mathbb{R}^3$. Similarly, for each channel in I_G, a saliency map is computed by using the co-occurrence histogram model. The saliency maps are finally fused by taking maximum operation across the channels.

Saliency maps M_{OM} and M_{OG} are used to capture the bottom-up temporal cues. To obtain M_{OM} and M_{OG}, optical flows are first extracted from a pair of consecutive video frames. The optical flows are extracted using the OpenCV implementation of the TV-L1 algorithm [13] [14]. The magnitude and gradient of the optical flows are then used to generate M_{OM} and M_{OG} respectively by using the co-occurrence histogram model. The co-occurrence histogram model captures both local and global motion rarity by capturing the occurrence and co-occurrence of the optical flows.

Objectness map M_{O} is used to capture the top-down objectness cues. To obtain M_{O}, 1000 detection windows with corresponding objectness scores are first extracted using the Objectness method [15]. Then, M_{O} is obtained by accumulating the 1000 detection windows with corresponding objectness scores.

An adaptive motion weight, α, is used to trade-off the contributions of the bottom-up spatial and temporal cues. Motion-compensated frame differencing is first performed between two consecutive video frames to handle camera movement. In motion-compensated frame differencing, a frame is first warped toward its consecutive frame by using their estimated optical flows. The differences between the warped frame and its consecutive unwarped frame are then used to compute the adaptive motion weight as

$$\alpha = \frac{\max_{\forall i \forall j} \left(\max_{i'=i-1}^{i'=i+1} \max_{j'=j-1}^{j'=j+1} D(i',j') \right)}{\mathrm{mean}_{\forall i \forall j} \left(\max_{i'=i-1}^{i'=i+1} \max_{j'=j-1}^{j'=j+1} D(i',j') \right)} - 1, \quad (6)$$

where D is the difference image between the warped frame and its consecutive unwarped frame. The motion weight is designed to be large when large foreground motion is present. i.e., when the maximum difference is much larger than the mean difference.

For a video, the final attention maps, F_t, are computed

$$F_t = (1-\beta)(S_{t-1}) + \beta(S_t), \quad (7)$$

where S_{t-1} and S_t are the attention maps for the video frames at time $t-1$ and t respectively. Temporal averaging of the attention maps is performed to ensure smoothness of the

