Second Screen Interaction: an Approach to Infer TV Watcher's Interest Using 3D Head Pose Estimation

Julien Leroy
UMONS
31 Boulevard Dolez
Mons, Belgium
julien.leroy@umons.ac.be

Francois Rocca
UMONS
31 Boulevard Dolez
Mons, Belgium
francois.rocca@umons.ac.be

Matei Mancas
UMONS
31 Boulevard Dolez
Mons, Belgium
matei.mancas@umons.ac.be

Bernard Gosselin
UMONS
31 Boulevard Dolez
Mons, Belgium
bernard.gosselin@umons.ac.be

ABSTRACT

In this paper, we present our ”work-in-progress” approach to implicitly track user interaction and infer the interest a user can have for TV media. The aim is to identify moments of attentive focus, noninvasively and continuously, to dynamicaly improve the user profile by detecting which annotated media have drawn the user attention. Our method is based on the detection and estimation of face pose in 3D using a consumer depth camera. This allows us to determine when a user is or not looking at his television. This study is realized in the scenario of second screen interaction (tablet, smartphone), a behavior that has become common for spectators. We present our progress on the system and its integration in the LinkedTV project.

Categories and Subject Descriptors

H.5 [**Information interfaces and presentation**]: User Interfaces

General Terms

Experimentation

Keywords

attention, head pose estimation, second screen interaction

1. INTRODUCTION

Using a second screen while watching a TV has become a more and more common behavior. These companion device are there to give you more interactivity with the broadcast media. In the LinkedTV project [1], one of our goals is to offer new possibilities for personalization of content provided to users, including the implicit analysis of human behavior. To do this, one of the tracks that we explore is the possibility of detecting moments of attention focus on the user's screens. This information is important because it can tell us when, what and how the media interest a user. Which ultimately will allow us to modify the profile of the user based on their viewing preferences without him having explicitly requested. To achieve this, we chose to base our solution on the head detection and pose estimation using a depth camera. This choice was made due to the democratization of this type of sensors and their arrival in the home through gaming platform [15]. TV manufacturers are slowly beginning to integrate cameras into their new systems, regarding the sensors we can see the willingness of the maker to miniaturize sensors such as PrimeSense new camera ”Capri” [17]. Thus, we can expect to see in the coming years 3D sensors directly integrated into televisions. In this context, we decided to develop as experimental setting the interaction with the second screen and to detect when a user switches his attention focus between the main and second screen.

2. RELATED WORKS

A simple movement of the head can be representative of many things: a person can point excessively his head to indicate an interesting element, nod to show his approval, tilt to express his perplexity, rapid movements will in turn represent a sign of surprise or alarm, etc. Movement and orientation of the head are important non-verbal cues that can convey rich information about a person's behavior and attention [18][12]. While estimating the orientation of the human head seems to be a simple task, in the context of computer vision, it is a complex problem that challenges researchers for decades without that a perfect solution emerges. Until recently, the literature has mainly focused on the automatic estimation of the poses based on standard images or videos. One of the major issues that must be addressed to obtain a good estimator is to be invariant to variables such as: camera distortions, light sources ... Other variables are directly related to the head like the shape, facial expressions, accessories (glasses). Many techniques have been developed over the years such as: appearance template methods, detector array methods, non linear array methods, manifold regression methods, flexible methods, geometric method, tracking method and hybrid methods. More information on these methods can be found in [16]. More recently, with the ap-

[1] LinkedTV EU project, http://www.linkedtv.eu

WWW 2013 Companion, May 13–17, 2013, Rio de Janeiro, Brazil.
ACM 978-1-4503-2038-2/13/05.

pearance of low cost depth sensor, more accurate solutions have emerged [6][8]. Based on the use of depth maps, those methods are able to overcome known problems on 2D images as illumination. In addition, they greatly simplify the spatial positioning of the head with a global coordinate system directly related to the metric of the analyzed scene. Many of these techniques are based on a head tracking method which unfortunately often requires initialization and also undergoes a drift. Another approach, based on the frame to frame analysis as the method developed by [9], provides robust and impressive results. It is this method that we base our approach on. Indeed, it is well suited to the TV scenario and the second screen. It is robust to illumination conditions that can be very variable in this case (dim light, television only source of light, ...) but is based on a type of sensor like the Microsoft Kinect.

