Real-time marker-less implicit behavior tracking for user profiling in a TV context

F. Rocca, P.-H. De Deken, F. Grisard, M. Mancas & B.Gosselin
Numediart Institute - University of Mons
Mons, Belgium
{francois.rocca, pierre-henri.dedeken, fabien.grisard,
matei.mancas, bernard.gosselin}@umons.ac.be

Abstract

In this paper, we present a marker-less motion capture system for user analysis and profiling. In this system we perform an automatic face tracking and head direction extraction. The aim is to identify moments of attentive focus in a non-invasive way to dynamically improve the user profile by detecting which media have drawn the user attention. Our method is based on the face detection and head-pose estimation in 3D using a consumer depth camera. This study is realized in the scenario of TV watching with second screen interaction (tablet, smartphone), a behaviour that has become common for spectators. Finally, we show how the analysed data could be used to establish and update the user profile.

Keywords: head pose estimation, viewer interest, face tracking, attention, user profiling

1. Introduction

In this work we will focus on the analysis of a user sitting in front of his television. It will give us information on the spectator behavior. What draw the user interest? The analysis of the interest that the user brings to his environment is significant for the user profiling. This information can be known or estimated by different methods. The best way is to get a rapid estimation of the interest based on gaze direction and also the duration of the gaze fixation in this direction. Once the interest is known on the different media segments, it is possible to update the user profiling.
In section 2, we present techniques for gaze estimation based on head pose and we will describe the marker-less method used in this study. Section 3 shows the descriptor used to defined the user interest. In section 4, we present the experimental setup which was used to estimate the interest and send these data to the profiling platform. Finally we conclude in section 5.

2. State of the art

The orientation of the head is less difficult to estimate than the direction of the eyes. The head direction and the eye-gaze direction are strongly correlated. Physiological studies have shown that the prediction of gaze is a combination of the direction of the eyes and the direction of the head [1]. In this work, the distance from the sensor can be up to a few meters, and at these distances, the eye tracking becomes very difficult to achieve. An initial study showed the link between eye-gaze direction and the head direction. In this study, the correlation was assessed qualitatively when user focuses his gaze on a map (Figure 1). The results were obtained with the eye tracking system FaceLAB [2] and show that the average error is 3 to 4 cm to a plane located 1 meter away, which means that the angular difference is very small. The direction of the head is intrinsically linked to the direction of the eyes. This is especially the case when still in a rotating area of the head comfortable for the

user. Therefore, the direction of the face gives a good indication on the look when it is not possible to clearly get the direction of the eyes.

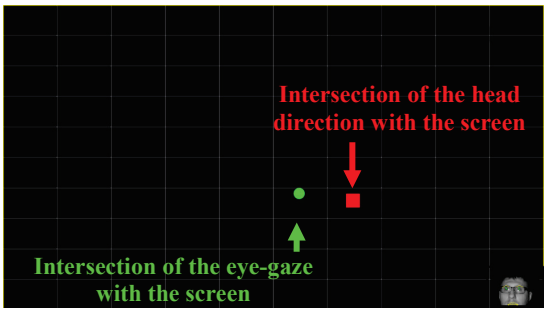

Figure 1. The head direction and the eye-gaze are highly correlated.

The gaze estimation can be achieved by calculating the orientation of the head, and these rotations have physiological limits and specific names (Figure 2). For an average user, the range of motion extending from the sagittal flexing of the head to the extension (head movement from the front to the rear) is about -60° to 70°. This movement is more commonly called "Pitch". Regarding the front lateral flexion (movement from right to left when looking ahead), it occurs around 40° in each direction and is called "Roll". The last movement, a horizontal axial rotation (head motion by looking horizontally from right to left), is around 78° in each direction [3] and is named "Yaw". All the motions of head rotation can be obtained by combining these angles.

In the animation industry, head pose estimation and head movements are almost exclusively captured with physical sensors and optical analysis. Physical sensors such as gyroscopes, accelerometers and magnetometers are placed on the head to compute the head rotation.

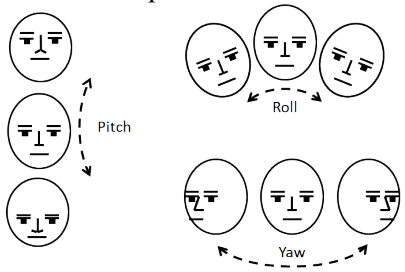

Figure 2. The 3 different degrees of freedom: pitch, roll and yaw [7].