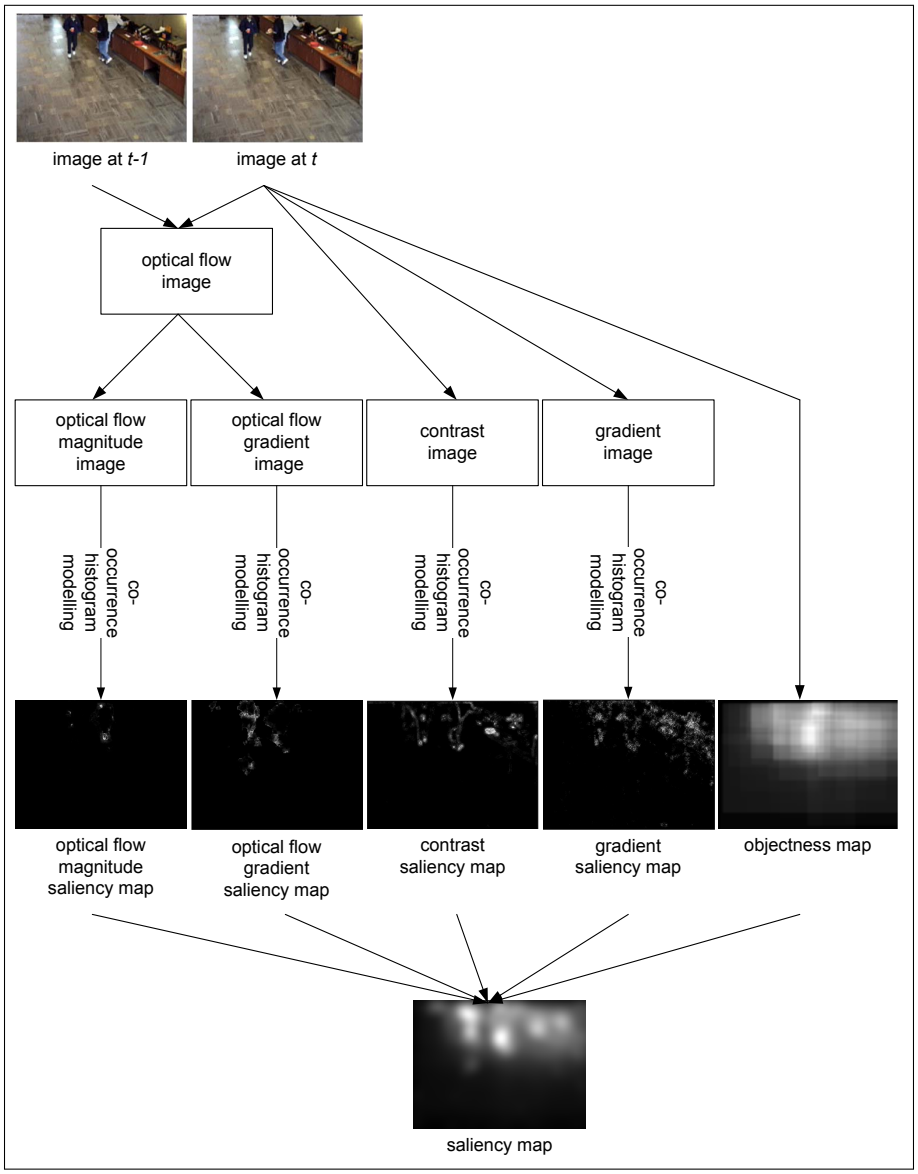


Fig. 1. An illustration of the proposed framework for a video frame.

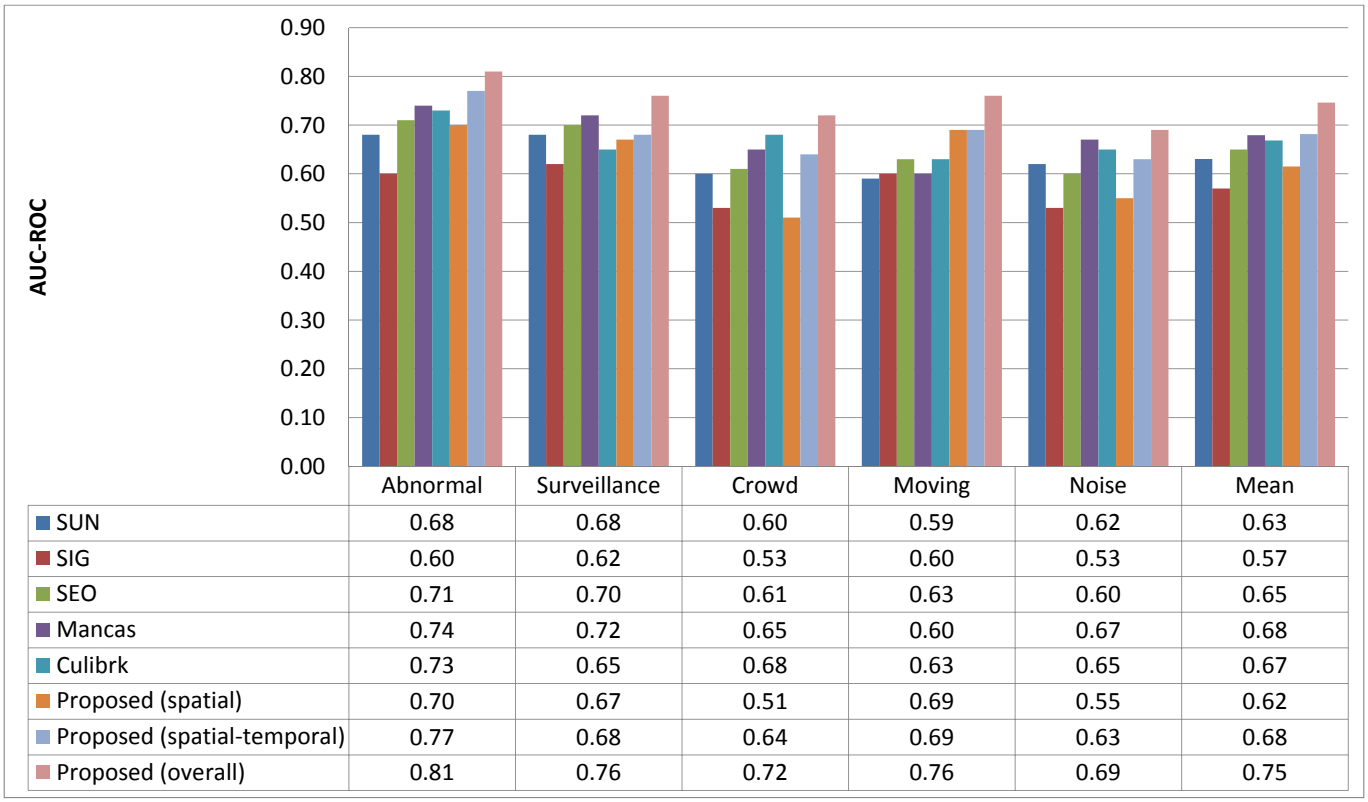
attention maps across the video frames. β is experimentally set to 0.9 in our implemented system.

