Novelty-based Spatiotemporal Saliency Detection for Prediction of Gaze in Egocentric Video

Patrik Polatsek, Wanda Benesova, *Member, IEEE*, Lucas Paletta, *Member, IEEE*, and Roland Perko

***Abstract*—The automated analysis of video captured from a first-person perspective has gained increased interest since the advent of marketed miniaturized wearable cameras. With this a person is taking visual measurements about the world in a sequence of fixations which contain relevant information about the most salient parts of the environment and the goals of the actor. We present a novel model for gaze prediction in egocentric video based on the spatiotemporal visual information captured from the wearer's camera, specifically extended using a subjective function of surprise by means of motion memory, referring to the human aspect of visual attention. Spatiotemporal saliency detection is computed in a bioinspired framework using a superposition of superpixel- and contrast based conspicuity maps as well as an optical flow based motion saliency map. Motion is further processed into a motion novelty map that is constructed by a comparison between most recent motion information with an exponentially decreasing memory of motion information. The innovative motion novelty map is experienced to be able to provide a significant increase in the performance of gaze prediction. Experimental results are gained from egocentric videos using eye-tracking glasses in a natural shopping task and prove a 6.48% increase in the mean saliency at a fixation in terms of a measure of mimicking human attention.**

***Index Terms*—Computer vision, feature extraction, image motion analysis, image processing, machine vision.**

I. Introduction

THE image-based analysis of egocentric video has gained increased interest with the increasing use of mass-marketed miniaturized wearable cameras, such as GoPro and Google Glass. A person is taking visual measurements about the world in a sequence of fixations which contain relevant information about the most salient parts of the environment and the goals of the actor. Prediction of gaze from the first-person perspective becomes increasingly relevant in order to interpret the continuous video stream in daily activities and deduce appropriate analytics and recommendations in the domains of health, social interaction analysis, traffic security, or in market research.

Location of human gaze and egocentric video analysis are intrinsically linked in the context of wearable vision. Recent research on first-person vision and egocentric video analysis [1] has recognized in an early phase the benefit of using attentional features for the purpose of activity recognition [2]. In the context of hand-eye coordination it has been exploited for video annotation that the distribution of both visual features and object occurrences in the vicinity of the gaze point is correlated with the verb-object pair describing the action [3]. Implicit cues from visual features, such as, hand location and pose, head and hand motion are useful features in this context [4]. Even in general settings on gaze prediction without significant focus on hand-eye coordination, camera motion estimation has been approved to represent a strong cue for gaze prediction [5]. In the frame of video summarisation, gaze provides the means to personalise the summary and provide a relevant feature for combinatorial optimisation [6].

Most work on gaze prediction has primarily focused on spatiotemporal attention detection based on classical, biologically inspired spatial approach from [7] which hierarchically decomposes visual features and processes according to the center-surround organization of ganglion cells. In the temporal information domain, optical flow methods are used to determine the motion information.

II. Related Work

The presented work is in the line of gaze prediction and intends to improve video-based human attention analysis by means of progressed saliency functionality, in particular in the frame of spatiotemporal saliency, by focusing on motion-based saliency.

Previous research on spatiotemporal saliency has focused on various aspects of temporal saliency computation. A spatiotemporal extension of the hierarchical model is defined in [8] where dynamic changes between consecutive video frames are expressed in two additional types of Gaussian pyramids and static and dynamic feature maps are fused into a spatiotemporal saliency map.

A spectral model proposed in [9] analyzes motion saliency in the frequency domain. Using an optical flow algorithm, an image is represented by flow magnitude and phase fields. Magnitude and orientation are separately processed to obtain the difference between the original and smoothed version of the log spectrum called the spectral residual.

Manuscript received December 07, 2015; accepted January 15, 2016. Date of publication January 28, 2016; date of current version February 15, 2016. This work was supported by Slovakian Grant VEGA 1/0625/14, the Austrian Research Promotion Agency under Contract 832045, Research Studio Austria FACTS, and by the Austrian Ministry for Transport, Innovation and Technology under project Collaborative Robotics. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Zhu Liu.

P. Polatsek and W. Benesova are with the Faculty of Informatics and Information Technologies Slovak University of Technology, Bratislava, Slovakia.