Another way for head pose estimation is marker-based optical motion capture. These systems are able to capture the subtlety of the motion because the markers are placed on the head of the actor and they are tracked through multiple cameras. The markers are often colored dots or infrared reflective markers and the cameras depend on the markers type. Accurate tracking requires multiple cameras and specific software to compute head pose estimation but these systems are very expensive and complex, and they need for precise positioning of markers and calibration (Optitrack [4], Qualisys [5]).

Marker-less tracking is another approach for face motion capture and a wide range of methods exists. Some marker-less equipment uses infrared cameras to compute tracking of characteristic points. For example, FaceLAB gives the head orientation and the position of lips, eyes and eyebrows [2]. But there are also algorithms using only a consumer webcam. We can cite Facetracker using PNP [6] and FaceAPI [2]. Marker-less systems use classical cameras or infrared cameras to compute tracking of characteristic points. Based on consumer infrared camera, we can cite the Microsoft KinectV1 SDK [7]. The KinectV1 SDK is free, easy to use and contains multiple tools for user tracking such as face tracking and head pose estimation. These tools combine 2D and 3D information obtained with the KinectV1 sensor. Based on 3D consumer sensor there are also methods using random regression forest for head pose estimation from only depth images [8].

In this work we choose to use the KinectV2 with the new version of the SDK [9]. The KinectV2 is composed by a color camera (1080p) and a depth sensor (512x424 pixels). The technology behind the new sensor is infrared TOF for time of flight. This sensor measures the time it takes for pulses of laser light to travel from the laser projector to a target surface, and then back to an image sensor. Based on this measure, the sensor gives a depth map. To achieve head pose, at least the upper part of the user's KinectV2 skeleton has to be tracked in order to identify the position of the head. The position of the head is located using the head pivot from the 3D skeleton only on the depth map. The head pose estimation is based on the face tracking and it is achieved on the color images. Consequently, the face tracking is dependent on the light conditions, even if KinectV2 is stable into darker light conditions.

3. Fixing duration and direction for interest measurements

Based on the gaze estimation, or in this case on the head direction, it is possible to measure the interest of a person to a specific element of his environment by calculating the intersection between the direction of the face and a 2D plane (or a 3D volume). In this case, the TV screen will be represented by a 2D plane and another 2D plane will be used for the second screen. The previous head-pose estimation will give an indication on what the user is looking at. In a visual attention to television context, a study showed that there are four types of behavior depending on the fixing duration [10]. This classification is given in Table 1.

Table 1. Attentive behavior over time.

Duration	≤1.5 sec.	1.5 sec. to 5.5 sec.	5.5 sec. to 15 sec.	>15 sec.
Behavior	Monitoring	Orienting	Engaged	Stares

These four measures of attention correspond to be firstly attracted by something with "monitoring behavior", and then intrigued, "orienting behavior", and more time passes more the user becomes interested, "engaged behavior", and beyond 15 seconds the user is captivated with a "staring behavior". These measures have been established for a TV watching and used to correctly describe the interaction with one or more screens.

4. Experimental setup

4.1 Placement

The purpose of the experiment is to get a maximum of information on user implicit behavior in front of the TV. Several users watch a TV (main screen) and need in the same time to focus on some of the content while playing to a tablet game (second screen). The sofa is installed 2 meters away from the TV which is equipped with a KinectV2 (Figure 3).

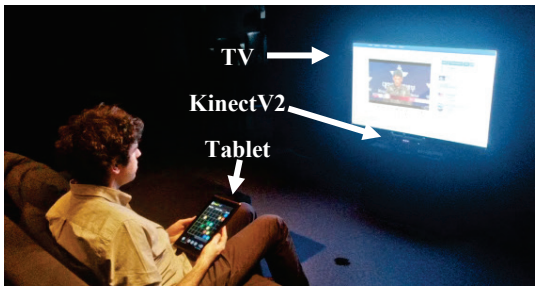

Figure 3. User watches the TV with a tablet in his hand (second screen). The head is about 2.5 meters from the TV.

4.2 User interaction and media

The system allows us to know when the user watches the TV (main screen) or the tablet (second screen) using the interest durations given in point 3. When the user does not watch TV, the image is hidden but the media runs because the user hears the sound. The user can use a keyboard to control the player and navigate into the video enrichment displayed next to the player. The video used for the tests is a mashup of CNN Student News. It has been enriched with links to related web pages that are displayed next to the video.

4.3 Behavior analyses

When the user comes into the field of view of the KinectV2, placed under the TV, his skeleton is tracked and the head orientation is estimated. The Tracker Computer performs the process and determines what the user is watching with an accuracy of a few centimeters: Screen 1 (video player or list of enrichement), Screen 2 or elsewhere. These informations are completed by attentive behavior over time and are sent to the Player Server (Figures 4 & 5). The televison displays the web page from the player server containing the media player accompanied by enrichments related to playing video segment [11]. The working flow structure is given on Figure 4.

