3D Attention: Measurement of Visual Saliency Using Eye Tracking Glasses

Lucas Paletta
DIGITAL – Inst. Information and Communication Technologies
JOANNEUM RESEARCH Forschungsgesmbh, Austria
lucas.paletta@joanneum.research

Katrin Santner
DIGITAL – Inst. Information and Communication Technologies
JOANNEUM RESEARCH Forschungsgesmbh, Austria
katrin.santner@joanneum.research

Gerald Fritz
DIGITAL – Inst. Information and Communication Technologies
JOANNEUM RESEARCH Forschungsgesmbh, Austria
gerald.fritz@joanneum.research

Heinz Mayer
DIGITAL – Inst. Information and Communication Technologies
JOANNEUM RESEARCH Forschungsgesmbh, Austria
heinz.mayer@joanneum.research

Johann Schrammel
CURE – Center for Usability Research & Engineering
Modecenterstr. 17, Obj. 2, Vienna, Austria
schrammel@cure.at

CHI 2013 Extended Abstracts, April 27–May 2, 2013, Paris, France.
ACM 978-1-4503-1952-2/13/04.

Abstract
Understanding and estimating human attention in different interactive scenarios is an important part of human computer interaction. With the advent of wearable eye-tracking glasses and Google glasses, monitoring of human visual attention will soon become ubiquitous. The presented work describes the precise estimation of human gaze fixations with respect to its environment, without the need of artificial landmarks in the field of view, and being capable of providing attention mapping onto 3D information. It enables full 3D recovery of the human view frustum and the gaze pointer in a previously acquired 3D model of the environment in real time. The key contribution is that our methodology enables mapping of fixations directly into an automatically computed 3d model. This innovative methodology will open new opportunities for human attention studies during interaction with its environment, bringing new potential into automated processing for human factors technologies.

Author Keywords
Human attention; gaze recovery; map building and localization; saliency maps.

ACM Classification Keywords
H.1.2 User/Machine Systems - Human information processing.

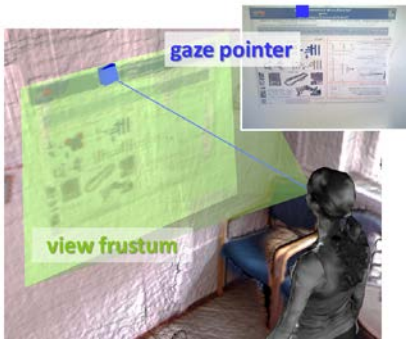


Figure 1. 3D gaze recovery: eye tracking glasses localize human fixations in video (above). The user's view frustum and gaze pointer is reconstructed in the 3D environment model.

Introduction

Understanding and estimating human attention is an important part of human computer interaction, and different computational attention modeling approaches have been proposed [12,13]. Today, human attention gets investigated directly in the space where the task of interest is performed [e.g. 11]. An important requirement to perform scientific measurements and interpretations using mobile eye tracking data is to measure the exact embedding of attention in its environmental context. Application objectives are marketing, usability engineering, to develop computational models of human attention for simulation studies or perception design in humanoid interactive robots. Furthermore, with the advent of Google glasses and increasingly affordable wearable eye-tracking, monitoring of human attention will soon become ubiquitous. The presented paper anticipates that wearable human factor services will provide enabling technologies for more precise estimation of human attention in human computer interaction. This work presents a novel methodology that enables to precisely estimate the position and orientation of human view frustum and gaze and from this enables to precisely analyze human attention in the context of the semantics of the local environment (objects, signs, scenes, etc.). Figure 1 visualizes how accurately human gaze is mapped into the 3D model for further analysis.

The methodology for the recovery of human attention in 3D environments is based on the workflow as sketched in Figure 2: 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 (i.e., GPSlam [1]). Secondly, the user's view is gathered with eye tracking glasses within the environment and localized from extracted local image descriptors [4]. 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, for example, being further evaluated in the frame of user behaviors of interest.

Performance evaluation of the methodology concentrates on the error of modeling and localization which remains limited within the size of the eye-tracking technology. The angular projection error is ≈0.6º within the chosen 3D model and therefore smaller than the calibration error of the eye-tracking glasses (≈1º). The Euclidean projection error is only ≈1.1 cm (avg.) and therefore enables to reconstruct attention on daily objects and activities.

Related Work

Reconstruction of the Environment.

Solving the SLAM problem involves constructing a map of a previously unknown environment while simultaneously localizing within this map. Early approaches relying on a single camera as the main sensor were based on filtering approaches (Civera et al. [2]) or Structure from Motion [3]. With the launch of range image devices (e.g. Microsoft's Kinect) that are based on structured light and directly providing per pixel depth information, large scale dense reconstruction of indoor environments has been proposed. Our map building method follows the approach of Pirker et al. [1], reconstructing large scale environments by fusing bundle adjustment techniques with occupancy grid mapping.

