

DELIVERABLE REPORT

D2.4.2

“Multisensory Usability Engineering”

MASELTOV

Mobile Assistance for Social Inclusion and Empowerment of Immigrants with Persuasive Learning Technologies and Social Network Services

Grant Agreement No. 288587 / ICT for Inclusion

collaborative project co-funded by the
European Commission - Information Society and Media Directorate-General
Information and Communication Technologies - Seventh Framework Programme (2007-2013)

Due date of deliverable:	31 December 2013 (month 24)
Actual submission date:	9 March 2014
Start date of project:	Jan 1, 2012
Duration:	36 months

Work package	WP4 – MULTISENSORY CONTEXT AWARENESS
Task	T4.2 – Multisensory Usability Engineering
Lead contractor for this deliverable	JR
Editor	Lucas Paletta (JR)
Authors	Lucas Paletta (JR), Michael Schwarz (JR), Helmut Neuschmied (JR), Martin Pszeida (JR), Gerald Lodron (JR), Stefan Ladstätter (JR), Patrick Luley (JR), , Florian Eyben (TUM), Felix Weninger (TUM), Björn Schuller (TUM), Stephanie Deutsch (CUR), Jan Bobeth (CUR), Manfred Tscheligi (CUR)
Quality reviewer	Nicoletta Bersia (TI)

Project co-funded by the European Commission within the Seventh Framework Programme (2007–2013)		
Dissemination Level		
PU	Public	X
PP	Restricted to other programme participants (including the Commission Services)	
RE	Restricted to a group specified by the consortium (including the Commission Services)	
CO	Confidential, only for members of the consortium (including the Commission Services)	















CONTACT

Contact for feedback on this report to the project coordinator / editor / author:

lucas.paletta@joanneum.at

Lucas Paletta
JOANNEUM RESEARCH Forschungsgesellschaft mbH
Steyrergasse 17
8010 Graz

© MASELTOV - for details see MASELTOV Consortium Agreement

MASELTOV partner			organisation name	country code
01	JR		JOANNEUM RESEARCH FORSCHUNGSGESELLSCHAFT MBH	AT
02	CUR		CURE CENTRUM FUR DIE UNTERSUCHUNG UND REALISIERUNG ENDBENUTZER- ORIENTIERTER INTERAKTIVER SYSTEME	AT
03	AIT		RESEARCH AND EDUCATION LABORATORY IN INFORMATION TECHNOLOGIES	EL
04	UOC		FUNDACIO PER A LA UNIVERSITAT OBERTA DE CATALUNYA	ES
05	OU		THE OPEN UNIVERSITY	UK
06	COV		COVENTRY UNIVERSITY	UK
07	CTU		CESKE VYSOKE UCENI TECHNICE V PRAZE	CZ
08	FHJ		FH JOANNEUM GESELLSCHAFT M.B.H.	AT
09	TI		TELECOM ITALIA S.p.A	IT
10	FLU		FLUIDTIME DATA SERVICES GMBH	AT
12	FUN		FUNDACION DESARROLLO SOSTENIDO	ES
13	DAN		VEREIN DANAIDA	AT
14	MRC		THE MIGRANTS' RESOURCE CENTRE	UK
15	PP		PEARSON PUBLISHING	UK

CONTENT

Contact	2
Executive summary	5
1. Human factors technologies – contribution to MASELTOV	6
1.1 The role of multisensory usability engineering in MASELTOV	6
1.2 First contribution: eye tracking for mobile interaction analysis	6
1.3 Second contribution: supporting online dialogue evaluation.....	6
1.4 Third contribution: 3D gaze recovery to measure embodied attention	7
1.5 Fourth contribution: application in user studies	7
2. Eye tracking for mobile interaction analysis	8
2.1 Introduction	8
2.2 Related work.....	9
2.3 Smartphone Eye Tracking System (SMET)	10
2.4 Conclusion – MASELTOV service aspects.....	14
3. Supporting online dialogue evaluation.....	16
3.1 Introduction	16
3.2 The GRAS ² database.....	16
3.3 Eye contact and acoustics	16
3.4 Conclusion – technological aspects	17
3.5 Conclusion – MASELTOV service aspects.....	17
4. 3D gaze recovery to measure embodied attention	19
4.1 Positioning of user attention in arbitrary environments.....	19
4.2 Conclusion – MASELTOV service aspects.....	20
5. Application in user studies	22
5.1 User study on mobile interaction in navigation applications	22
5.2 User study on mobile dialogue evaluation	22
5.3 User study on interaction with environments	22
6. Conclusions	22
7. Acknowledgments	23
8. References	24

