3D Head Pose Estimation for TV setups

Julien Leroy, Francois Rocca, Matei Mancas, and Bernard Gosselin

University of Mons (UMONS), Faculty of Engineering (FPMs),
20, Place du Parc, 7000 Mons, Belgium

Abstract. In this paper, we present an architecture of a system which aims to personalize the TV content to the viewer reactions. The focus of the paper is on a subset of this system which identifies moments of attentive focus in a non-invasive and continuous way. The attentive focus is used to dynamically improve the user profile by detecting which displayed media or links have drawn the user attention. Our method is based on the detection and estimation of face pose in 3D using a consumer depth camera. Two preliminary experiments were carried out to test the method and to show its link to viewer interest. This study is realized in the scenario of a TV with a second screen interaction (tablet, smartphone), a behaviour that has become common for spectators.

Keywords: attention, head pose estimation, second screen interaction, eye tracking, Facelab, future TV, personalization

1 Introduction

One of the goals of future TV is to offer new possibilities for personalization of content provided to users, including the implicit analysis of human behaviour. To achieve the personalization goal several factors need to be taken into account: explicit interactions (pause, play, skip, click on a link, etc.), implicit interactions (looking to the TV or not) and context information (date, time, social networks, number of viewers, etc.). In this paper, we focus on implicit interaction and more specifically on a solution of head detection and pose estimation using a low-cost depth camera. This choice was made due to the democratization of this type of sensors and their arrival in the home through gaming platforms [17]. Moreover, TV manufacturers begin to integrate cameras into their new systems, regarding the sensors we can see the willingness of the makers to miniaturize sensors such as PrimeSense new camera "Capri" [21]. Thus, we can expect to see in the coming years 3D sensors directly integrated into televisions.
The next section provides information about the related work, section 3 details the implemented algorithm and two experiments. Section 4 relates the first results of the first experiment, while section 5 focuses on the second experiment. Section 6 provides some cues about the analysis of the results for media personalization and it is followed by the conclusion section.

2 Related work

Movement and orientation of the head are important non-verbal cues that can convey rich information about a person's behaviour and attention [24][12]. 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, illumination, face shape and expressions or features (glasses, beard). 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 [18]. More recently, with the arrival 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 or low contrast backgrounds. In addition, they greatly simplify the spatial positioning of the head with a global coordinate system directly related to the metric of the analysed 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. This method is well suited for a living room and TV scenario. It is robust to illumination conditions that can be very variable in this case (dim light, television only source of light, etc.) but is based on a 3D sensor like the Microsoft Kinect. The paper proposes a entire system of optimized head pose extraction.

3 Head pose estimation

3.1 Algorithm

The proposed system is based on the head detection and pose estimation on a depth map. Our goal is to achieve 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 a 3D system is that it uses only geometric information on the point cloud and is independent of the illumination issues which can dramatically change in front of a device like a TV. The proposed system can even operate in the dark or in rapidly varying light conditions, which is not possible with face tracking systems working on RGB images. In addition, the use of 3D data provide more stable results than 2D data which can be mislead by projections of the 3D world on 2D images.

Figure 1 shows the global pipeline of the head pose estimation sub-system. First, the 3D point cloud is extracted from a Kinect sensor using the PCL library [20]. In a second step people face is detected and localized (the blue larger boxes in Figure 1). Those boxes are computed from the head of the skeleton extracted from the depth maps by using the OpenNI library [19]. The skeleton head provides the 3D coordinates of the area where a face might be located. The smaller