L. Paletta and R. Perko are with the Joanneum Research, Graz, Austria.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/LSP.2016.2523339

A generic multimodal approach on spatiotemporal attention is presented by [10] which combines visual, aural and linguistic attention models for video summarization, estimating dynamic saliency from motion vectors between macro blocks.

Another approach on spatiotemporal attention is proposed by [11] where center-surround feature processing and the theory known as discriminant saliency are combined. The set of features are spatiotemporal patches represented by a dynamic texture model. The discriminant feature selection is modelled by maximizing the mutual information between a set of features and class labels.

Another center-surround discriminant method [12] uses histogram differences between a center and surrounding regions to measure spatiotemporal saliency. The model computes color and edge orientation histograms and temporal gradients produced as intensity differences between frames.

The following work is related to our method by using a superpixel approach for video processing. Spatiotemporal saliency detection model proposed in [13] processes input videos at the superpixel level that fits to the boundaries of salient objects. The method computes optical flow-based motion and color histograms locally at superpixel level and globally at frame level. It exploits the assumption that moving salient objects generate a higher difference between superpixel-level and frame-level motion histograms. The second assumption lies in the temporal coherence of movements. Hence, it measures the similarity between a superpixel in the current frame and its correlated superpixels in the previous frame based on their color histograms and their mutual distance. Spatial and temporal pixel-level saliency maps are finally derived from superpixel-level measures and adaptively fused by considering their mutual consistency.

A further multimodal information-based spatiotemporal model presented in [14] measures saliency as rarity of color in CIE Lab space, orientation using Gabor filters and motion detected by optical flow. Ideas for the model are improved in [15]. The measurements are based on hyperhistograms built using a sliding cube as temporal concatenations of multiple histograms. A saliency map is finally enhancement by a SLIC algorithm.

Authors in [16] extract high-frequency signals representing the parvocellular-like retinal output for static saliency and low-frequency signals representing the magnocellular-like output for temporal saliency. Dynamic estimation of saliency includes a camera motion compensation and optical flow computations. A dynamic saliency map is derived from modules of motion flow vectors defined for each pixel.

The proposed work extends the concept of surprise to egocentric vision processing. A motion event is a process evolving in time due to which motion perception is also guided by our memory. The goal of this paper is to propose a novel motion detection method and provide this additional information within a novel spatiotemporal saliency model for the prediction of gaze in egocentric video. In particular, we are interested in the interpretation of video from eye-tracking glasses from which we take the opportunity to directly compute the performance of the novel method in estimating locations of human point-of-regards. Using the assumption of motion coherency within a superpixel we implement the theory using the hierarchical approach presented in [17]. The resulting superpixel-based spatiotemporal saliency model performs information fusion using the information from static saliency maps with a motion saliency map and a motion novelty map characterizing unexpected events in dynamically changing scenes.

III. Motion Novelty for Spatiotemporal Saliency

The proposed model is based on the *Hierarchical Superpixel-based Saliency Model* for the detection of bottom-up saliency [17]. The model segments input images into superpixels using a *SLIC* [18] algorithm.

The novel method is a combination of a standard hierarchical and a superpixel approach. Superpixel segmentation allows us to partially involve object-based attention in our model, something which is absent in standard hierarchical methods.

The algorithm is inspired by *Feature Integration Theory* [19] and consequently hierarchically processes all features using *Gaussian pyramids* with 6 layers. Center-surround organization of human ganglion cells is modelled as a difference between finer and coarser levels of the pyramid. The center is represented by scales $c \in \{0, 1, 2\}$ and the surround scales are $s = c + \delta$, where $\delta \in \{1, 2, 3\}$. Each pyramid layer consists of a *superpixel map* representing the locations of all superpixels and a *set of superpixel histograms*.

B. Spatial Saliency Map

A spatial form of the saliency model SM_S integrates *intensity*, *color* and *orientation*. More details about the used algorithm can be found in the publication [17].

B. Temporal Saliency Map

Motion processing requires 2-channel dense *optical flow maps* characterizing the angle and the magnitude of flow vectors.

Using the flow maps, each superpixel is represented by a *histogram of flow orientations* H_o and a *histogram of flow magnitudes* H_m, both with 90 bins. In order to compare superpixels on different pyramid layers, each superpixel is characterized by a flow vector with 2 parameters–*orientation* φ and *magnitude* r: $\mathbf{v} = [\overline{\varphi}_{H_o}, \overline{r}_{H_m}]$, where $\overline{x}_H$ denotes the mean value of a histogram H.

