OBJECTIVE PREDICTION OF VISUAL SALIENCY MAPS IN EGOCENTRIC VIDEOS FOR CONTENT-ACTION INTERPRETATION

Hugo Boujut, Vincent Buso, Jenny Benois-Pineau

OBJECTIVE PREDICTION OF VISUAL SALIENCY MAPS IN EGOCENTRIC VIDEOS FOR CONTENT-ACTION INTERPRETATION

H. Boujut, V. Buso, J. Benois-Pineau

University of Bordeaux, LaBRI, UMR 5800, F-33400 Talence, France

ABSTRACT

Extraction of visual saliency from video is in the focus of intensive research nowadays due to the variety and importance of application areas. In this paper we study the relation between subjective saliency maps, recorded on the basis of gaze-tracker data in a new upcoming video content: the egocentric video recorded with wearable cameras. On the basis of physiological research and comparing the subjective maps of an Actor performing activities of everyday life and a Viewer who interprets the video after it has been recorded, we identify the temporal shift between these two saliency maps. Using this relation we propose an ”à la carte” prediction of saliency maps of an Actor for the beginning of actions by an objective saliency model we previously developed. All the components of objective saliency: spatial, temporal and central bias are merged in this prediction. The commonly used quality metrics for pixel-based saliency prediction such as Pearson Correlation Coefficient, Normalized Scan Path and Area Under Curve show the good correspondence of predicted maps for Actor and Viewer. This research seems to us promising for content interpretation coming from mobile video recording devices.

1. INTRODUCTION

Automatic extraction of visually salient areas from video content is a strong research direction because of the needs in various fields of video analysis such as video quality assessment (VQA), Region-of-Interest (ROI) detection for advanced video coding and finally for efficient video content interpretation and object recognition in video. Since recently a new video content is massively coming into practice: the egocentric video recorded by body-weared cameras by sportsmen, or in the framework of behavioral studies for neurodegenerative diseases [1], or for entertainment purposes. In this case the problem of visual saliency detection is posed in a new and challenging way. Two different persons are involved in video acquisition and video interpretation process: the Actor who is wearing the video camera and the Viewer who is interpreting the video. Their visual saliencies are not the same. Indeed according to the physiological studies [2, 3], the human gaze anticipates the motor action of limbs when fulfilling an activity. When the viewer interprets the video acquired with wearable devices, he is much more interested in the action recorded and hence his saliency is different from that one of an actor. In various problems of interpretation of video content, such as studies of neurodegenerative diseases [1], there is a need to predict a physiologically normal saliency map of an actor and to do this in an automatic way. In this paper we study the relation between visual saliency maps from Actor and Viewer and propose a prediction of actor saliency map from an objective saliency model built upon previous research in [4, 5, 6]. This research has become possible due to the availability of a new video dataset recorded by a camera on looking glasses with an integrated eye-tracker [7]. The rest of the paper is organized as follows. In Section 2 we propose the study of relationships between the Actor's and Viewer's saliencies realized on a subjective saliency maps, obtained with eye-trackers. In Section 3 we propose a method of adaptation of objective saliency maps of viewer to retrieve the actor saliency maps for the beginning of actions. The experimental results are also presented in these sections. Section 4 concludes this work and outlines its perspectives.

2. STUDYING ACTORS' AND VIEWERS' POINTS OF VIEW

In this section we firstly explicit the methodology of building subjective saliency maps or in other words ”visual attention maps” and their comparison. Furthermore we estimate the temporal relation between subjective saliency maps of Actor and Viewer using manual and automatic metrics.

2.1. Subjective saliency maps building method

The subjective saliency maps in images and videos are built from eye position measurements in image/video plan. Indeed the attractors such as contrast, motion (in video), colors make the humans fixate some narrow areas in the video plan. With the help of eye-trackers the gaze projection in video frames can be recorded. There are two reasons for which eye positions cannot be directly used to represent the areas of visual attention. First, the eye positions are only spots on the frame and do not represent the field of view. Secondly, in the case of Viewers to get accurate results, the eye positions of several experimental subjects observing video content, are recorded. These positions vary from one subject to another and represent sparse discrete maps. In order to determine the areas of visual attraction in images and videos, we need dense maps. The method proposed by D. S. Wooding [8] has become the reference [9] since it fulfils these two constraints. In this method a two dimensional Gaussian is applied at the center of every eye-fixations. The Gaussian spread σ is set to an angle of 2° to reproduce the fovea projection of the screen as proposed in [10]. Then the Gaussians are summed-up and the final map is normalized. No matter for which recording of fixations is the eye-tracker used for, Wooding's method can be applied. Hence in our work we apply this method to build both Actor's and Viewer's attention maps from the eye-recordings. We remind that the Actor data is obtained by the eye-tracker worn by the actor and hence the data of only one subject is recorded for each video, while several Viewers observe the same video to simulate video interpretation conditions.