EXECUTIVE SUMMARY

Task 4.2 Multisensory Usability Engineering (Lead: JR (13 PM); CUR (2 PM))

In order to develop a correct understanding of the independent expert's interactions, the evaluation of the functional prototypes that add in to the qualitative means of usability engineering, this task considers applying methods for quantitative assessment of the performance, using state-of-the-art of mobile eye tracking hardware and innovative methods in the interpretation of mobile eye tracking videos, in correlation with other sensors, such as, audio, motion, and psychophysiological sensors.

This second and final report of results in Task 4.2 describes the full extent of the development of innovative technologies, together with giving an account on the plan about the user studies which will be performed in the remainder of the project duration, in the frame of WP9, i.e., field trials and evaluation.

A key issue is certainly the provision of a completely new technology to become capable to perform an automated interpretation of large user studies with eye tracking, even in challenging outdoors studies. This technology, aiming at capturing the interaction experience using smartphones, enables (i) to perform the evaluation of the usability of interfaces under specific consideration of the attention processes within, (ii) to enable to deduce facts on cultural diversity from behaviour and attentive perception, and (iii) to monitor the tracks of experts and from this analyse the mobility.

The report is structured, as follows. Section 1 presents an overview on the developed human factors technologies and their key contribution to MASELTOV. Section 2 describes in detail the novel technology for mobile interaction and the implications on large user studies. Section 3 presents details of the individual technologies involved in supporting dialogue evaluation. Section 4 presents an outline of the work plan for user studies which are out of scope of work package WP4 on multisensory context awareness. These user studies will be drawn in the frame of WP9, field trials and evaluation, as part of the iterative evaluation of user interfaces and services. Section 5 concludes with a discussion and an outlook on promising future pathways for technology development.

1. HUMAN FACTORS TECHNOLOGIES – CONTRIBUTION TO MASELTOV

1.1 THE ROLE OF MULTISENSORY USABILITY ENGINEERING IN MASELTOV

Cultural diversity is a key theme in the project since MASELTOV involves the entering of an immigrant culture into a local context, supporting the language based communication between immigrants and local citizen, and in this sense focuses on the aspects of information retrieval and personal communication under very different cultural contexts.

The objective of employing multisensory usability engineering in MASELTOV was to develop novel technologies that index into a more profound understanding of human-computer interfaces in the context of multicultural diversity, from the viewpoint of the requirements of immigrants.

In the following, we present the three major technologies that have been developed in MASELTOV, together with the application context, and an outlook on the user studies that will further on be applied using these novel technologies.

1.2 FIRST CONTRIBUTION: EYE TRACKING FOR MOBILE INTERACTION ANALYSIS

Usability research on interaction with mobile devices has been thoroughly investigated in recent years and established a relevant application field in eye tracking research. However, current technologies for the mapping of point-of-regards (PORs) to mobile displays do not enable natural interaction with the mobile device: users are conditioned to interact with tightly mounted displays, or are distracted by markers in the field of view. We propose a novel approach for usability engineering that enables fully natural interaction with the mobile device, using eye tracking glasses to capture the POR-annotated video, with computer vision for highly accurate tracking of mobile devices. Our approach enables 2D information recovery of the POR on the display and continuous dynamic attention heat mapping of the visual task. In a benchmark test we achieve a mean accuracy of POR localization on the display of $\approx 1.5 \pm 0.9$ mm and conclude that the methodology could open new avenues for eye tracking research in the field of mobile applications.

1.3 SECOND CONTRIBUTION: SUPPORTING ONLINE DIALOGUE EVALUATION

An important aspect in the evaluation of short dialogues is how attention is manifested by eye-contact between subjects. Eye-contact in principle might serve as an indicator for mutual understanding, for displaying that two communicators are of an equal social value hence indicating social inclusion, and for the level of concentration in the person who leads the dialogue. All this factors have not been explored in full detail but it is assumed that eye-contact will play a specific, to-be-researched role in the mentioned characteristics.