III. Experimental Results

We evaluated our proposed visual attention model on the ASCMN [16] and UCF Sports Action [17] [18] [19] [20] datasets containing videos with eye-tracking annotations. The ASCMN dataset contains five categories of videos, including "Abnormal" (videos with abnormal motion), "Surveillance (surveillance videos), "Crowd" (videos with crowd motion), "Moving" (videos from moving cameras), and "Noise" (videos with sudden salient motion). On the other hand, the UCF Sports Action dataset contains broadcast videos. The diversity of videos allows comprehensive testing of the visual attention models under different conditions. As in [16], the area under curve - receiver operating characteristic (AUC-ROC) was used for our evaluation. We used the evaluation functions provided in the ASCMN dataset for our evaluation on both datasets. A score of 1 corresponds to perfect prediction while a score of 0.5 indicates chance level.

The experimental results on the ASCMN dataset are shown in Figure 2. The proposed (overall) model denotes the visual attention model with bottom-up spatial-temporal saliency and top-down objectness. The proposed (spatial) model denotes the visual attention model with only bottom-up spatial saliency while the proposed (spatial-temporal) model denotes the visual attention model with bottom-up spatial-temporal saliency. Our proposed models were compared against some state-of-the-art saliency models, mainly SUN [2], SIG [3], SEO [6], Mancas [7], and Culibrk [8] models.

The proposed (overall) model out-performed the state-of-the-art methods on the ASCMN dataset. In general, the spatial



	Abnormal	Surveillance	Crowd	Moving	Noise	Mean
SUN	0.68	0.68	0.60	0.59	0.62	0.63
SIG	0.60	0.62	0.53	0.60	0.53	0.57
SEO	0.71	0.70	0.61	0.63	0.60	0.65
Mancas	0.74	0.72	0.65	0.60	0.67	0.68
Culibrk	0.73	0.65	0.68	0.63	0.65	0.67
Proposed (spatial)	0.70	0.67	0.51	0.69	0.55	0.62
Proposed (spatial-temporal)	0.77	0.68	0.64	0.69	0.63	0.68
Proposed (overall)	0.81	0.76	0.72	0.76	0.69	0.75

Fig. 2. AUC-ROC scores on the ASCMN dataset.

saliency models, i.e., SUN, SIG and proposed (spatial) models, obtained lower AUC-ROCs. This is followed by the spatial-temporal models, i.e., SEO, Mancas, Culibrk, and proposed (spatial-temporal) models, which obtained higher AUC-ROCs. Furthermore, among the spatial-temporal models, the SEO model which considers local motion rarity obtained slightly lower AUC-ROC than the Culibrk and Mancas models which consider global motion rarity. Our proposed (spatial-temporal) model which considers both local and global motion rarity obtained highest AUC-ROC among the spatial-temporal models, suggesting the effectiveness of modelling both local and global motion rarity.

The proposed models using spatial, spatial-temporal, and overall features obtained average AUC-ROCs of 0.62, 0.68, and 0.75 respectively on the ASCMN datastset. The improvement of 0.06 (0.68-0.62) shows the effectiveness of the temporal cues for visual attention modelling. The improvement of 0.07 (0.75-0.68) further shows the effectiveness of the objectness cues for visual attention modelling.

The experimental results on the UCF Sports Action dataset are shown in Table I. Our proposed models were compared against some state-of-the-art saliency models where source codes are publicly available. The proposed model using spatial-temporal features and proposed model using overall features achieved AUC-ROCs of 0.80 and 0.84 respectively, out-performing other state-of-the-art models. Similarly, the proposed (overall) model achieved higher AUC-ROC than the proposed models using spatial and spatial-temporal features only, suggesting the effectiveness of the temporal and top-down objectness cues for visual attention modelling.

TABLE I
AUC-ROC SCORES ON THE UCF SPORTS ACTION DATASET.