2.2. Comparison of saliency maps

The normalized saliency maps of Actor and Viewer can be compared with help of dedicated metrics. A good survey has recently been published in [11] about them. From this survey and anterior work [12] we retained the Normalized Scan Path, the Pearson correlation coefficient (PCC) and the ROC area, or the Area Under Curve(AUC) as most frequently used and suitable for the comparison of pixel-based saliency maps. Since results prove the scores to be highly correlated between these metrics (table 1, 2), only the AUC is displayed in this paper.

In AUC the problem is limited to a two-class prediction (binary classification). Pixels of one saliency map which is considered as "ground truth" as well as those of the predicted saliency map are labelled either as fixated or not fixated. A ROC curve plotting the false positive rate as a function of the true positive rate is used to present the classification result. The metric consists in computing the area under this ROC curve.

2.3. Experiments and Results

In this section we compared the actors' and viewers' points of view using different approaches: manually and automatically. The GTEA corpus and eye-tacker recording of viewers' gazes are explained before comparing the results of these two methods.

2.3.1. Corpus description

For this work, a dataset containing the eye locations of the persons performing the actions (Actors) is needed in order to compare their gaze-recordings with the gaze coordinates of the people watching these actions on video (Viewers). Along with their paper [7], the authors have publicly released two datasets. The GTEA gaze dataset has been obtained using the Tobii eye-tracking glasses. The videos and gaze locations are recorded thanks to a camera and infrared light system integrated to the glasses. The videos are at a 15fps rate and a 640×480 pixel resolution. For the gaze location, two points per frame are recorded (30 samples per second). The subjects are asked to prepare a meal for themselves based on the different ingredients placed on the table in front of them. In total 17 videos of 4min average are available, performed by 14 different participants. The different noticeable actions related with the preparation of a meal (e.g. spread jam, take milk, . . .) are listed in [7].

2.3.2. Eye tracker setup

In order to get the eye location of the people watching the videos provided by the authors of [7], an eye–tracker experiment has been performed. The gaze positions have been recorded with a HS-VET 250Hz from Cambridge Research Systems Ltd at a rate of 250 eye positions per second. The experiment conditions and the experiment room were compliant with the recommendation ITU-R BT.500-11 [12]. Videos were displayed on a 23 inches LCD monitor with a native resolution of 1920x1080 pixels. To avoid image distortions, videos were not re-sized to the screen resolution. A mid-gray frame was inserted around the displayed video. 31 participants have been gathered for this experiment, 9 women and 22 men. For 3 participants some problems occurred in the eye-tracking recording process and so they have been discarded.

2.3.3. Human-based comparison of actions beginning

For our first comparison between actors and viewers, we manually annotated the moments when each of both sides focused on the beginning of a new action for 8 of the videos provided by the GTEA dataset. To decide weather a party was indeed focusing on a new action, we used the gaze provided by GTEA and the gaze recorded by our Eye-Tracker experiment. We considered the focusing of viewer's or actor's gaze on an object of interest related to a new action to be an acknowledgment of the realization from the corresponding party that a new action is happening. Since most of the actions cannot be considered as starting at a specific frame number, the results are an average value of every 4 frames to avoid the noise induced by manual annotation. Results are displayed in figure 1. From this histogram one can clearly notice a peak of time difference between the realization of actions for the two parties. Indeed most of the actions are acknowledged by the viewer around 8 frames later than the actor ($\simeq 533ms$ which corresponds with the findings of [13, 1]. This difference in frames/time will later on be referred as time-shift.