Let $\mathbf{v}_c(x, y)$ and $\mathbf{v}_s(x, y)$ be flow vectors of superpixels on a center and surround pyramid layer at location (x, y), a value of a motion feature map can be expressed as the magnitude of the vector difference:

$$FM_M(x, y) = \|\mathbf{v}_s(x, y) - \mathbf{v}_c(x, y)\|. \tag{1}$$

Using the law of cosines:

$$FM_M(x, y) = \sqrt{r_s^2 + r_c^2 - 2r_s r_c \cos\gamma}, \tag{2}$$

where r_i is the magnitude of the flow vector v_i at (x, y) scaled into the range $\langle 0, 0.5\rangle$ and γ indicates the angle between the corresponding vectors, as shown in Fig. 1. The highest value of

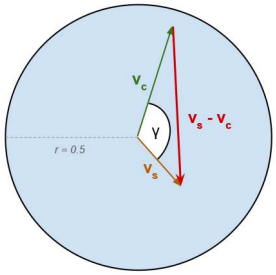

Fig. 1. Motion difference between center c and surround s scales of a pyramid.

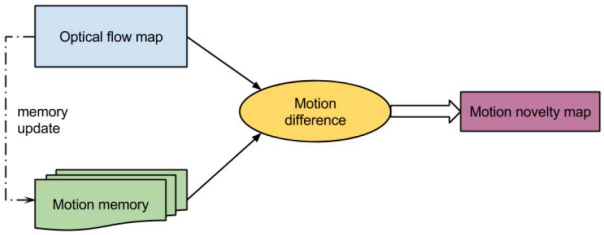

Fig. 2. Motion novelty is defined as the difference between the actual optical flow map and the accumulated flow map called the motion memory.

a feature map occurs when vectors at the same location have the opposite directions and the maximum velocity magnitudes.

Extracted feature maps are subsequently normalized and linearly combined into a temporal saliency map:

$$SM_T = \sum_i \mathcal{N}(FM_{M_i}). \tag{3}$$

A normalisation factor $\mathcal{N}$ multiplies a map by $(M-m)^2$, where M represents the global maximum and m denotes the average of all local maxima in rectangular blocks.

c. Motion Novelty Map

Motion feature maps represent the motion saliency of a current video frame. To determine dynamic changes in a scene, we build a *motion novelty map* considering not only a single optical flow map but also a subsequence of several previous flow maps.

Temporal changes in motion are detected using a *motion memory*, which is updated with each incoming optical flow map. The update can be defined as:

$$MEM_{t+1} = (1-\eta)MEM_t + \eta o_{t+1}, \tag{4}$$

where MEM_t represents a motion memory at time t, o_t is an optical flow map and a *learning rate* $\eta = 0.05$.

Before the memory update, an actual flow map is compared with the motion memory (Fig. 2).

Each pixel of both compared images represents a vector $v = [\varphi, r]$, where φ and r define magnitude and orientation at the pixel position.

A pixel-by-pixel comparison between the memory and the flow map is analogous to the motion difference in a motion feature map (Section III-B). Motion novelty is then defined by the following formula:

$$NM_{M_t}(x,y) = \|\mathbf{v}_{MEM_t}(x,y) - \mathbf{v}_{o_t}(x,y)\|, \tag{5}$$

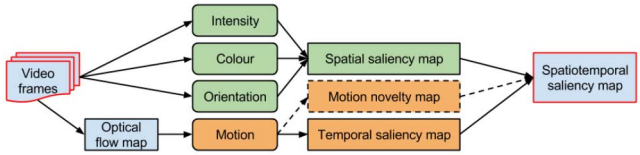

Fig. 3. General scheme of spatiotemporal saliency model.