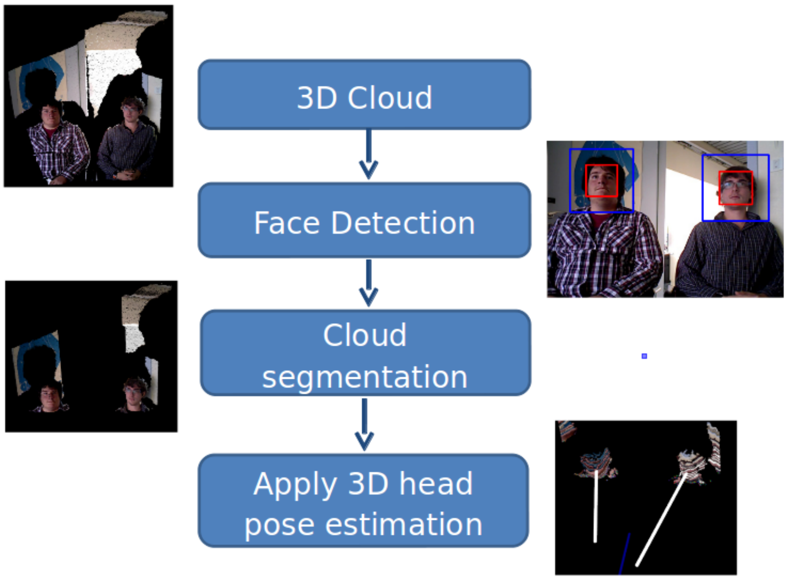

Fig. 1. Algorithm pipeline: 3D cloud extraction from the RGBD sensor, face localization and detection, 3D cloud segmentation and pose estimation.

red boxes are 2D face detection which can be used for face analysis, but this issue is not in the focus of this paper. Once the 3D position of the head is extracted, the 3D cloud is segmented to optimize the last 3D head pose estimation step. The segmentation eliminates a lot of the points of the 3D clouds where the chances to find a face are very low and therefore boosts the computational efficiency of the method.

The 3D head pose estimation used here follows the development in [16] which is improved by 4 in terms of computation time due to the 3D point cloud segmentation. The 3D pose estimation algorithm is based on the approach in [7][10] and implemented in the PCL library [2]. This solution relies on the use of a random forest [3] extended by a regression step. This allows us to detect faces and 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 only once and it is not user-dependent. One initial training is enough to handle multiple users without any additional configuration or re-training. This is convenient in a setup where a wide variety of people can watch TV. 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 poses (±75 degrees yaw and ±60 degrees pitch) and generalizes the detection step.

During the test 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. This step of the algorithm provides the head position and a rough head orientation on any new individual without the need of re-training. We then apply a final processing step which consists in registering a generic face cloud over the region corresponding to the estimated position of the head. This last step greatly stabilizes the final head position result.

3.2 Experiment

Our experimental setting consists of:

- a 46 inch HD TV,
- a sofa, located at 2.5m from the TV,
- 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 (Figure 4). The Kinect is located between the viewer and the TV which is not very convenient and it can be subject to viewer face occlusion when using second screen devices. Therefore the final setup will use the second generation Kinect which has a better resolution and should be capable to capture head motion when ideally located on top of the TV.
Within this setup, we performed two scenarios. The first one consists in detecting the head direction of a person watching TV in his living room. The idea is to discriminate between 1) watching TV, 2) watching the second screen (tablet, smartphone), 3) watching outside the TV, 4) watching out of the TV setup (no face detection but viewer detection). We asked participants to solve various puzzles on a tablet with increasing difficulty to keep them focused on the second screen like on the Fig. 2. The broadcast media is a zapping, a series of short clips of news, sports, politics, buzz, etc. In addition to this test, we also performed a second scenario. We used a commercial eye-tracking (Facelab 5 [23]) system which is able to measure both head direction and eye gaze direction. The eye-tracker was located at 1.80m from the TV screen and the viewer at 2.30m from the same TV as in experiment 1. The purpose of this second test was both to asses the 3D camera-based head detection, and also to have a first idea about the relationship between the head direction and eye direction.