In this study we provide a first analysis whether such visual attention is evident in the acoustic properties of a speaker's voice. We thereby introduce the multi-modal GRAS² corpus (a database including eye tracking and audio information, “Graz Real-Life Affect in the Street & Supermarket”), which was recorded for analysing attention in human-to-human interactions of short daily-life interactions with strangers in public places in Graz, Austria. Recordings of four test subjects equipped with eye tracking glasses, three audio recording devices, and motion sensors are contained in the corpus. We describe the robust identification of speech segments from the subjects and other people in an unsupervised manner from multi-channel recordings – this study has been managed, implemented by major part by Technische Universität München, and during a visit of Prof. Björn Schuller (project ASC-INCLUSION

under the coverage of DGEI clustering, Task 10.2) when he was employed at JOANNEUM RESEARCH in Graz, Austria (August-October 2012). We then discuss correlations between the acoustics of the voice in these segments and the point of visual attention of the subjects. A significant relation between the acoustic features and the distance between the point of view and the eye region of the dialogue partner is found. Further, we show that automatic classification of binary decision eye-contact vs. no eye-contact from acoustic features alone is feasible with an Unweighted Average Recall of up to 70%.

1.4 THIRD CONTRIBUTION: 3D GAZE RECOVERY TO MEASURE EMBODIED ATTENTION

The study of human attention in the frame of interaction studies has been relevant for usability engineering and ergonomics for decades. Today, with the advent of wearable eye-tracking and Google glasses, monitoring of human attention will soon become ubiquitous. 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 centred measurements directly into an automatically computed 3D model. We apply RGB-D SLAM (a method that applies *S_imultaneous Localization And Mapping* by use of 2D and 2½D information) and descriptor matching methodologies for the 3D modelling, localization and fully automated annotation of ROIs (regions of interest) within the acquired 3D model. This methodology brings new potential into automated processing of human factors, opening new avenues for attention studies, bringing immigrants behaviour into context with the specific environments of immigrants.

1.5 FOURTH CONTRIBUTION: APPLICATION IN USER STUDIES

We provide a brief account on the plan for user studies as they will be implemented in order to take advantage of the novel human factors technologies in order to extract novel aspects of diversity engineering.

2. EYE TRACKING FOR MOBILE INTERACTION ANALYSIS

2.1 INTRODUCTION

The investigation of mobile human-computer interfaces by means of eye tracking research has been continuously developed in recent years. The evaluation of interaction designs in mobile applications requires the investigation of eye movements, similar to the use case of usability analysis of static websites and interaction schemes [Jacob & Karn, 2003].

Mobile computing technology allows having electronic devices available whenever and wherever we want. However, designers and developers of mobile applications like palmtop computers, PDAs, and mobile phones have to face unique challenges, because location and environment are usually less predictable than in desktop applications [Barnard et al. 2007]. Similar restrictions apply also to public displays. Furthermore mobile computing devices have the common problem of rather small visual displays and limited input techniques, wherefore performance is often substantially worse than in the desktop context (e.g., [Neerincx & Streefkerk, 2003]. Multitasking and support for task interruption are of high relevance, since in a mobile context the frequency of distracting events is much higher than for a desktop application and tasks with interruptions take longer to complete on a mobile device than with a desktop application [Nagata 2003]. Therefore special interest is on the increased competition with regard to attracting the users' attention and on interaction as a non-primary task in a certain context [Schrammel et al., 2011].