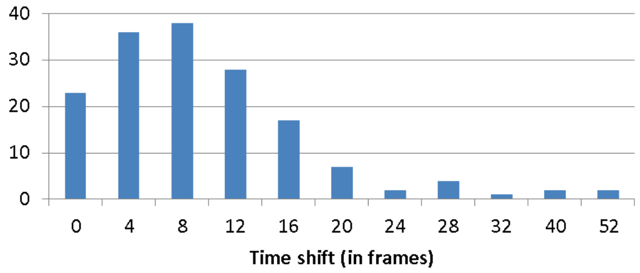

Fig. 1: Histogram displaying the differences of frames between the viewer's and actor's focus on a new action

AUC/NSS	AUC/PCC	PCC/NSS
0.996	0.997	1.0

Table 1: Correlation scores between the three different metrics for section 2.3.4

2.3.4. *Comparison of Actor's and Viewer's saliency maps*

After looking at the previous manual annotation results (1) confirming our expectations one can wonder weather this time-shift phenomenon is still observable when comparing two subjective saliency models. Based on the three metrics described in 2.2 we compared the similarity of saliency maps between actors and viewers computed using the method introduced in 2.1 for the frames belonging to the beginning of actions. The corresponding results are given by Figure 2. The AUC scores are displayed for different values of time-shift between actors' (fixed) and viewers' (varying in time) saliency maps. The NSS and PCC metrics are not displayed since the scores are highly correlated with AUC: see table 1. The computation of these three metrics clearly brings to the same conclusion pointed out in 2.3.3: the actors' saliency maps show more correspondence with those of the viewers when the latter are considered with a time-shift. An also noticeable and expectable result to be extracted from this figure is that the standard deviation (grey bars) gets lower when the correspondence score gets higher (around 14 frames $\simeq 933ms$ time-shift).

3. ADAPTATION OF OBJECTIVE SALIENCY MAPS TO RETRIEVE THE ACTOR'S SALIENCY MAPS

In the current literature, all the automatic saliency maps models are proposed aiming to approach the viewer's one in the best manner. Sections 2.3.3, 2.3.4 have both by manual and automatic calculations showed that the viewer's and actor's points of view are indeed more correlated when shifted in time. In this section we tackle a new problem: based on the previous results can we adapt the objective saliency maps automatically extracted from signals to match those of the actor?

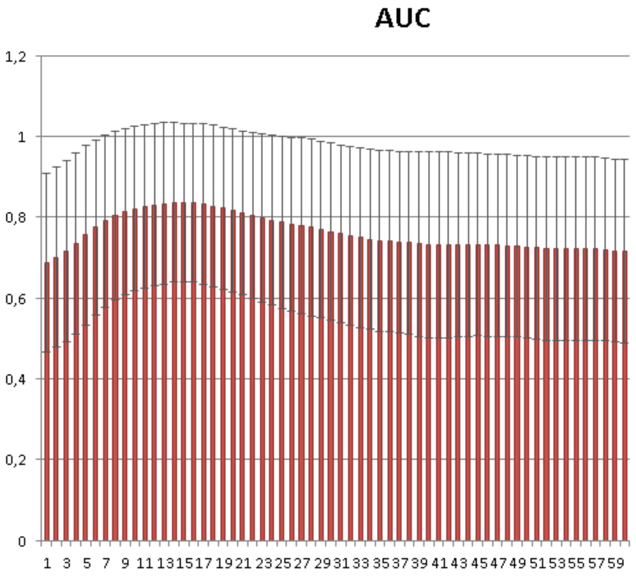

Fig. 2: AUC scores between actor's and viewer's saliency maps for different time-shifts (in frames)

3.1. Objective saliency maps

To delimit the area of video analysis in video frames to the regions which are potentially interesting to human viewers we need to model visual saliency on the basis of video signal features. Here we follow the spatio-temporal-geometric building process of subjective saliency maps as described in [6]. In this section we briefly introduce the three components of the overall objective visual saliency map proposed in [6]. Spatial saliency map is built from local color contrast features in each frame according to the method M.Z. Aziz et al. [14]. The features expressing the Hue, Saturation and Intensity contrast (HSI system) in each pixel are pooled together by a simple mean operator and normalized with regard to the maximal value in a frame. The temporal saliency map is estimated from the residual motion in the frame after compensation of global motion according to the complete first order affine model. In the temporal saliency computation we use the selectivity of human visual system with regard to the magnitude of residual motion, studied by S.J. Daly [15]. Indeed the low and strong magnitude of residual motion are thresholded and only residual motion vectors in the range of $]0°/s, 80°/s[$ degrees per second contribute to the map. This map represents the energy of residual motion vectors normalized by its maximum as well. Finally, in [6], a geometric saliency map construction is proposed in function of camera position on a given video corpus. In case of the camera worn on glasses, we can reasonably admit the so-called central bias hypothesis reported in [3, 16, 17]. Indeed the movement of the head constantly replaces the object of interest near the camera field center, since the video camera is attached to the

