DOI: 10.2390/biecoll-saga2013_8

In Pfeiffer and Essig (2013): Proceedings of the First International Workshop on Solutions for Automatic Gaze-Data Analysis 2013 (SAGA 2013). DOI: 10.2390/biecoll-saga2013_0

3D Gaze Recovery in Large Environments Using Visual SLAM

Lucas Paletta, Katrin Santner, Albert Hofmann, Georg Thallinger

DIGITAL – Institute for Information and Communication Technology
JOANNEUM RESEARCH Forschungsgesellschaft mbH, 8010 Graz, Austria

1 Introduction

This work describes a multi-component vision system that enables pervasive mapping of human attention. The key contribution is that our methodology enables full 3D recovery of the gaze pointer, human view frustum and associated human centered measurements directly into an automatically computed 3D model. We apply RGB-D SLAM and descriptor matching methodologies for the 3D modeling, localization and fully automated annotation of ROIs (regions of interest) within the acquired 3D model. This methodology enables fully automated processing of human attention, without artificial landmarks, in indoor natural environments.

2. Implementation

This work presents a computer vision system methodology that, *firstly*, enables to precisely estimate the 3D position and orientation of human view frustum and gaze [1] and from this enables to precisely analyze human attention in the context of the semantics of the local environment (objects [10], signs, scenes, etc.). *Secondly*, the work describes how ROIs (regions of interest) are automatically mapped from a reference video into the model and from this prevents from state-of-the-art laborious manual labeling of tens / hundreds of hours of eye tracking video data. This provides a scaling up of nowadays still small sketched attention studies. With the presented methodology, extended natural environments, such as shop floor departments, analysis of navigation guidance, and human-robot interaction, can be studied first time in large scale, statistically significant usability studies.

For a spatio-temporal analysis of human attention in the 3D environment, we firstly build a spatial reference in terms of a three-dimensional model of the environment using RGB-D SLAM methodology [2]. Secondly, the user's view is gathered with eye tracking glasses (ETG) within the environment and localized from extracted local descriptors [3]. Then ROIs are marked on imagery and automatically detected in video and further mapped into the 3D model. Finally, the distribution of saliency onto the 3D environment is computed for further human attention analysis, such as, evaluation of the attention mapping with respect to object and scene awareness. Saliency information can be aggregated and further evaluated in the frame of user behaviors of interest. The performance evaluation of the present-

ed methodology firstly refers to results from a dedicated test environment [4] demonstrating very low projection errors, enabling to capture attention on daily objects and activities (package logos, cups, books, pencils).

3 Gaze recovery in 3D without artificial markers

Human Attention Analysis in 3D. 3D information recovery of human gaze has in principle been targeted by Munn et al. [2] who introduced monocular eye-tracking and triangulation of 2D gaze positions of subsequent key video frames, obtaining observer position and gaze pointer in 3D with angular errors of ≈3.8°. Pirri et al. [3] achieved accuracy indoors about ≈3.6 cm at 2 m distance to the target compared to our ≈0.9 cm [1]. Previous attempts focused on single 3D point recovery. Our approach maps fixation within a 3D environment model with the possibility of real-time tracking of attention with mass marketed eye-tracking hardware.

Visual Map Building and Camera Pose Estimation. For realistic environment modeling we make use of an RGB-D sensor providing per pixel color and depth information at high frame rates. Our environment consists of a sparse point-cloud, where each landmark [4] is attached for data association during pose tracking. Estimated camera poses are stored in a 6DOF manner. Incremental camera pose tracking assuming an already existing map is done by keypoint matching followed by a least-square optimization routine minimizing the reprojection.

Densely Textured Surface Generation. For realistic environment visualization, user interaction and subsequent human attention analysis, a dense, textured model of the environment is constructed. Depth images are integrated into a 3D occupancy grid [5] using the previously corrected camera pose estimates.

3D Gaze Recovery from Monocular Localization. To estimate the proband's pose, SIFT keypoints are extracted from ETG video frames and a *full 6DOF pose* is estimated using the perspective n-Point algorithm [6].

Automated 3D Annotation of Regions of Interest. Annotation of ROIs in 2D or even 3D information usually causes a process of massive manual interaction. In order to map objects of interests, such as, logos, package covers, etc. into the 3D model, we first use logo detection in the high resolution scanning video to search for occurrences of predefined reference appearances, using vocabulary trees [4].

Semantic Mapping of Attention. The automatic detection of ROIs in 3D enables statistical evaluations, such as on ROIs called AOI hit, which states for a raw sample or a fixation that its coordinate value is inside the ROI [7]. From this, the dwell time distribution for ROIs can be plotted over all participants, and some of the captured fixations are related to human object recognition which is known to trigger from 100 ms of observation / fixation [8].