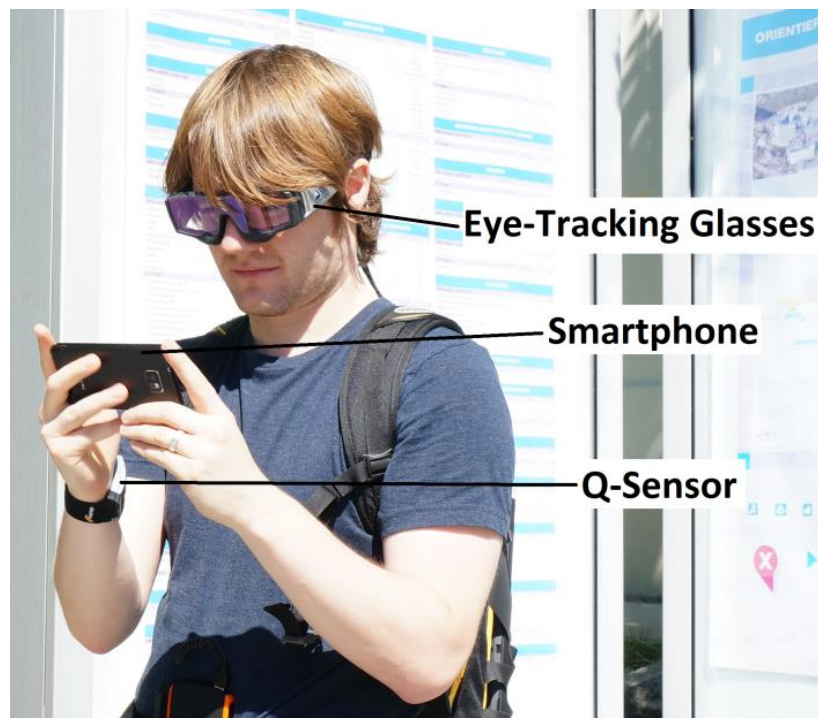


Figure 1 User participating in the benchmark study with typical sensor setup: eye tracking glasses, smartphone and wristband sensor for physiological data. We propose in this paper the fully automated recovery of POR localization on mobile devices. Future work will make use of GPS to map attention data add psychophysiological data for more profound task analysis. In the picture, the MASELTOV navigation component is tested by an expert user. A large user study with more than 20 persons, among them immigrants and a local user based control group, will be run with the developed technology described in this report.

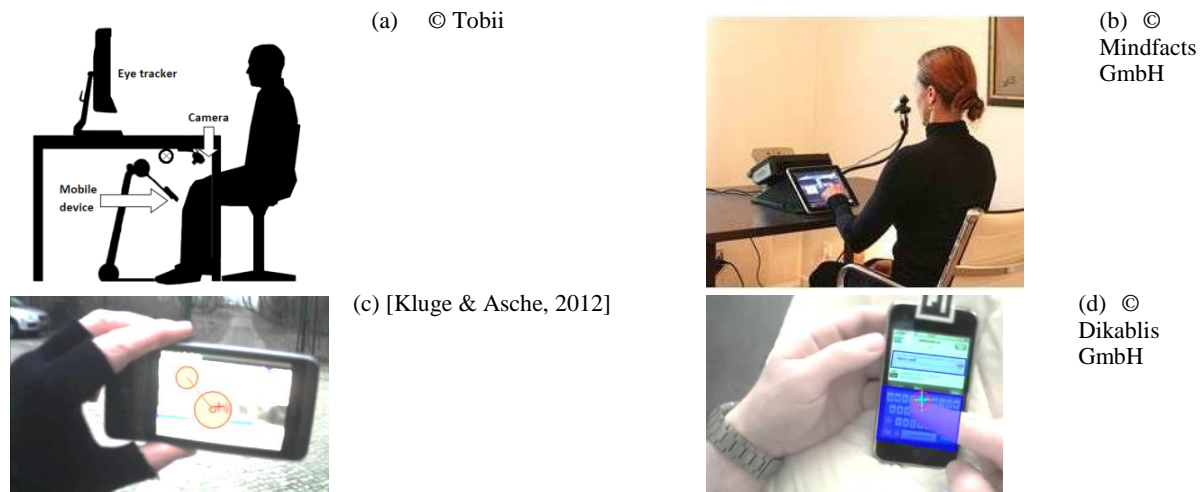


Figure 2 Experimental setups used in usability oriented eye tracking research on mobile interaction. (a) Tobii setup of phone mounted underneath the table. (b) Mindfacts setup on top of the table. (c) Manual Annotation with BeGaze (SMI). (d) Automated annotation with markers.

However, current technologies for the mapping of point-of-regards (PORs) to mobile displays do not enable natural interaction with the mobile device. Mobile users in usability research are either conditioned to act with tightly mounted displays or distracted by markers in the view (Figure 2d). Markers are artificial landmarks that are placed in the scene, consequently they can be easily detected by computer vision methods, however, one relies on placing them properly, knowing where the action is, and the user is certainly distracted by their appearance since they are easily detectable. These methods capture attention behavior that is dependent on the distraction of marker patterns or forbid to act in urban spaces appropriately, which is a major concern for usability analysis.

We propose a novel approach that enables fully natural interaction with the mobile device and its application (Figure 1), using eye tracking glasses (ETG) to capture the POR annotated video, and computer vision for highly accurate tracking of the mobile device. Our approach enables 2D information recovery of the POR on the display and from this continuous dynamic heat mapping of the visual task. In a benchmark test we achieve a mean accuracy of POR localization on the display of $\approx 1.5 \pm 0.9 \text{ mm}$ and conclude that the methodology could open new avenues for eye tracking research in the field of mobile services.