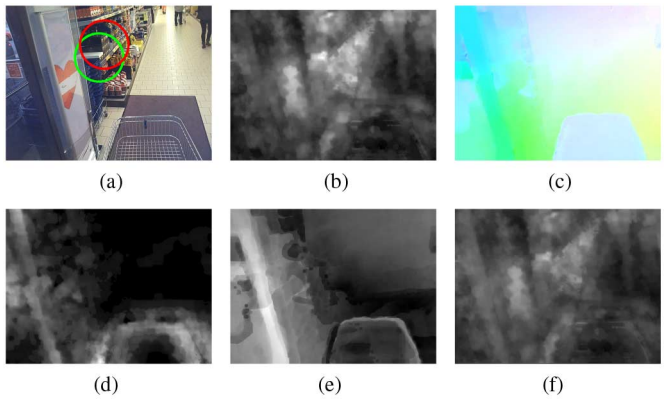

Fig. 4. Fusion of saliency maps into a spatiotemporal saliency map with motion novelty using a motion rate $\lambda = 0.25$. The most salient location is marked by a green circle and a fixation by a red circle. (a) Video frame. (b) Spatial saliency map. (c) Optical flow map. (d) Temporal saliency map. (e) Motion novelty map. (f) Spatiotemporal saliency map.

where $\mathbf{v}_{MEM_t}(x,y)$ and $\mathbf{v}_{o_t}(x,y)$ are vectors at (x,y) in the motion memory and the actual flow map, respectively.

D. Spatiotemporal Fusion

Fusion of a spatial saliency map SM_S and a temporal saliency map SM_T results in a single spatiotemporal map:

$$SM = (1-\lambda)SM_S + \lambda SM_T, \tag{6}$$

where λ denotes a motion saliency rate. Considering temporal changes in a video sequence obtained from a motion novelty map NM_M, the equation for a final saliency map has the following formula (Fig. 3):

$$SM = (1-\lambda)SM_S + \frac{\lambda}{2}SM_T + \frac{\lambda}{2}NM_M. \tag{7}$$

An example of a spatiotemporal saliency map with motion novelty is included in Fig. 4.

IV. Experimental Results

The evaluation of most attention models is based on eye-tracking data from simple videos displayed on a screen. However, such conditions cannot simulate real human visual attention, and motion perception is completely different when surrounding objects as well as an observer may move. Hence, we have decided to use an evaluation video dataset captured under real conditions for the evaluation of our saliency model.

The dataset (2 videos, 860 frames in total of 1280×960 size) was recorded using eye-tracking glasses at a shopping mall (mostly one fixation per frame). Viewers were asked to

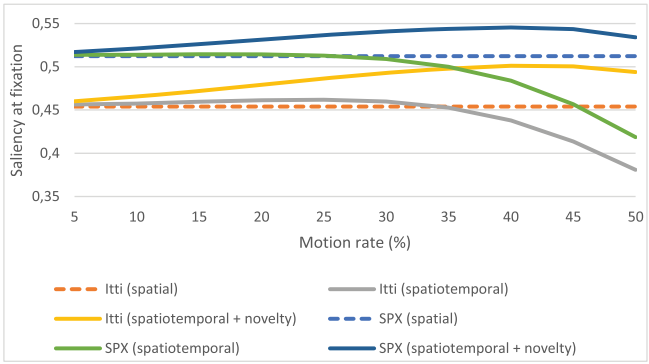

Fig. 5. Mean saliency at a fixation using different motion rates. Dashed lines represent the performance of spatial saliency models without motion processing. A superpixel-based temporal saliency map and a motion novelty map have been added into a novel superpixel-based (SPX) model as well as a standard spatial saliency model inspired by [7].

Fig. 6. Examples in which a saliency map correctly predicts fixation locations. The most salient location is marked by a green circle and a fixation by a red circle ($\lambda = 0.25$). (a) Video frame. (b) Optical flow map. (c) Saliency map.

find two particular products in a store. Eye-tracking data are supplemented by dense optical flow maps based on [21].

In order to test the video dataset, the superpixel-based [17] and a standard hierarchical saliency map based on [7] are fused with the novel dynamic superpixel-based saliency map. The fusion into a spatiotemporal map is expressed in Equation (6), and also in Equation (7) when a motion novelty map is considered.

As an evaluation metric, the average of saliency values at fixation locations have been used. We have investigated the effect of changing a motion rate λ in a saliency map. The results compared with spatial saliency modes are visualized in Fig. 5.

For visualization purposes the most salient location in the resulting saliency map is labeled with a green circle and a fixation location with a red circle. Figures included in this section represent saliency maps obtained from the proposed saliency model with motion novelty. A sample saliency map in which the most salient location of the novel spatiotemporal saliency approximately equals to the human fixation is depicted in Fig. 6.