3. HEAD POSE ESTIMATION

Given the technical limitations we impose to stick to a scene of interaction with a television as realistic as possible, it is not possible to access the orientation of a person's eyes. This is mainly due to a too large distance between the TV and the user, the gaze normally consists of two components: it is a combination of both the direction of the eyes and the pose of the head. But as we can't access the eyes, one of the hypotheses on which we rely to detect changes in visual focus is that, at the TV setup distance (more than two meters from the main screen), the gaze of a person is considered to be similar to the direction of his head. As stated in [16], "[...]*Head pose estimation is intrinsically linked with visuel gaze estimation ... By itself, head pose provides a coarse indication of gaze that can be estimated in situations when the eyes of a person are not visible*[...]". Several studies rely and validate this hypothesis as shown in [1]. Therefore, we will detect visual attention switch and attention focus by studying the behavior of the orientation of the head of a person. Like said, the analysis system is based on the detection and pose estimation of heads on a depth map. Our goal is to achieve a head tracking in real time and estimate the six degrees of freedom (6DOF) of the detected head (spatial coordinates, pitch, yaw and roll). The advantage of such a system is that it uses only geometric information and is independent of the brightness. It can operate in the dark, which is rarely possible with face tracking systems working on color image which are highly dependent on the illumination. This approach was chosen because it fits well in the scenario of TV interaction. In addition, the use of 3D data will simplify the integration of future contextual information about the scene. The method used is based on the approach developed in [7][10] and implemented in the PCL library [2]. This solution relies on the use of random forest regression. Random forests [3] are a very popular technique in computer vision for their capability of handling large training sets, high power of generalization and fast computing time. In our case the random forest are extended by using a regression step. This allows us to simultaneously detect faces but also to estimate their orientations on the depth map. The method consists of a training stage during which we build the random forest and an on-line detection stage where the patches extracted from the current frame are classified using the trained forest. The training process is done once an ti is not requested for any user. The training stage is based on the BIWI dataset [10] containing over 15000 images of 20 people (6 females and 14 males). This dataset covers a large set of head pose (+-75 degrees yaw and +-60 degrees pitch) and generalizes the detection step. A leaf of the trees composing the forest stores the ratio of face patches that arrived to it during training as well as two multi-variate gaussian distributions voting for the location and orientation of the head. A second processing step can be applied, it consists in registering a generic face cloud over the region corresponding to the estimated position of the head. This refinement step can greatly increase the accuracy of the head tracker but requires more computing resources.

4. EXPERIMENT

4.1 Scene description

Our experimental setting consists of:

- a 46 inches HD TV,
- a sofa, located at 2.5m from the television,
- a 3D camera positioned at 80 cm from the sofa and low enough to not obstruct the field of vision of the viewer,
- a 10 inches tablet that plays the role of a second screen.

These parameters allow us to calibrate our tracking system and reconstruct a simplified virtual 3D scene.

4.2 Scenario

The studied scenario consists in detecting the moments of transission between the interaction with the second screen and the broadcast media on the TV. To do this, we asked participants to solve various puzzles on a tablet with increasing difficulty to keep them focus on the second screen like on the Fig. 1. The broadcast media is a zapping, a series of short clips on news, sports, politics, buzz, etc...