2.2 RELATED WORK

Eye tracking devices have already been used in various combinations with smartphones. Mapping of PORs to mobile displays is usually performed - in indoor studies - by measuring with tightly mounted phones (Figure 2a,b). For example, [Cauchard et al., 2011] screw the mobile device under a table, and enables evaluation of mobile interaction only when sitting (Figure 2a). [Chynal et. al. 2012] investigated and proved the significance of eye tracking in mobile applications usability testing, using a head mounted eye tracker. Outdoors mobile applications, such as navigation, require to focus the user's attention quickly on the desired interface functionality, since the frequency of distraction through events and context changes may cause inattention of the user to relevant information on the mobile display [Rohs et al., 2007]. For these circumstances, [Kluge & Asche, 2012] developed an eye tracking pilot test on validating a smartphone based pedestrian navigation system, describing the relationship between reality and navigation instructions. Evaluation of eye movements was done on the basis of massive manual intervention using the SMI BeGaze Software.

Alternative approaches use the smartphone camera directly with its rear view on the user to estimate the eye gaze on the mobile phone [Miluzzo et al., 2010]. However, so far this method achieves accuracies only in the few centimeters range, is vulnerable to harsh illumination conditions, and, furthermore, does not provide a video about the backstage environment as well.

[Fritz & Paletta, 2010; Mardanbegi, D. & Hansen, 2011] demonstrated the use of static display localization in eye tracking tasks. However, their approach is rather specific and does not apply to the localization of mobile devices in eye tracking videos. [Gehring et al., 2012] mentions the concept to detect mobile devices but does not provide quantitative information on experimental data. [Giannopoulos et al., 2012] mentioned Dikablis based marker tracking (Figure 2d) which principally distracts the user's attention.

2.3 SMARTPHONE EYE TRACKING SYSTEM (SMET)

We propose a complete software toolbox package in terms of a smartphone eye tracking (SMET) system that will enable fully automated analysis of attention in user studies. Figure 3 depicts a schematic sketch of the information flow in the system: data capture is applied with non-invasive wearable interfaces. Synchronisation and image analysis provide then a correlated data stream with smartphone events. Geometric transformation and heat mapping provide then the basis for attention analysis.