The best performance with the novel spatiotemporal saliency method occurs when $\lambda = 0.40$. Moving objects are briefly perceived by the observer but not continuously. The human gaze is mostly directed at objects in motion at the beginning of the encounter. Afterwards these objects are not tracked anymore. Using motion novelty, the saliency of a continuously moving trolley can be decreased appropriately, as shown in Fig. 7. However, the proposed motion novelty map does not track objects, it just learns about the motion direction and magnitude at a given location. Hence, the motion novelty map considers as

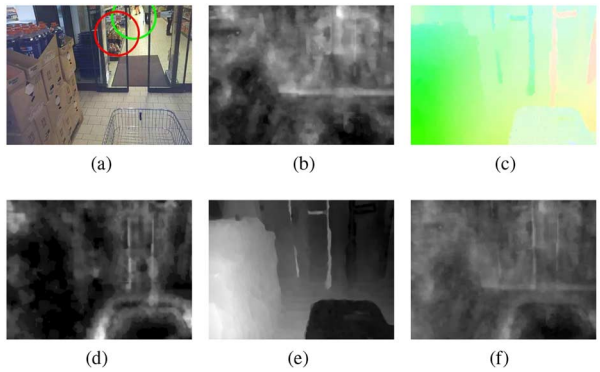

Fig. 7. Motion novelty reduces a final saliency value at locations where motion has not been changed. The most salient location is marked by a green circle and a fixation by a red circle ($\lambda = 0.5$). (a) Video frame. (b) Spatial saliency map. (c) Optical flow map. (d) Temporal saliency map. (e) Motion novelty map. (f) Spatiotemporal saliency map.

TABLE I
EXPERIMENTAL RESULTS OF COMPARED SALIENCY MODELS USING THE AUC AND NSS SCORE

Metric	SPX	Itti [7]	Liu [13]
AUC	**0.7050**	0.6754	0.6953
NSS	0.7264	0.6289	**0.7636**

novelty objects that are moving to another location or objects that change motion.

The performance of the novel model has been compared with a spatial hierarchical saliency detection algorithm [7] and a spatiotemporal saliency detection algorithm based on a superpixel segmentation [13] using the Normalized Scanpath Saliency (*NSS*) and the Area Under the ROC Curve (*AUC*) [22]. Saliency maps have been smoothed with a Gaussian kernel with standard deviations σ from 0.01 to 0.13 in image width (in steps of 0.01) and optimal σ values producing the maximum AUC have been taken. The highest NSS score is achieved by the model described in [13], but the novel model ($\lambda = 0.4$, with motion novelty) outperforms both compared models using the AUC (Table I).

Experimental results also shows 6.48% increase in the mean saliency at a fixation in terms of a measure of mimicking human attention.

V. CONCLUSION

In this paper, we have proposed a motion memory and novelty approach in the context of superpixel-based spatiotemporal saliency detection. The method is particularly beneficial for the prediction of gaze in egocentric video. Spatial saliency computation is extended by motion saliency extracted from optical flow maps. Eventually, novelty in the motion perception refers to subjective human saliency measures in the frame of surprise, highly useful in a world of unexpected dynamic events. Fusion of spatial, motion and novelty based attentional factors results in a novel spatiotemporal saliency model to predict human gaze better in egocentric video from daily tasks than comparable models do under evaluation of a performance measure that quantifies success with the capacity to match with human saliency in real environments.

REFERENCES

[1] A. Betancourt, P. Morerio, C. S. Regazzoni, and M. Rauterberg, "An overview of first person vision and egocentric video analysis for personal mobile wearable devices," *arXiv preprint arXiv:1409.1484*, 2014.

[2] K. Yamada, Y. Sugano, T. Okabe, Y. Sato, A. Sugimoto, and K. Hiraki, "Attention prediction in egocentric video using motion and visual saliency," *Adv. Image and Video Technology*, Berlin, Germany, Springer, 2012, pp. 277–288.

[3] A. Fathi, Y. Li, and J. M. Rehg, "Learning to recognize daily actions using gaze," *Computer Vision–ECCV 2012*, Berlin, Germany: Springer, 2012, pp. 314–327.

[4] Y. Li, A. Fathi, and J. M. Rehg, "Learning to predict gaze in egocentric video," *IEEE Int. Conf. Computer Vision (ICCV), 2013*, 2013, pp. 3216–3223, IEEE.

[5] K. Matsuo, K. Yamada, S. Ueno, and S. Naito, "An attention-based activity recognition for egocentric video," *IEEE Conf. IEEE Computer Vision and Pattern Recognition Workshops (CVPRW), 2014*, 2014, pp. 565–570.