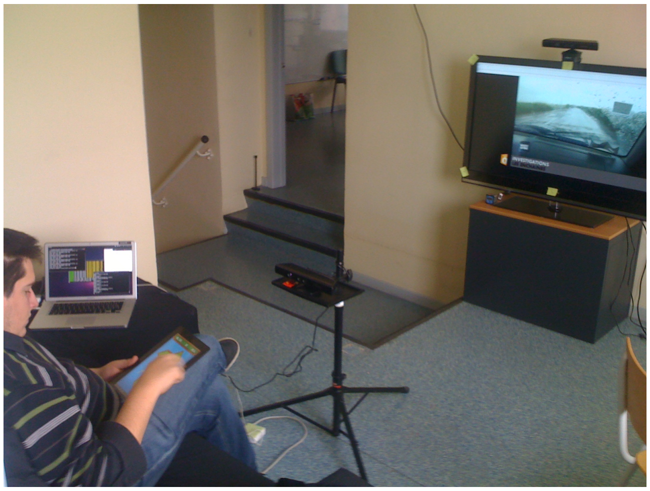


Figure 1: Setup of the experiment with the user playing a puzzle game on the second screen while a TV show is broadcast on the main screen. The camera in the middle of the scene tracks head movements (looking at the main screen or not).

4.3 Performance

Two operating options are developed. In our experimental setting, the data are processed offline, which allows firstly to enable the functions of refinement (precision mode) to dramatically increase the accuracy of temporal and spatial observations. Each frame can be treated at 1.5 frames/sec on a Macbook Pro with an Intel Core i5 2.53GHz. Online operation can be envisaged without refinement (coarse mode), in this case performances can reach above 10 frames/sec.

4.4 System

To detect if a user watches TV or not, we reconstruct a virtual simplified model of the real scene. Therefore, knowing the 6DOF position of the face of the person detected, it is possible to estimate the point of intersection between the TV virtual model and the orientation of the head. In this way, we can synchronize annotated media with the head tracker and estimate (+- 10 cm, on our 46" TV in precision mode) where the user is looking. With a bit more of precision and stability, it would be possible to determine which part of the scene attracts his interest.

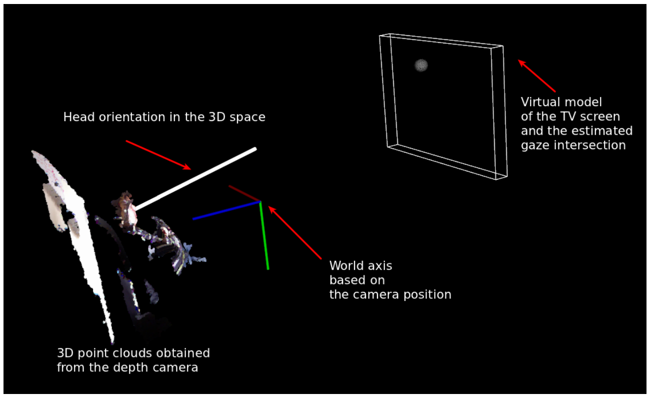


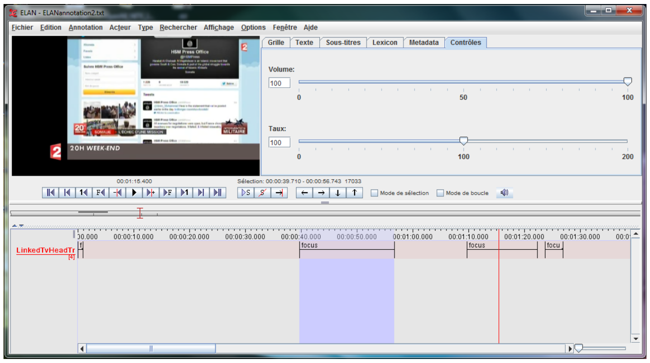
Figure 2: 3D rendering of our system. On the left, we can observe the 3D point cloud obtained with the depth camera. The head pose estimation algorithm is applied on this cloud, if a face is detected, we retrieve a vector of the head direction and compute an estimation of where the user is watching on the virtual screen.