AUC/NSS	AUC/PCC	PCC/NSS
0.945	0.947	1.00

Table 2: Correlation scores between the three different metrics for section 3.3

glasses. Therefore, the gaze directed on the object of interest in the scene can be expressed by a simple Gaussian with half screen height spread centered on the frame center. Different fusion processes between these three cues have been applied for comparison based on [18, 12]. According to the previous research [6] in this work we use the squared Minkovsky pooling reinforced by multiplicative pooling, Eq (1):

$$S_{sp-t-g}^{sq}(t) = S_{sp}(t) \cdot S_t(t) \cdot S_g(t) + \frac{S_{sp}^2(t) + S_t^2(t) + S_g^2(t)}{3} \tag{1}$$

3.2. Time-shift based model adaptation

The previously presented objective saliency models have been designed to locate the areas of interest in videos. Since one can conclude based on the previous results (figures 1, 2) that actors have a tendency to focus on the areas of interest before the viewers, it is fair to assume that the automatic saliency maps can be adapted to match the actor's one for the beginning of actions by taking into account this shift in time. We firstly compared the AUC, PCC, and NSS scores when comparing different automatic saliency maps with either those of the viewers or the actors for the frames corresponding to the beginning of actions. According to the results in figure 2 where the highest score is computed with a time-shift of 14 frames ($\simeq 933ms$), we then computed the same metrics scores when comparing actors' saliency maps with the automatic ones shifted by 14 frames.

3.3. Results

Results of the AUC scores computed for the three different comparisons described in section 3.2 are showed in figure 3. As for section 2.3.4, the NSS and PCC scores are not displayed since highly correlated with AUC (see table 2)

Firstly, as can be expected, we can see the difference of scores when comparing the automatic saliency maps to the actor's one versus the viewer's. The results demonstrate that the objective maps correlate more with the viewer indeed, the scores of correspondence with the actor's being low.

Another important point is that the automatic saliency maps showing the highest scores for all three different comparisons is the square with geometric model introduced in 2.3.4 by Eq (1).

Finally the results obtained with the new automatic Time-shift based model (section 3.2) display higher scores indeed when compared to the subjective actor's saliency maps.

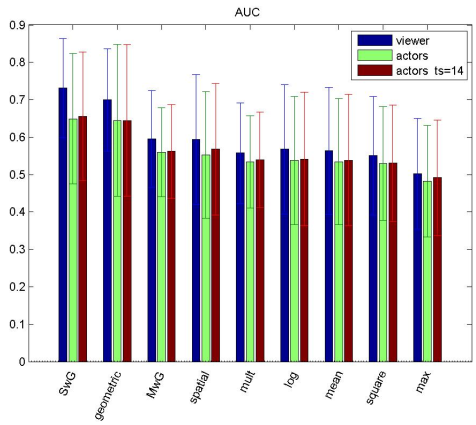

Fig. 3: AUC scores for the comparison between different models of automatic saliency maps (SwG stands for square with geometric, MwG stands for multiplication with geometric, mult stands for multiplication). In blue: viewers vs automatic, in green: actors vs automatic, in red: actors vs the time-shifted adapted model of automatic saliencies.

4. CONCLUSION AND PERSPECTIVES

Hence in this paper we proposed a new approach for prediction of visual saliency maps in the upcoming "egocentric" video content from the cameras worn by persons. Accordingly to the research results in vision and motor control we formulated the assumption of temporal shift of visual saliency between the person executing different activities, i.e. Actor and the Viewer who interprets this content a posteriori. Psychovisual experiments confirm this assumption. Based on these results we proposed "à la carte" prediction of visual saliency maps by an objective model we previously developed completing it by temporal integration in the case of Actor's maps. This research, we believe, opens a new and exciting perspective of interpretation of visual content from various points of view. This is necessary in various medical and physiological studies, specifically for neurodegenerative diseases such as Alzheimer and Parkinson disease, but also could be applied to the content coming from more wide ranges of mobile devices. In the immediate perspective of this work we see the fine-grain studies of activities to refine the prediction results according to the activities taxonomy.

5. ACKNOWLEDGMENTS

This research is supported by the EU FP7 PI Dem@Care project #288199. We also would like to thank M.L. Latash - professor of kinesiology, PSU - for fruitful discussions on gaze-movement interaction.