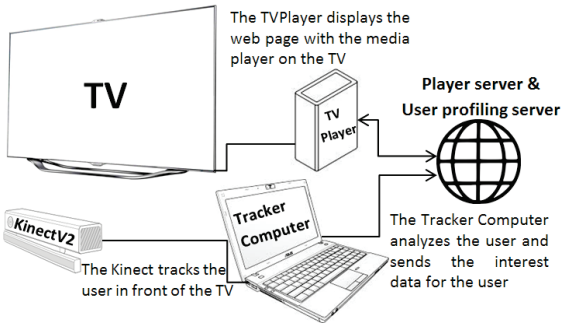

Figure 4. Overall working flow.

4.4 User Profiling

The data coming from tracking is related to each video segment through the server player (Figure 5). The User Profiling module receives measures of interest (or disinterest) for each video segment. A score of interest could be calculated for each keyword from the profiling list. This score list allows to establish the user profile and to know by what the user is interested.

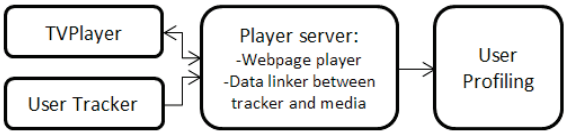

Figure 5. User interest is sent from the User Tracker to the User Profiling trough the Player Server.

At the end of each session we get events timeline: the interest value for each screen, player control, etc. This will update the list of the score on the user profile (Figure 6).

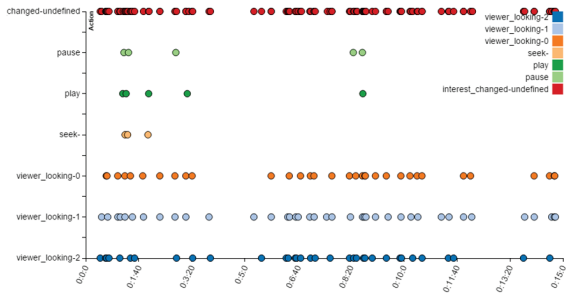

Figure 6. Event timeline for one session of 15 minutes

5. Conclusion

In this paper, we have described a marker-less motion capture system for implicit behaviour analysis in a TV context using a consumer depth camera. This system allows us to establish and update a user profile using user interest based on head pose estimation and using the duration of fixation separated into 4 levels of attention. The aim is to identify moments of attentive focus in a non-invasive way to dynamically improve the user profile by detecting which parts of the media have drawn the user attention.

Acknowledgements

This work is supported by the LinkedTV Project funded by the European Commission through the 7th Framework Program (FP7- 287911) and by the funds “Region Wallonne” WBHealth for the Project RobiGame.

References

[1] S. Langton, H. Honeyman, and E. Tessler, “The influence of head contour and nose angle on the perception of eye-gaze direction,” Perception and Psychophysics, vol. 66, no. 5, pp. 752–771, 2004

[2] Seeing Machines. FaceLAB5 & FaceAPI; Face and eye tracking application, 2011.

[3] V. Ferrario, et al. “Active range of motion of the head and cervical spine: a three-dimensional investigation in healthy young adults,” J. Orthopaedic Research, vol. 20, no. 1, pages. 122–129, 2002.

[4] OptiTrack, Optical motion tracking solutions. www.optitrack.com/ accessed on 23/03/2015

[5] Qualisys. Products and services based on optical mocap. http://www.qualisys.com accessed on 19/03/2015

[6] F. Rocca, et al. Head Pose Estimation by Perspective-n-Point Solution Based on 2D Markerless Face Tracking. Intelligent Technologies for Interactive Entertainment: 6th International ICST Conference, 2014

[7] Microsoft Kinect Software Development Kit, www.microsoft.com/en-us/kinectforwindowsdev/Start.aspx, on 20/02/2015

[8] G. Fanelli, J. Gall, and L. Van Gool. Real time head pose estimation with random regression forests. CVPR, 617-624, 2011.

[9] KinectV2 SDK, https://msdn.microsoft.com/en-us/library/dn799271.aspx, accessed on 18/03/2015

[10] R. Hawkins, S. Pingree, et al. What Produces Television Attention and Attention Style?. Human Communication Research, 31(1), pages 162-187, 2005

[11] J. Kuchař and T. Kliegr. GAIN: web service for user tracking and preference learning - a smart TV use case. RecSys '13. In Proceedings of the 7th ACM conference on Recommender systems. ACM, New York, NY, USA, 467-468. 2013