



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

Hugo Boujut, Vincent Buso, Jenny Benois-Pineau, Yann Gaëstel,
Jean-François Dartigues

► To cite this version:

Hugo Boujut, Vincent Buso, Jenny Benois-Pineau, Yann Gaëstel, Jean-François Dartigues. Visual saliency maps for studies of behavior of patients with neurodegenerative diseases: Observer's versus Actor's points of view. *Innovation in Medicine & Healthcare*, Jul 2013, Piraeus, Greece. 5p. hal-00853853

HAL Id: hal-00853853

<https://hal.science/hal-00853853>

Submitted on 23 Aug 2013

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

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

Hugo Boujut¹, Vincent Buso¹, Jenny Benois-Pineau¹, Yann Gaestel², Jean-François Dartigues²

¹Laboratoire Bordelais de Recherche en Informatique (LaBRI), University of Bordeaux 1, France

²ISPED, U897 INSERM, Bordeaux, France

Abstract—Finding the salient regions in videos has been a very active topic. In this work we compare the modelisation of visual attention on egocentric video recordings for two different points of view. 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). This question is of importance because the building of objectives saliency maps has always been optimized with regard to the viewer's. However, its relation with actor's saliency map has never been considered in the past. 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 adjustment of observer saliency map to the actor saliency map.

Keywords—

1. INTRODUCTION

Computer vision approaches become more and more popular in the context of studies of behavior of patients attained by neurodegenerative diseases.

Since recently a new video content is massively coming into practice: the egocentric video recorded by body-worn 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. This research is pioneery in the sense that we try to provide medical practitioners with a tool to automatically predict saliency maps of "normal" actors for the future automatic comparisons with saliency maps of patients attained by neurodegenerative diseases.

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 [4].

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), 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 [5] has become the reference [6] 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 [7]. 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 [8] about them. From this survey and anterior work [9] 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.

of dependency of signals. 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 [4], 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. The different noticeable actions related with the preparation of a meal (e.g. spread jam, take milk...) are listed in [4].

3.2. Eye tracker setup

In order to get the eye location of the people watching the videos provided by the authors of [4], 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 [9]. 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 1. 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 ($\approx 533ms$) which corresponds with the findings of [10] [1]. This difference in frames/time will later on be referred as time-shift.

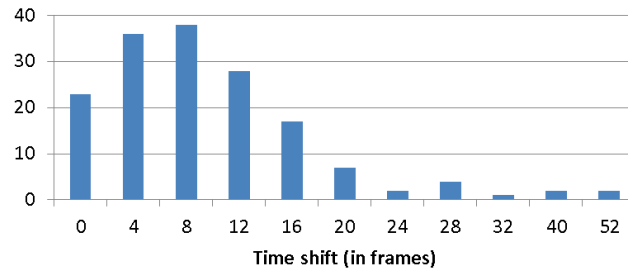


Figure 1-Histogram displaying the differences of frames between the viewer's and actor's focus on a new action

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

After looking at the previous manual annotation results (Figure 1) 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 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 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.

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 $\approx 933ms$ time-shift).

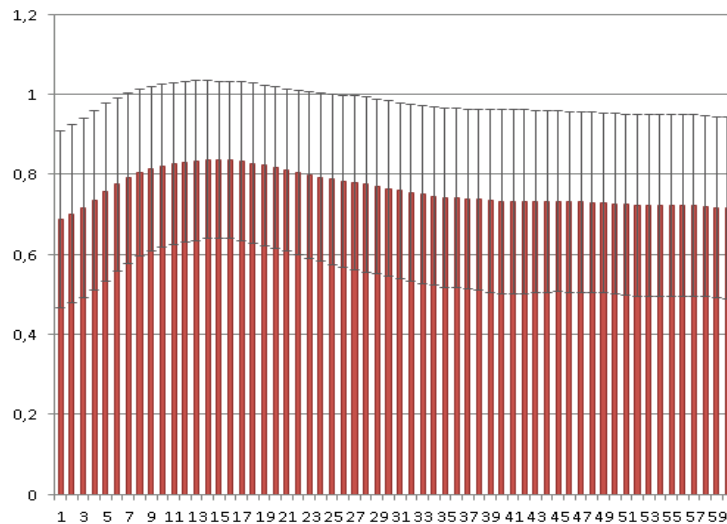


Figure 2 - AUC scores between actor's and viewer's saliency maps for different time-shifts (in frames)

Tableau 1 - Correlation scores between the three different metrics

AUC/NSS	AUC/PCC	PCC/NSS
---------	---------	---------

0.996	0.997	1.0
-------	-------	-----

CONCLUSION AND PERSPECTIVES

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 a 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] 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.
- [2] 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.
- [3] 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.
- [4] 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, Éd., Springer Berlin Heidelberg, 2012, pp. 314-327.
- [5] 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.
- [6] A. T. Duchowski, *Eye Tracking Methodology: Theory and Practice*, Second Edition, Springer-Verlag London Limited, 2007.
- [7] D. C. Hood et M. A. Finkelstein, «Sensitivity to Light,» K. R. Boff, L. Kaufman et J. P. Thomas, Éd., New York, NY, John Wiley & Sons, 1986, pp. 5-1.
- [8] O. Le Meur et T. Baccino, «Methods for comparing scanpaths and saliency maps: strengths and weaknesses.,» *Behav Res Methods*, 2012.

- [9] 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.
- [10] M. F. Land et M. Hayhoe, «In what ways do eye movements contribute to everyday activities?,» *Vision research*, vol. 41, n° 25-26, pp. 3559-3565, 2001.