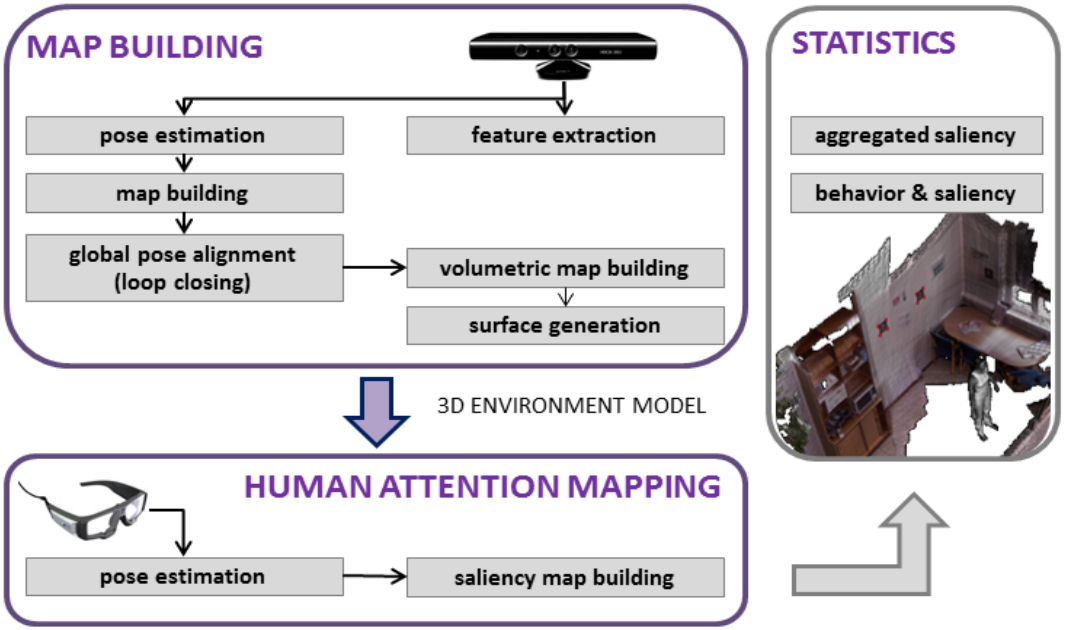


Figure 2. Schematic sketch of the workflow for the recovery of 3D human gaze: The metric 3D environment model is built with RGB-D information. Pose estimation is achieved on the user track with image descriptor matching. From trajectories one can deduce statistical facts.

Reconstruction of Human Pose.
Localization of the eye-tracking glasses' camera is mandatory to represent an accurate 6 DOF (degrees of freedom) pose of the test subject within the environment model. A 6 DOF pose can be computed by a perspective n-Point algorithm [7] or least-squares optimization techniques on the re-projection error. In case of large scale maps, image retrieval techniques are less costly and should to be used instead [5].

Analysis of Human Attention.
Munn et al. [8] introduced monocular eye-tracking and triangulation of 2D gaze positions of subsequent key frames of the eye-tracking video. They reconstructed only single 3D points without the reference to a complete 3D model and achieved an angular error of ≈3.8° compared to ≈0.6° to our method. Voßkühler et al. [9] proposed to analyze 3D gaze movements of freely moving observers. However, they need a special head tracking unit for intersecting the gaze ray with a digitized model of the surrounding. Pirri et al. [10] proposed gaze estimation in 3D space with a special, not mass marketed stereo rig that is required in addition to a commercial eye-tracking device. The achieved accuracy indoor is ≈3.6 cm at 2 m distance to the target compared to ≈0.9 cm at the same distance of our proposed workflow. Furthermore, attention is not mapped and cannot be tracked within a 3D model of the environment. In contrast to previous proposals, we present a straight forward solution of mapping fixation distributions onto saliency maps within a model of the environment, using mass marketed eye-tracking hardware.

3D Map Building.

Color and depth images are collected within the environment with an RGB-D device. Feature-based, visual SLAM is performed resulting in sparse three-dimensional point clouds together with camera pose estimates. Then a dense, textured surface is built by integrating depth measurements into a volumetric model according to [1].

Visual SLAM and Camera Pose Estimation
Visual SLAM aims at estimating the camera pose (pose tracking) whilst simultaneously constructing the

previously unknown environment (mapping). Assuming an already existing map, per frame pose tracking is done by key point detection and matching against those key points already in the map. Since images can be captured at high frame rates, a guided search technique is used for correspondence estimation to speed up the matching process and to guarantee robust matches. The new camera pose is then estimated by minimizing the re-projection error between previously estimated 2D-3D correspondences using robust least-squares optimization. As a result, the problem of scale-drift - which is inevitably in monocular SLAM – gets drastically reduced. Loop closure detection is handled by a vocabulary tree search [5], followed by a geometric consistency check. Loop-closure correction follows the pose-graph optimization procedure presented in [3].

