InImpact: The Journal of Innovation Impact: ISSN 2051-6002 : http://www.inimpact.org
Special Edition on Innovation in Medicine and Healthcare : Vol. 6. No. 1 : pp.74-78 : imed13-024

Visual saliency maps for studies of behavior of patients with neurodegenerative diseases: Observer's versus Actor's points of view

Hugo Boujut[1], Vincent Buso[1], Jenny Benois-Pineau[1], Yann Gaestel[2], Jean-François Dartigues[2]

[1]Laboratoire Bordelais de Recherche en Informatique (LaBRI), University of Bordeaux 1, France
[2] ISPED, U897 INSERM, Bordeaux, France

Abstract—We are interested in finding the relation between the visual saliency maps of the viewer of visual content and the actors (person executing the actions) in the context of studies of neurodegenerative diseases such as Alzheimer's disease. From results of eye-trackers worn by the actors and used when recording observers, and on the basis of hand-eye interactions from motor control studies we established a time shift between actor's and viewer's saliency maps. This time shift corresponds to the latency of hand-eye interaction. The method is based on adequate normalization of saliency maps and computation of similarity metrics for pixel based saliency. This finding gives good perspectives for automatic prediction of a normal actor saliency map from observer saliency map.

.

Keywords—Visual attention, visual saliency, neurodegenerative diseases.

1. Introduction and Motivation

Computer vision tools become more and more popular in the context of studies of behavior of patients attained by neurodegenerative diseases [1]. Taking into account that these patients are fragile, the measurement devices have to be as less intrusive as possible. In the ideal case, the studies of behavior of such patients have to be conducted in the ecological situation, when they are performing required activities in their usual non-stressful environment.

Since recently a new video content is massively coming into practice: the egocentric video recorded by body-worn cameras. The research conducted in the context of Alzheimer's patients behavior studies with these devices [2] showed that such patients easily accept the device and even forget it. In a study of difficulties of spatial vision and also of visio-motor coordination we are interested in visual attention of the patients, which in computer vision domain can be called "subjective visual saliency". This term designates a visual attention map which can be built in the video frames recorded with a wearable camera corresponding to the patient's point of view. In the present paper we make a pioneering attempt to establish a relationship between saliency maps of the test subject – that is a person who executes the required activities and the saliency map of the person who observes recorded video a posteriori, using videos recorded with wearable cameras. We will further call these two parties an "Actor" and a "Viewer". The motivation for this research is as follows.

In computer vision, automatic prediction of visual attention maps of humans observing visual content from image signal has been a well studied subject since the early works by Itti [3].

Automatic prediction of Viewer visual attention maps is therefore possible. Then, in order to predict the "normal" visual attention map for the purpose of comparing it with the recorded visual attention map of a tested patient, we first need to establish the relationship between two visual attention maps: those of an Actor and of a Viewer using video content recorded with wearable cameras.

The visual saliencies of Actor and Viewer are not the same. Indeed according to the physiological studies [4] [5], the human gaze anticipates the motor action of limbs when fulfilling an activity. Hence in this paper we study the relationship between these two visual attention subjective saliency maps

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 [6].
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. Experimental setup and results are presented in Section 3. 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), and 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 [7] has become the reference [8] since it fulfills 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 [9]. 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 [10] about them. From this survey and anterior work [11] 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. The PCC is a straightforward application of statistical analysis. It was used in several studies and is particularly adapted for comparison of pixel-based saliency maps. The NSS is a Z-score that compares two scanpaths. This method is widely used in the research community since it is suitable for scanpath and pixel-based saliency maps. Finally, The AUC is also a popular metric in the research community but is suitable only for pixel-based saliency map comparison. 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 labeled 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.
Since results prove the scores to be highly correlated between these metrics (Tableau 1), only the AUC is displayed in this paper.

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-tracker recording of viewers' gazes are explained before comparing the results of these two methods.

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 [6], 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 640x480 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.

3.2. Eye tracker setup

In order to get the eye location of the people watching the videos provided by the authors of [6], 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 [11]. 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.

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 whether 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 2. From this histogram one can clearly notice a peak of time difference between the realizations of actions from 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 [12] [13]. This difference in frames/time will later on be referred as time-shift.

Tableau 1 - Correlation scores between the three different metrics

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

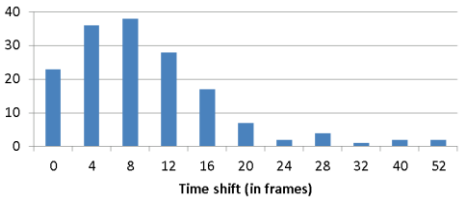

Figure 2-Histogram displaying the differences of frames between the viewer's and actor's

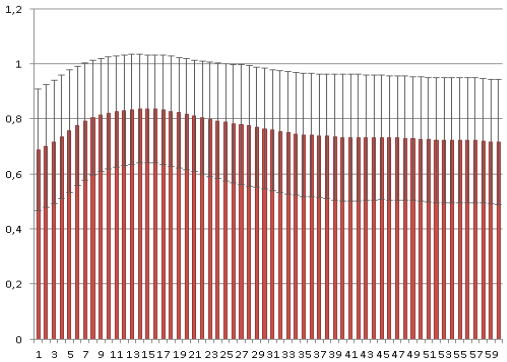

Figure 1 - AUC scores between actor's and viewer's saliency maps for different time-shifts (in