6. REFERENCES

[1] C.C. Gonzalez and M.R. Burke, "The brain uses efference copy information to optimise spatial memory," *Experimental Brain Research*, vol. 224, pp. 189–197, 2013.

[2] C. Prablanc, J.F. Echailler, E. Komilis, and M. Jeannerod, "Optimal response of eye and hand motor systems in pointing at a visual target," *Biol. Cybernetics*, vol. 35, pp. 113–124, 1979.

[3] Michael Dorr, Thomas Martinetz, Karl R. Gegenfurtner, and Erhardt Barth, "Variability of eye movements when viewing dynamic natural scenes.," *Journal of vision*, vol. 10, no. 10, 2010.

[4] O. Le Meur, P. Le Callet, and D. Barba, "Predicting visual fixations on video based on low-level video features," *Vision Research*, vol. 47, no. 19, pp. 1057–1092, Sept 2007.

[5] P. LeCallet, O Lemeur, D. Barba, and D Thoreau, "Bottom - up visual attention modeling: Quantitative comparison of predicted salience maps with observers eye-tracking data," *ECVP - European Conference on Visual Perception*, 2004.

[6] Hugo Boujut, Jenny Benois-Pineau, and Remi Megret, "Fusion of multiple visual cues for visual saliency extraction from wearable camera settings with strong motion," in *Computer Vision – ECCV 2012. Workshops and Demonstrations*, Andrea Fusiello, Vittorio Murino, and Rita Cucchiara, Eds., vol. 7585 of *Lecture Notes in Computer Science*, pp. 436–445. Springer Berlin Heidelberg, 2012.

[7] Alireza Fathi, Yin Li, and JamesM. Rehg, "Learning to recognize daily actions using gaze," in *Computer Vision – ECCV 2012*, Andrew Fitzgibbon, Svetlana Lazebnik, Pietro Perona, Yoichi Sato, and Cordelia Schmid, Eds., vol. 7572 of *Lecture Notes in Computer Science*, pp. 314–327. Springer Berlin Heidelberg, 2012.

[8] David Wooding, "Eye movements of large populations: Ii. deriving regions of interest, coverage, and similarity using fixation maps," *Behavior Research Methods*, vol. 34, pp. 518–528, 2002, 10.3758/BF03195481.

[9] Andrew T. Duchowski, *Eye Tracking Methodology: Theory and Practice, Second Edition*, Springer-Verlag London Limited, 2007.

[10] Donald C. Hood and Marcia A. Finkelstein, "Sensitivity to light," in *Handbook of perception and human performance, Volume 1: Sensory processes and perception*, K. R. Boff, L. Kaufman, and J. P. Thomas, Eds., chapter 5, pp. 5–1–5–66. John Wiley & Sons, New York, NY, 1986.

[11] O. Le Meur and T. Baccino, "Methods for comparing scanpaths and saliency maps: strengths and weaknesses.," *Behav Res Methods*, 2012.

[12] Sophie Marat, Tien Ho Phuoc, Lionel Granjon, Nathalie Guyader, Denis Pellerin, and Anne Guérin-Dugué, "Modeling spatio-temporal saliency to predict gaze direction for short videos," *International Journal of Computer Vision*, vol. 82, no. 3, pp. 231–243, 2009, Département Images et Signal.

[13] M. F. Land and M. Hayhoe, "In what ways do eye movements contribute to everyday activities?," *Vision research*, vol. 41, no. 25-26, pp. 3559–3565, 2001.

[14] M.Z. Aziz and B. Mertsching, "Fast and robust generation of feature maps for region-based visual attention," *IEEE Transactions on Image Processing*, vol. 17, no. 5, pp. 633 –644, 2008.

[15] S. J. Daly, "Engineering observations from spatiovelocity and spatiotemporal visual models," in *IS&T/SPIE Conference on Human Vision and Electronic Imaging III*, 1 1998.

[16] D. Parkhurst, K. Law, and E. Niebur, "Modeling the role of salience in the allocation of overt visual attention," *Vision Research*, vol. 42, pp. 107–123, 2002.

[17] Benjamin W. Tatler, "The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions," *Journal of Vision*, vol. 7, no. 14, pp. 1–17, Nov. 2007.

[18] H. Boujut, J. Benois-Pineau, T. Ahmed, O. Hadar, and P. Bonnet, "A metric for no-reference video quality assessment for hd tv delivery based on saliency maps," in *Multimedia and Expo (ICME), 2011 IEEE International Conference on*, july 2011, pp. 1 –5.