Model	AUC-ROC
SUN	0.73
SIG	0.68
SEO	0.78
Proposed (spatial)	0.75
Proposed (spatial-temporal)	0.80
Proposed (overall)	**0.84**

IV. CONCLUSION

In this paper, we proposed a visual attention model that incorporates both bottom-up spatial-temporal saliency and top-down objectness. By using the co-occurrence histogram model to model temporal saliency, both local and global motion rarity can be captured. Objectness also helps to boost the modeling performance greatly. The proposed model showed good performance on the ASCMN and UCF Sports Action datasets. In extension of this work, other means of motion information besides optical flow will be investigated. The role of objectness for attention modelling will also be more thoroughly investigated.

REFERENCES

[1] R. A. Abrams and S. E. Christ, "Motion onset captures attention," *Psychological Science*, vol. 14, no. 5, pp. 427–432, 2003. [Online]. Available: http://dx.doi.org/10.1111/1467-9280.01458

[2] L. Zhang, M. H. Tong, T. K. Marks, H. Shan, and G. W. Cottrell, "SUN: A Bayesian framework for saliency using natural statistics," *Journal of Vision*, vol. 8, no. 7, p. 32, 2008.

[3] X. Hou, J. Harel, and C. Koch, "Image signature: Highlighting sparse salient regions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 1, pp. 194–201, Jan 2012.

[4] S. Lu and J.-H. Lim, "Saliency modeling from image histograms," in *European Conference on Computer Vision (ECCV)*, 2012, pp. 321–332.

[5] S. Lu, C. Tan, and J.-H. Lim, "Robust and efficient saliency modeling from image co-occurrence histograms," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 1, pp. 195–201, Jan 2014.

[6] 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.

[7] M. Mancas, N. Riche, J. Leroy, and B. Gosselin, "Abnormal motion selection in crowds using bottom-up saliency," in *IEEE International Conference on Image Processing (ICIP)*, Sep 2011, pp. 229–232.

[8] D. Culibrk, M. Mirkovic, V. Zlokolica, M. Pokric, V. Crnojevic, and D. Kukolj, "Salient motion features for video quality assessment," *IEEE Transactions on Image Processing*, vol. 20, no. 4, pp. 948–958, Apr 2011.

[9] J. Hosang, R. Benenson, and B. Schiele, "How good are detection proposals, really?" in *British Machine Vision Conference (BMVC)*, 2014.

[10] K.-Y. Chang, T.-L. Liu, H.-T. Chen, and S.-H. Lai, "Fusing generic objectness and visual saliency for salient object detection," in *IEEE International Conference on Computer Vision (ICCV)*, 2011.

[11] T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to predict where humans look," in *IEEE International Conference on Computer Vision (ICCV)*, 2009.

[12] S. Goferman, L. Zelnik-Manor, and A. Tal, "Context-aware saliency detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 10, pp. 1915–1926, Oct 2012.

[13] J. Sanchez Perez, E. Meinhardt-Llopis, and G. Facciolo, "TV-L1 optical flow estimation," *Image Processing On Line*, vol. 3, pp. 137–150, 2013.

[14] C. Zach, T. Pock, and H. Bischof, "A duality based approach for realtime TB-L1 optical flow," in *In Ann. Symp. German Association Patt. Recogn*, 2007, pp. 214–223.

[15] B. Alexe, T. Deselaers, and V. Ferrari, "What is an object?" *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 0, pp. 73–80, 2010.

[16] N. Riche, M. Mancas, D. Culibrk, V. Crnojevic, B. Gosselin, and T. Dutoit, "Dynamic saliency models and human attention: A comparative study on videos," in *Computer Vision - ACCV 2012*, ser. Lecture Notes in Computer Science, K. Lee, Y. Matsushita, J. Rehg, and Z. Hu, Eds. Springer Berlin Heidelberg, 2013, vol. 7726, pp. 586–598.

[17] M. Rodriguez, J. Ahmed, and M. Shah, "Action MACH a spatio-temporal maximum average correlation height filter for action recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2008, pp. 1–8.

[18] K. Soomro and A. Zamir, "Action recognition in realistic sports videos," in *Computer Vision in Sports*, ser. Advances in Computer Vision and Pattern Recognition, T. B. Moeslund, G. Thomas, and A. Hilton, Eds. Springer International Publishing, 2014, pp. 181–208.

[19] S. Mathe and C. Sminchisescu, "Dynamic eye movement datasets and learnt saliency models for visual action recognition," in *European Conference on Computer Vision (ECCV)*, 2012.

[20] S. Mathe and C. Sminchisesc, "Actions in the eye: Dynamic gaze datasets and learnt saliency models for visual recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 7, pp. 1408–1424, Jul 2015.