4 Results of the first scenario: head from 3D camera

Each frame can be processed up to 8 frames/sec on a Macbook Pro with an Intel Core i5 2.53GHz. This speed is enough to extract head direction and basic features like direction change and speed. In addition, the algorithm proposed here also works on a recorded 3D video (.oni format). In this case the processing

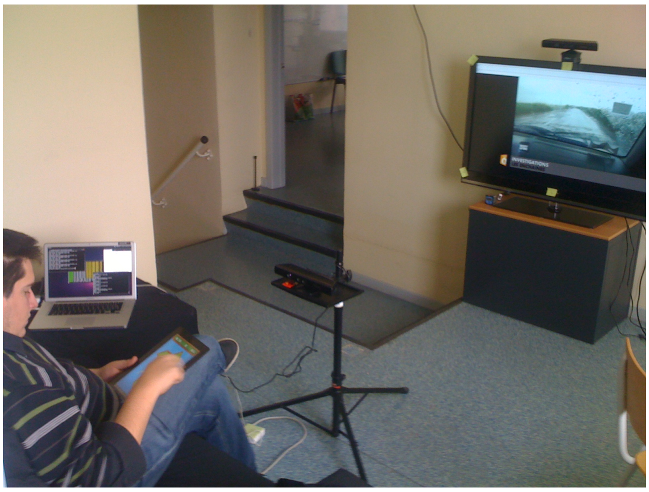

Fig. 2. Setup of the experiment with the user playing a puzzle game on the second screen (tablet) while a TV show is displayed on the main screen. The camera in the middle of the scene tracks head movements.

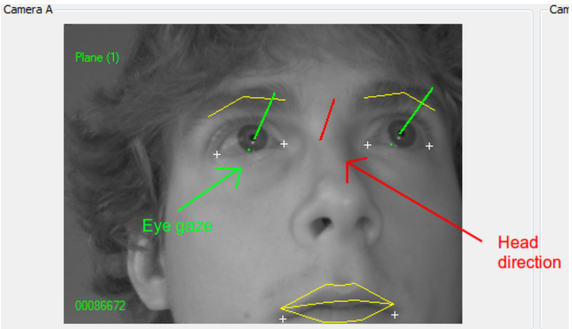

Fig. 3. Second experiment: Facelab interface showing head direction as a red vector between the eyes and eye gaze as two green vector located on the eyes.

speed can be the same as the framerate (30 fps). Within the TV viewer profile personalization application, the use of pre-recorded video is possible as the head pose data is only sent when the context changes (viewers enter/leave, etc.) as explained in section 6.

To detect if a user watches TV or not, we reconstruct a virtual simplified model of the real scene (Figure 4). Therefore, knowing the 6DOF position of the face of the person detected, the camera position and the TV position it is possible to estimate the point of intersection between the TV and the orientation of the head. In this way, we can synchronize annotated media with the head tracker

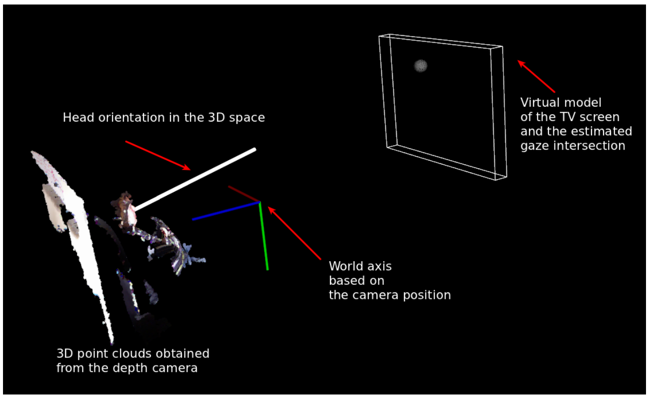

Fig. 4. 3D rendering of our system. On left: 3D point cloud from depth camera and head direction vector. On right: 3D model of the TV and intersection point between TV and viewer head position.

and estimate (±10 cm, on our 46" TV) where the user is looking.

Depending on the camera position, user head direction can be detected towards the main screen or the second screen (tablet) as in Figure 5, top images. To be able to achieve this measure, the 3D camera must see the viewer face 1) with no occlusions due to the second screen, 2) with a pitch angle which is small enough for the algorithm (±60 degrees pitch).