4 Experimental results

Eye Tracking Device. The mass marketed SMI™ eye-tracking measure the gaze pointer for both eyes with 30 Hz. The gaze pointer accuracy of 0.5°–1.0° and a tracking range of 80°/60° horizontal/vertical assure a precise localization of the human's gaze in the HD 1280x960 scene video with 24fps. We recorded data on a shop floor covering an area of about 8x20m² (Figure 1). We captured 2366 RGB-D images, reconstructed the model from 41700 natural visual landmarks.

5 Conclusions and future work

We present a complete system for (i) wearable data capturing, (ii) automated 3D modeling, (iii) automated recovery of human pose and gaze, and (iv) automated ROI based semantic interpretation of human attention. The presented system is a significant step towards a mobile mapping framework [9] for quantitative analysis of human attention measures [7,10] in natural environments (Figure 2). Future work will focus on improved tracking of the human pose across image blur and uncharted areas as well as study human factors in the frame of stress and emotion in the context of the 3D space.

Figure 1. Hardware (a) for the 3D model building process (Kinect and HD camera), (b) study with packages, (c) 3D environment model, a large shop floor.

ACKNOWLEDGMENTS: The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement No. 288587 (MASELTOV) as well as from the Austrian FFG via contracts No. 832045 (FACTS) and No. 836270 (EVES), and by the Provinc. Gov. of Styria (NeoAttrakt). We kindly thank INTERSPAR Graz and CITYPARK GmbH for the permission to capture the data.

References

1. Paletta, L., Santner, K., Fritz, G., Mayer, H., & Schrammel, J.: 3D Attention: Measurement of Visual Saliency Using Eye Tracking Glasses, *Proc. CHI 2013*.
2. Munn, S. M., & Pelz J. B.: 3D point-of-regard, position and head orientation from a portable monocular video-based eye tracker. *Proc. ETRA 2008*, pp. 181-188.
3. Pirri, F., Pizzoli, M., & Rudi, A.: A general method for the point of regard estimation in 3D space. *Proc. CVPR 2011*, pp. 921-928.
4. Nistér, D. & Stewénius, H.: Scalable Recogn. with a Vocabulary Tree, Proc. CVPR 2006.
5. Marks, T. K., Howard, A., Bajracharya, M., Cottrell, G. W. & Matthies, L.: Gamma-SLAM: Using stereo vision and variance grid maps for SLAM, *Proc. ICRA 2008*.
6. Lepetit V., Moreno-Noguer F. and Fua P.: EPnP: An Accurate O(n) Solution to the PnP Problem, *International Journal of Computer Vision*, pp. 155-166, 2009.
7. Holmqvist, K., Nyström, M., Andersson, R., Dewhurst, R., Jarodzka, H., and van de Weijler, J.: *Eye Tracking*, Oxford University Press, 2011, pp. 187.
8. Grill-Spector, K. and Sayres, R. Object Recognition: Insights From Advances in fMRI Methods, Current Direct. *Psychol. Science*, Vol. 17, No. 2. 2008, pp. 73-79.
9. Paletta, L., Santner, K., Fritz, G., Hofmann, A., Lodron, G., Thallinger, G., and Mayer, H. FACTS - A Computer Vision System for 3D Recovery and Semantic Mapping of Human Factors, *Proc. ICVS 2013*, Springer-Verlag, LNCS 7963, pp. 62-72.
10. Fritz, G., Seifert, C., Paletta, L., and Bischof, H. (2005). Attentive object detection using an inform. theoret. saliency measure, *Proc. WAPCV 2004*, Springer-Verlag, LNCS 3368.

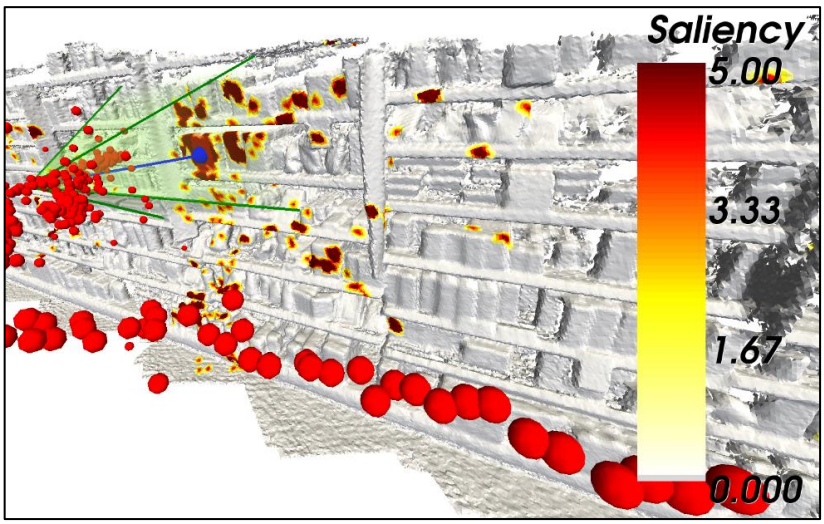

Figure 2. Mapping of user saliency onto the acquired 3D model and automated recovery of the trajectory of ETG camera positions (spheres), as well as recovery of frustum (green lines) and gaze pointer (blue).