4.5 Observation

Based on the detection of focus on the main screen, we can have a clue about what parts of the media have attracted the substained viewer attention. This can be done by measuring the duration of a fixation on the screen. For such events, we will weigh differently parts of the media: a greater weight for the part that attracts attention, a lower weight to the parts viewed when the gaze is steady and finally a negative weight for the part that causes loss of attention. In a second phase, a more detailed analysis, based on the actual behavior of screen transition can be achieved and give us information about the stimilus that caused this change. Two behaviors should be highlighted:

- The classical case is that the user's attention is drawn from the second screen and stays focused on the main screen for a long time.
- The user has the attention drawn on the main screen but quickly returns to the second screen. In this case, it is interesting to note the attentive mechanisms that have been taking place. A very short and abrupt shift of attention may mean that the original stimulus is bottom-up. Typically, such a response can be observed when a flash of light or a surprising sound occurs on the screen but has little semantic interest to the user. If the attention is substained, it's more likely due to the fact semantic (or top-down) attention is needed. By distinguishing this behavior, we will allocate a smaller weight to the event than if the stimulation was of top-down origin that involves a more important an attentive and cognitive process, as recognizing specific voices and faces or known subjects.

To realize these observations, the system can export the tracking information in different file formats to use annotation software like ELAN [19] Fig. 3. Based on this, we build a database of characteristic behaviors.


ELAN - ELANannotation2.txt

Volume: 100

Taux: 100

00:01:15.400

Sélection: 00:00:39.710 - 00:00:56.743 17033

LinkedTvHeadTr [4]: focus, focus, focu

Figure 3: Integration of the head tracker observation into the ELAN interface and the synchronised media

4.6 Integration into the LinkedTV framework

The data collected through our system is sent to a content personalization framework. To do this, at the end of each user session (example: when a user leaves the interaction zone), a report containing the tracking data will be sent as REST [11] query to the remote personalization module called GAIN (General Analytics INterceptor)[13] .

4.7 Future works

First, we will improve the overall performance of our system with a new training set to extend the distance of detection and the integration a 2D face recognition module that should enable us to facilitate the tracking of multiple users and accelerate the processing by limiting areas of the cloud to analyze. In a second step, we will focus on the discremination of bottom-up and top down attention. As previously stated two types of stimuli can attract the attention of a person: top-down and bottom-up. The top down, in this case, is more important because it demonstrates a cognitive interest for the broadcast media content. It is therefore important to detect attention switches due to top-down stimuli. Various studies have been conducted [14][4][5] and they show that it is possible by classifying head trajectories, based on their speed and amplitude, to distinguish attention switch due to bottom-up stimuli and those due to top-down information. In a third step, we will deal with the study of the concept of joint attention in front of the television. The headtracker lets us track more than one person (Fig. 4) so we can make some assumptions on where the persons are looking and if they gaze at the screen if a part drawn the attention of all the users.

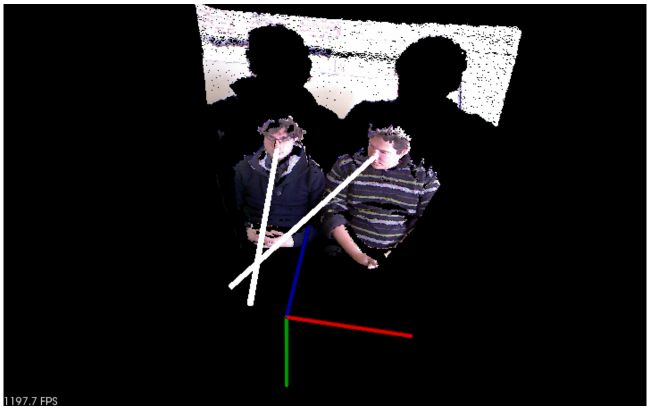


Figure 4: Multiple head detection and orientation estimation. We can detect if the users are possibly looking at the same thing. By analysing how their gaze cross, we can infer a process of joint attention.