3.4. Comparison of Actor's and Viewer's saliency maps

After looking at the previous manual annotation results (Figure 2) confirming our expectations one can wonder whether this time-shift phenomenon is still observable when comparing two subjective saliency models.
Based on the three metrics introduced in 2.2 we compared the similarity of saliency maps between actors and viewers computed using the method introduced in 1 for the frames belonging to the beginning of actions.
The corresponding results are given by Figure 1. 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 Tableau 1) meaning the histograms have the same shape for all three metrics. This brings to the same conclusion pointed out in 3.3: the actors' saliency maps show more correspondence with those of the viewers when the latter are considered with a time-shift.

CONCLUSION AND PERPECTIVES

Hence in this paper we were interested in the relation between Actor's and Viewer's subjective saliency maps in egocentric video. The rationale of this research was egocentric video understanding in large-scale epidemiological studies of behavior of patients with Alzheimer disease. As the Actor's saliency map is never available in such a context, all automatic saliency maps are usually optimized with regard to the Observer's saliency map. The analysis of video content from Actor's point of view is one of the points which are of interest for medical practitioners. Furthermore, some fundamental computer vision problems such as recognition of manipulated objects in such new emerging content from Actor's point of view is an interesting question to address. Based on the studies of hand-eye interactions in pointing tasks in lab environment and in everyday life action studies in natural environment for some activities, we emitted the hypothesis of time shift between these two saliency maps. We showed that indeed this time shift does exist. Based on these results the immediate perspective of this work would be to develop an optimization of the state of the art automatic saliency maps building algorithms using this temporal shift in order to find mathematical ways to recreate those of the Actors.

ACKNOWLEDGMENT

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.

REFERENCES

[1] S. Verheij, D. Muilwijk, J. J. Pel, T. J. van der Cammen, F. U. Mattace-Raso and J. van der Steen, "Visuomotor Impairment in Early -Stage Alzheimer's Disease: Changes in Relative Timing of Eye and Hand Mouvements," *Journal of Alzheimer's Disease,* no. 30, pp. 131-143, 2012.

[2] S. Karaman, J. Benois-Pineau, R. Megret, V. Dovgalecs, J.-F. Dartigues et Y. Gaestel, «Human Daily Activities Indexing in Videos from Wearable Cameras for Monitoring of Patients with Dementia Diseases,» chez *ConferenceProceedings of the 2010 20th International Conference on Pattern Recognition*, Washington, DC, USA, 2010.

[3] L. Itti, C. Koch et E. Niebur, «A Model of Saliency-Based Visual Attention for Rapid Scene Analysis,» *IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE,* vol. 20, n° %111, pp. 1254-1259, 1998.

[4] C. Prablanc, J. F. Echailler, E. Komilis et M. Jeannerod, «Optimal response of Eye and Hand Motor systems in Pointing at a Visual Target,» *Biol. Cybernetics,* vol. 35, pp. 113-124, 1979.

[5] M. Dorr, T. Martinetz, K. R. Gegenfurtner et E. Barth, «Variability of eye movements when viewing dynamic natural scenes.,» *Journal of vision,* vol. 10, n° %110, 2010.

[6] A. Fathi, Y. Li et J. Rehg, «Learning to Recognize Daily Actions Using Gaze,» vol. 7572, A. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato et C. Schmid, Éds., Springer Berlin Heidelberg, 2012, pp. 314-327.

[7] D. Wooding, «Eye movements of large populations: II. Deriving regions of interest, coverage, and similarity using fixation maps,» *Behavior Research Methods,* vol. 34, n° %14, pp. 518-528, 2002.

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

[9] D. C. Hood et M. A. Finkelstein, «Sensitivity to Light,» K. R. Boff, L. Kaufman et J. P. Thomas, Éds., New York, NY, John Wiley & Sons, 1986, pp. 5-1.

[10] O. Le Meur et T. Baccino, «Methods for comparing scanpaths and saliency maps: strengths and weaknesses.,» *Behav Res Methods,* 2012.

[11] S. Marat, T. Ho Phuoc, L. Granjon, N. Guyader, D. Pellerin et A. Guérin-Dugué, «Modeling spatio-temporal saliency to predict gaze direction for short videos,» *International Journal of Computer Vision,* vol. 82, n° %13 , pp. 231-243, 2009.

[12] M. F. Land et M. Hayhoe, «In what ways do eye movements contribute to everyday activities?,» *Vision research,* vol. 41, n° %125-26, pp. 3559-3565, 2001.

[13] C. C. Gonzalez et M. R. Burke, «The brain uses efference copy information to optimise spatial memory,» *Experimental Brain Research,* vol. 224, n° %12, pp. 189-197, 2013.