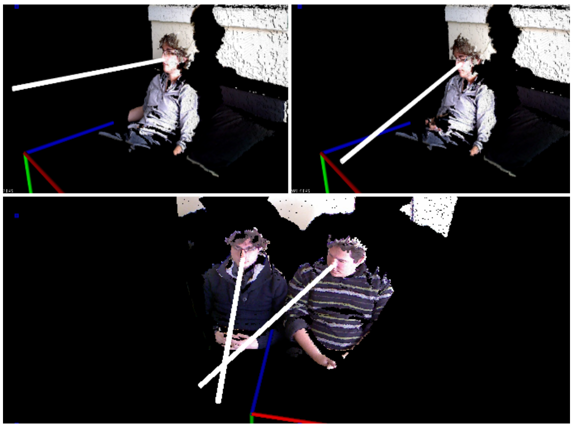

Fig. 5. Top-images: switch between main screen (left) and second screen (right). Bottom-image: Multiple head detection and orientation estimation.

Moreover, the algorithm can detect several users (as many as possibly detected in the camera field of view) and compute all users head directions as in Figure 5, bottom image. This feature allows us to check potential joint attention on the main TV screen.

5 Results of the second scenario: attention from head

The easier way to measure people overt [25] attention is to measure eye gaze or direction. Given the technical limitations of camera distance, it is not possible to access the viewer's eyes orientation. We than hypothesise that, at the TV setup distance (more than two meters from the main screen), the gaze of a person is considered to be close to the direction of his head. As stated in [18], "[...]*Head pose estimation is intrinsically linked with visual 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].
In the second experiment we firstly qualitatively compared the Facelab head direction detection with the proposed algorithm. The results are similar, and the proposed approach seems even to be more reactive to head movements, while the one of Facelab needs large head movements. However, a quantitative comparison is not simple due to the framerate difference between the two systems.
In a second step we compared the head direction with the eye gaze using again the Facelab system (Figure 6). The first results we obtained are consistent with the literature and show that there is a correlation between eye gaze and head direction. This correlation is higher when the gaze goes far from the image centre and for more dynamical content (fast moving videos). The head direction does not exactly follow the eye gaze which is much faster to attend events occurring on the TV screen, but the head direction accompany the gaze in a smoother way. The head and eye movements work together to both minimize their motion (effort) and maximizing the acquisition of interesting information in the scene. In this optimisation process, the head mechanics naturally act like a smoother while eye reactions can be much faster.

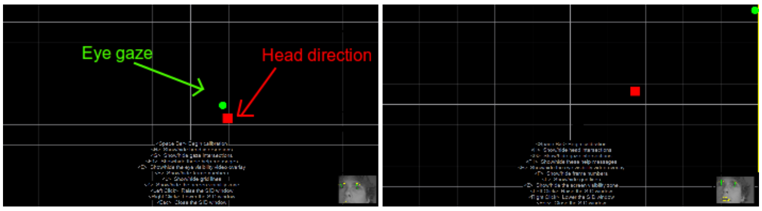

Fig. 6. Green circle: eye gaze, Red square: head direction. Left image: both are very close, Right image: the eye quickly shifted to the top-right corner and the head position followed in the same direction.

6 From attention to content personalization

Based on our preliminary studies, we can say that head direction might be a rough approximation of eye gaze, thus of overt attention. Attention is a phenomenon based on two competing precesses: the top-down and bottom-up attention [22]. Bottom-up attention is a generic approach also known as stimulus-driven or exogenous attention. Furthermore, it relies on the information innovation that the features extracted from the image can bring in a given spatial context. The top-down component of attention, which is also known as task-driven or endogenous attention, integrates specific knowledge that the viewer could have in specific situations (tasks, models of the kind of scene, recognized objects, etc.).
While bottom-up attention will be engaged each time that surprising images/motion or sounds arise, top-down attention shows that the viewer is specifically interested to the content displayed. Thus, for content personalization, the most important is to extract the viewer interest in terms of top-down attention.
To detect the top-down attention of the viewer, several features can be extracted.

- 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. This is a sign of sustained attention which shows that cognitive processes are engaged and it is not only a bottom-up attention due to surprising events [13]. A first feature is thus the time spent looking at the screen after a head position change.
- [15][4][5] 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. The speed of the head direction change and the total angle of the head motion are additional features of the kind of attention which the viewer uses.
- In case of several users, joint attention (see Figure 5, bottom image) which is stable during enough time shows a discussion or a common subject of interest. Joint attention is an additional cue for top-down attention.