[6] J. Xu, L. Mukherjee, Y. Li, J. Warner, J. M. Rehg, and V. Singh, "Gaze-enabled egocentric video summarization via constrained submodular maximization," *Proc. the IEEE Conf. Computer Vision and Pattern Recognition*, 2015, pp. 2235–2244.

[7] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Patt. Anal. Mach. Intell.*, vol. 20, no. 11, pp. 1254–1259, Nov. 1998, http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=730558

[8] L. Itti, N. Dhavale, and F. Pighin, "Realistic avatar eye and head animation using a neurobiological model of visual attention," *Proc. SPIE, 48th Annu. Meeting*, 2004, pp. 64–78.

[9] C. Loy, T. Xiang, and S. Gong, "Salient motion detection in crowded scenes," *5th Int. Symp. Communications Control and Signal Processing (ISCCSP), 2012*, May 2012, pp. 1–4.

[10] Y.-F. Ma, X.-S. Hua, L. Lu, and H.-J. Zhang, "A generic framework of user attention model and its application in video summarization," *IEEE Trans. Multimedia*, vol. 7, no. 5, pp. 907–919, 2005.

[11] V. Mahadevan and N. Vasconcelos, "Spatiotemporal saliency in dynamic scenes," *IEEE Trans., Patt. Anal. Mach. Intell.*, vol. 32, no. 1, pp. 171–177, 2010.

[12] W. Kim, C. Jung, and C. Kim, "Spatiotemporal saliency detection and its applications in static and dynamic scenes," *IEEE Trans., Circuits Syst. Video Technol.*, vol. 21, no. 4, pp. 446–456, 2011.

[13] Z. Liu, X. Zhang, S. Luo, and O. Le Meur, "Superpixel-based spatiotemporal saliency detection," *IEEE Trans., Circuits Syst. Video Technol.*, vol. 24, no. 9, pp. 1522–1540, 2014.

[14] M. Decombas, N. Riche, F. Dufaux, B. Pesquet-Popescu, M. Mancas, B. Gosselin, and T. Dutoit, "Spatio-temporal saliency based on rare model," *20th IEEE Int. Conf. Image Processing (ICIP), 2013*, 2013, pp. 3451–3455, IEEE.

[15] I. Cassagnea, N. Riche, M. Décombas, M. Mancas, B. Gosselin, T. Dutoit, and R. Laganiere, "Video saliency based on rarity prediction: Hyperaptor," *2015 23rd Eur. Signal Processing Conf. (EUSIPCO)*, Nice, France, Sep. 2015, pp. 1521–1525.

[16] S. Marat, T. H. Phuoc, L. Granjon, N. Guyader, D. Pellerin, and A. Guérin-Dugué, "Modelling spatio-temporal saliency to predict gaze direction for short videos," *Int. J. Comput. Vis.*, vol. 82, no. 3, pp. 231–243, 2009.

[17] P. Polatsek and W. Benesova, "Bottom-up saliency model generation using superpixels," *Proc. 31st Spring Conf. Computer Graphics*, New York, NY, USA, 2015, pp. 121–129, http://doi.acm.org/10.1145/2788539.2788557 ser. SCCG '15.

[18] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Susstrunk, "Slic superpixels compared to state-of-the-art superpixel methods," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2274–2282, Nov. 2012, http://dx.doi.org/10.1109/TPAMI.2012.120

[19] A. M. Treisman and G. Gelade, "A feature-integration theory of attention," *Cognitive Psychology*, vol. 12, pp. 97–136, 1980.

[20] C. Zach, T. Pock, and H. Bischof, "A duality based approach for realtime tv-l1 optical flow," *Proc. 29th DAGM Conf. Pattern Recognition*, Heidelberg, Germany, 2007, pp. 214–223.

[21] A. Chambolle and T. Pock, "A first-order primal-dual algorithm for convex problems with applications to imaging," *J. Math. Imag. Vis.*, vol. 40, no. 1, pp. 120–145, 2011.

[22] A. Borji, H. R. Tavakoli, D. N. Sihite, and L. Itti, "Analysis of scores, datasets, and models in visual saliency prediction," *IEEE Int. Conf. Computer Vision (ICCV), 2013*, 2013, pp. 921–928, IEEE.