Dense Surface Generation
For human attention analysis and realistic environment modeling, a dense environment representation is necessary. Therefore, we decided to form an occupancy grid, which discretizes the environment into equally sized voxels. Every depth measurement provided by the sensor is integrated into the volume using the previously estimated camera poses. Hereby, we follow a pyramidal mapping approach where all voxels inside the cameras view frustum are updated in parallel using a 20 fps GPU implementation. Surface computation is done by a marching cubes algorithm and simple per vertex texture mapping.

Human Attention Analysis

Monocular Localization and 3D Gaze Estimation
To estimate the test users' pose within the previously reconstructed area, we use SIFT key points [4] extracted from the images. First, every image descriptor is matched against all descriptors represented in the sparse point cloud map resulting in 2D-3D correspondences. These are filtered through geometric verification (e.g. RANSAC based fundamental matrix estimation). Finally, a 6 DOF pose is estimated using the perspective n-Point algorithm [7].

Given human pose and image gaze position, we are interested in its fixation point within the 3D map. Thus we compute the intersection of the viewing ray through the gaze position with the triangle mesh of the model, using an object oriented bounding box tree [6].

3D Saliency Computation
Saliency maps are widely used in eye-tracking studies to visualize the distribution of fixation hits on the environment as well as the amount of time one particular area was fixated. Fixations are integrated over time and typically project into a 2D reference image, i.e., on websites or shopping shelves. Our methodology enables mapping of fixations directly into an automatically computed 3D model. Figure 4a depicts a saliency map directly computed onto the 3D model, allowing for automatic generation of visualizations without manual interaction, in contrast to today's state-of-the-art methods.

Experiments

To evaluate our systems performance and applicability we captured visual information of a room (6x3 m^2) using Microsoft's Kinect. We also attached four artificial markers on the walls the positions of which have been measured by a tachymeter to receive three-dimensional ground-truth data. Afterwards, two test

Figure 3. Eye-tracking glasses (ETG) used in the experiments. (a) ETG video frame. (b) ETG, (c) test user.

Marker ID	Mean Euclidean Error [cm]
# 1	16.46
# 2	34.13
# 3	2.71
# 4	3.74

Table 1. Euclidean distance error of marker re-projection.

persons wearing the calibrated eye-tracking glasses moved randomly through the room.

The mass marketed SMI™ eye-tracking glasses (Figure 3) were used by the subjects for the gathering of the gaze data. This non-invasive video based binocular eye tracker with automatic parallax compensation measures 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 assures a precise localization of the human's gaze in the HD 1280x960 scene video with 24fps. For all experiments an accurate (less than 0.5° validation error) three point calibration was performed and the gaze positions within the HD scene video frames were used for further processing.

Performance Evaluation of Reconstruction

For accuracy evaluation the resulting textured environment model of the test scenery is compared to the ground truth data acquired by the tachymeter. To align the reconstructed marker corners with those measured by the tachymeter a rigid body transformation is estimated using a RANSAC routine. Euclidean errors of the markers are given in Table 1.