Based on the detection of focus on the main screen and the nature of the attention attracted by the media (sustained, bottom-up), it is possible to provide, for each media segment a weight of interest that the user implicitly expressed. This weight, mixed along with other contextual cues (time and date, number of viewers, the presence of children or not, etc.) and the explicit actions of the viewer (skip, play, stop, explore links, etc.) provides a good idea about the subsets of the media which are interesting for the viewer. This information let an ontology-based system propose to the viewer media which are close to the viewer interests depending on the context (viewing alone during the WE will most of the time be different from viewing in family during week days).

The data collected through the system is sent to a content personalization framework. At each change in context (new viewers entering, viewers leaving, kids, coming or leaving, etc.) the logs of the head focus for the viewers is sent (1: focus on the main screen, 2: focus on the second screen, 3: focus out of screens,

4: no head detected (the viewer is talking to another one or looking back ...)). In addition to the head focus, for the first two modes, the kind of attention (bottom-up or top-down) is also sent. These logs will be used to modify the viewer profile in the given context. Some feature combinations will provide cues about a positive interest of the viewer (look to the main screen - mode 1 and top-down attention, look to second screen - 2 and top-down attention), others about a negative interest (look to the walls - mode 3) and others will provide a neutral result (not enough to know about the viewer interest, keep previous score like mode 4 or modes 1 and 2 with bottom-up attention).
To summarize, the system described in this paper, at the end of each user session (context change: when a user leaves the interaction zone, when a second user comes in), the logs containing the tracking data will be sent as REST [11] query to the remote personalization module called GAIN (General Analytics INterceptor) [14] which will use rule-based learning algorithms to change viewer profile accordingly.

7 Conclusions

In this paper, we presented a system architecture and two preliminary experiments on an implicit behaviour analysis system based on a 3D head tracker. This tool is optimized compared to previous publications and it is designed to feed a personalization framework capable of processing behavioural data to dynamically enhance a user profile. The preliminary results show that it is possible to extract implicit information and that head direction can provide cues about viewer interest which can be used in future TV personalization. In the future, 1) more extensive tests will be conducted to confirm the preliminary findings of our two experiments and 2) additional information will be provided concerning the kind of attention (bottom-up or top-down) which is crucial information to asses real viewer interest.

8 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)

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. J. Henderson. Regarding scenes. *Current Directions in Psychological Science*, 16:219–222, 2007.
14. T. K. Jaroslav KUCHAR. Gain: Analysis of implicit feedback on semantically annotated content. WIKT 2012, pages 75–78, 2012.
15. 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.
16. J. Leroy, F. Rocca, M. Mancas, and B. Gosselin. Second screen interaction: an approach to infer tv watcher's interest using 3d head pose estimation. In *Proceedings of the 22nd international conference on World Wide Web companion*, WWW '13 Companion, pages 465–468, Republic and Canton of Geneva, Switzerland, 2013. International World Wide Web Conferences Steering Committee.
17. Microsoft. Kinect sensor. http://www.xbox.com/kinect.
18. 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.
19. OpenNI. Open natural interfaces. http://www.openni.org/.
20. PCL. Point cloud library. http://pointclouds.org/.
21. PrimeSense. Capri sensor. http://www.primesense.com/news/primesense-unveils-capri.
22. N. Riche, M. Mancas, M. Duvinage, B. Gosselin, and T. Dutoit. Rare2012: A multi-scale rarity-based saliency detection with its comparative statistical analysis. *Signal Processing: Image Communication*, 2013.
23. Seeingmachine. Facelab5. http://www.seeingmachines.com/product/facelab/.
24. A. Vinciarelli, M. Pantic, and H. Bourlard. Social signal processing: Survey of an emerging domain. *Image and Vision Computing*, 27(12):1743 – 1759, 2009.
25. R. D. Wright and L. M. Ward. *Orienting of attention.* Oxford University Press, 2008.