5. CONCLUSIONS

In this paper, we presented the first stage of our implicit behavior analysis system based on a 3D head tracker. This tool is designed to feed a personalization framework capable of treating behavioral data to dynamically enhance the profile of a user. The first results show that it is possible to extract implicit information in an efficient way on where and when people look at their TV. In the future, additional information will be provided concerning joint attention and the kind of attention (bottom-up or top-down). This kind of information extracted by our system is very important in the study of media interest for the viewers and future TV personalization.

6. ACKNOWLEDGMENTS

This work is supported by the Integrated Project LinkedTV (www.linkedtv.eu) funded by the European Commission through the 7th Framework Programme (FP7-287911)

7. REFERENCES

[1] K. Abe and M. Makikawa. Spatial setting of visual attention and its appearance in head-movement. *IFMBE Proceedings*, 25/4:1063–1066, 2010.

[2] A. Aldoma. 3d face detection and pose estimation in pcl. September 2012.

[3] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[4] A. Doshi and M. M. Trivedi. Head and Gaze Dynamics in Visual Attention and Context Learning. pages 77–84, 2009.

[5] A. Doshi and M. M. Trivedi. Head and eye gaze dynamics during visual attention shifts in complex environments. 12:1–16, 2012.

[6] G. Fanelli, M. Dantone, J. Gall, A. Fossati, and L. Gool. Random Forests for Real Time 3D Face Analysis. *International Journal of Computer Vision*, 101(3):437–458, Aug. 2012.

[7] G. Fanelli, M. Dantone, J. Gall, A. Fossati, and L. Gool. Random forests for real time 3d face analysis. *International Journal of Computer Vision*, 101:437–458, 2013.

[8] G. Fanelli, J. Gall, and L. Van Gool. Real time head pose estimation with random regression forests. *Cvpr 2011*, pages 617–624, June 2011.

[9] G. Fanelli, J. Gall, and L. Van Gool. Real time 3d head pose estimation: Recent achievements and future challenges. In *Communications Control and Signal Processing (ISCCSP), 2012 5th International Symposium on*, pages 1–4, 2012.

[10] G. Fanelli, T. Weise, J. Gall, and L. V. Gool. Real time head pose estimation from consumer depth cameras. In *Proceedings of the 33rd international conference on Pattern recognition*, DAGM'11, pages 101–110, Berlin, Heidelberg, 2011. Springer-Verlag.

[11] R. T. Fielding and R. N. Taylor. Principled design of the modern web architecture. *ACM Trans. Internet Technol.*, 2(2):115–150, May 2002.

[12] A. Gaschler, S. Jentzsch, M. Giuliani, K. Huth, J. de Ruiter, and A. Knoll. Social behavior recognition using body posture and head pose for human-robot interaction. *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 2128–2133, Oct. 2012.

[13] T. K. Jaroslav KUCHAÅŸ. Gain: Analysis of implicit feedback on semantically annotated content. WIKT 2012, pages 75–78, 2012.

[14] A. Z. Khan, G. Blohm, R. M. McPeek, and P. Lefèvre. Differential influence of attention on gaze and head movements. *Journal of neurophysiology*, 101(1):198–206, Jan. 2009.

[15] Microsoft. Kinect sensor. http://www.xbox.com/kinect.

[16] E. Murphy-Chutorian and M. M. Trivedi. Head pose estimation in computer vision: a survey. *IEEE transactions on pattern analysis and machine intelligence*, 31(4):607–26, Apr. 2009.

[17] PrimeSense. Capri sensor. http://www.primesense.com/news/primesense-unveils-capri.

[18] A. Vinciarelli, M. Pantic, and H. Bourlard. Social signal processing: Survey of an emerging domain. *Image and Vision Computing*, 27(12):1743 – 1759, 2009.

[19] P. Wittenburg, H. Brugman, A. Russel, A. Klassmann, and H. Sloetjes. Elan: a professional framework for multimodality research. In *In Proceedings of Language Resources and Evaluation Conference (LREC)*, 2006.