Evaluation of Gaze Recovery

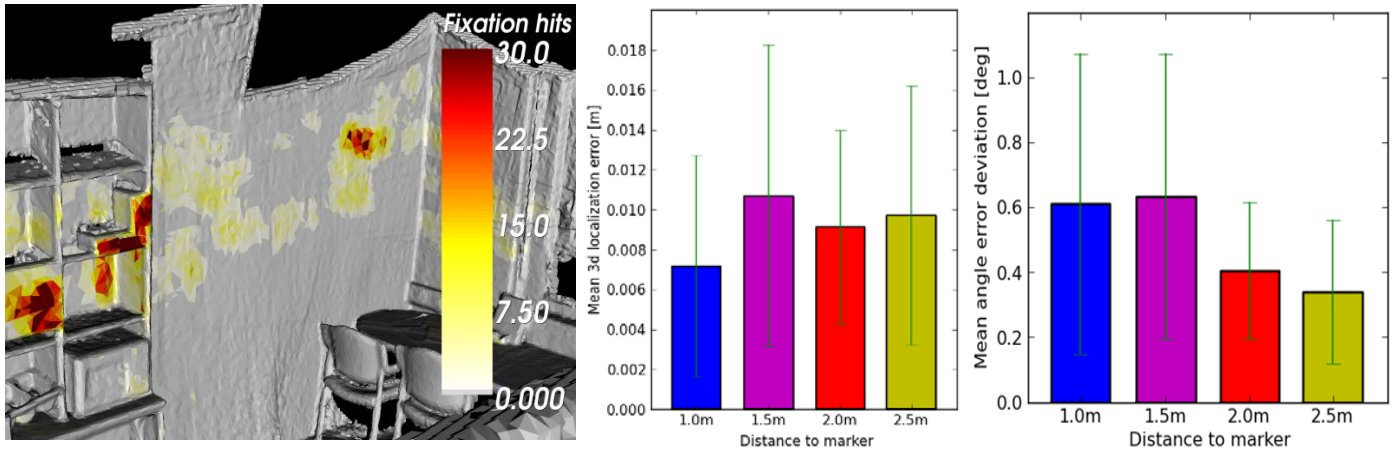
In order to evaluate the attention methodology, videos of two different test users wearing the eye-tracking device have been captured (2200 and 3400 video frames). Human pose estimation and attention computation has been performed as described above. In order to evaluate localization and attention computation accuracy, ETG reference frames were captured in front of three different markers, at four distances (1-2.5m). First, an Euclidean error measure is computed in the three-dimensional space around the marker: we detect and compute the 3D location of the four marker corners. The distance between intersection points and marker corners in the model defines the Euclidean error measure (Figure 4b). Second, an angular error is computed. This measure is based on the pixel distance between the detected marker and the re-projections, using the estimated pose of the reconstructed marker corners (Figure 4c). Because of the less textured test scenery, we find less widespread key points in the image – required for accurate pose estimates - when being closer and more if being more distanced, hence both error measures decrease with distance.

Conclusion

In this work we investigated the innovative contribution of visual SLAM and visual localization for the full and precise recovery of human gaze in 3D environments. We demonstrated that the chosen methodology produces a very small re-projection error that enables to recover the attention semantics of daily activities. This work will be highly relevant for the studying of human visual attention. The aggregation of fixation hits into 3D saliency maps on the environment eventually enables various interpretation avenues, mapping human interest on the local infrastructure, such as, objects and affordances of interest.

Acknowledgements

This work has been partly funded by the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement n°288587 MASELTOV and by the Austrian Research Promotion Agency (FFG) under contract n°832045 Research Studio Austria FACTS.



(a) (b) (c)

Figure 4. (a) Fixation hits projected onto 3D model, (b) mean localization error of the gaze pointer and (c) mean angular error.

References

[1] Pirker, K., Schweighofer, G., Rüther, M., Bischof, H.: GPSlam: Marrying Sparse Geometric and Dense Probabilistic Visual Mapping, *Proc. BMVC*, 2011.

[2] Civera, J. and Davison, A.J. and Montiel, J.: Inverse Depth Parametrization for Monocular SLAM*, IEEE Transactions on Robotics*, vol. 24, pp. 932—945, 2008.

[3] H. Strasdat, A.J. Davison, J.M.M. Montiel, and K. Konolige: Double Window Optimisation for Constant Time Visual SLAM, *Proc. IEEE ICCV*, 2011.

[4] Lowe, David G.: Distinctive image features from scale-invariant keypoints, *IJCV*, vol. 60, pp. 91-110.

[5] Nister, D. and Stewenius, H.: Scalable Recognition with a Vocabulary Tree, *Proc. CVPR*, 2006.

[6] Gottschalk S. & Lin M. C. & Manocha D.: OBB-Tree: A Hierarchical Structure for Rapid Interference Detection, *Proc. 23rd Annual Conference on Computer Graphics and Interactive Techniques*, 1996.

[7] Lepetit V., Moreno-Noguer F. and Fua P.: EPnP: An Accurate O(n) Solution to the PnP Problem, *IJCV*, pp. 155-166, 2009.

[8] Munn, S. M., and Pelz J. B.: 3D POR, position and head orientation from a portable monocular video-based eye tracker. *Proc. ETRA*, pp. 181-188, 2008.

[9] Voßkühler A., Nordmeier V. and Herholz S.: Gaze3D - Measuring gaze movements during experimentation of real physical experiments. *Proc. ECEM*, 2009.

[10] Pirri, F., Pizzoli, M., Rudi, A.: A general method for the POR estimation in 3D space. *Proc. CVPR*, 2011.

[11] Schrammel, J., Döbelt. S., Paletta, L., Tscheligi, M., Attentional Behavior of Users on the Move Towards Pervasive Advertising Elements, in eds., Müller, Alt, Michelis, *Pervasive Advertising*, Springer, 2011.

[12] Paletta, L., and Tsotsos, J.K., Eds., *Attention in Cognitive Systems*, LNAI 5395, Springer, 2009.

[13] Judd, T., Ehinger, K., Durand, F., and Torralba, A., Learning to predict where humans look, *Proc. IEEE ICCV*, 2009.