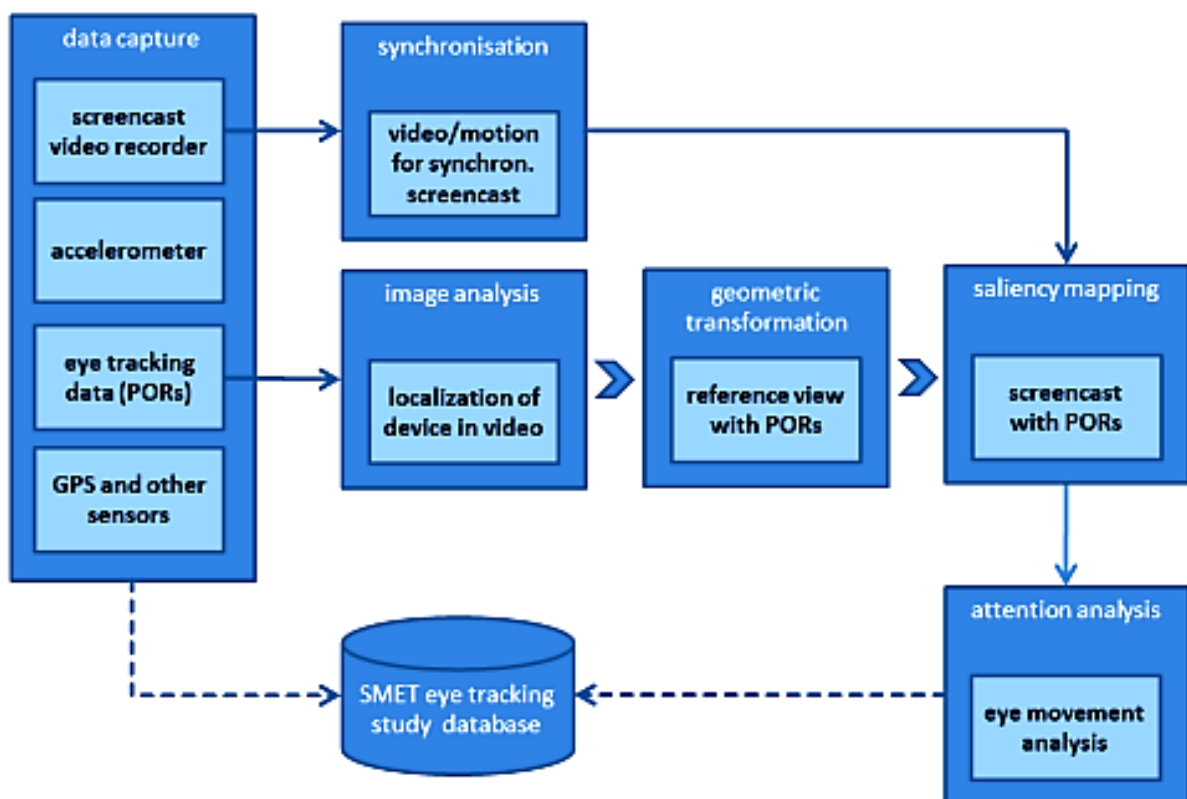


Figure 3 Information processing in the SMET (smartphone eye tracking) system.

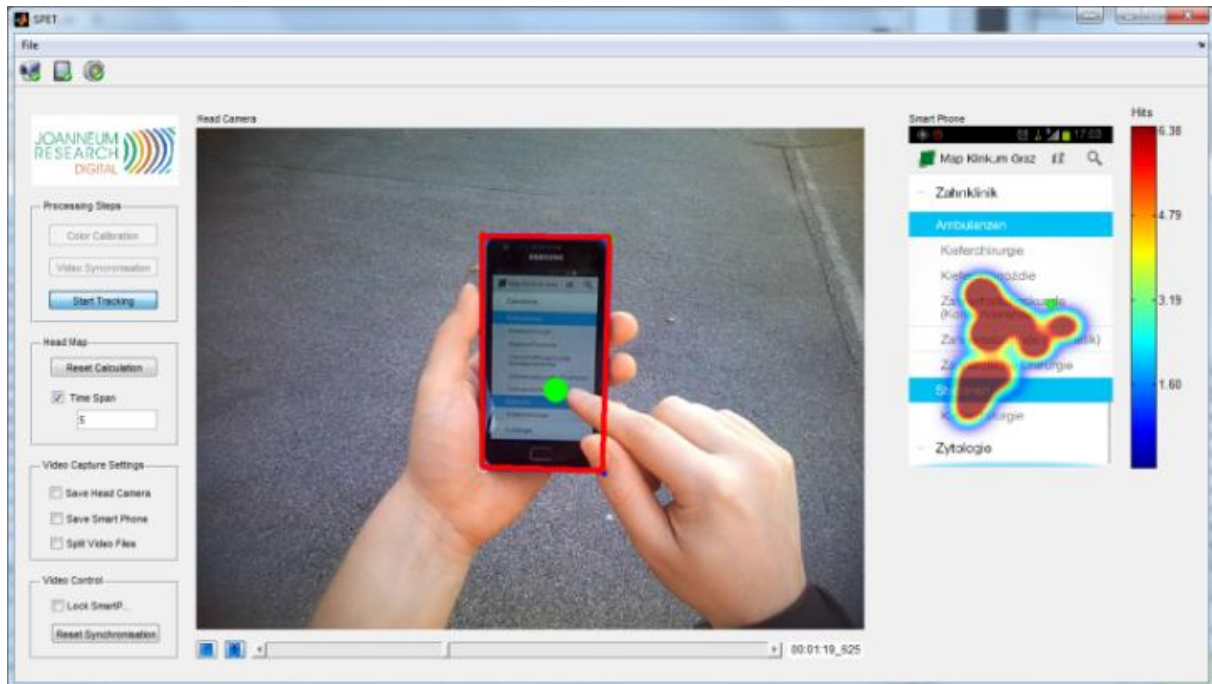


Figure 4 *Graphical User Interface of the SMET application.*

2.3.1 THE GRAPHICAL USER INTERFACE

The Graphical User Interface (GUI) of the SMET application (Figure 4) requires three kinds of input data: the scene camera based video, the smart phone video, and the captured POR data. Firstly, points that lie on image area about the coloured phone case are manually selected for initialization, but further then the colour thresholds which are required by the tracking algorithm are automatically determined. The two videos are then synchronized in a semi-automated manner. After that, the automated smartphone tracking process is processed. Once the phone image area is tracked, an affine transformation matrix is computed to assign the POR location in the scene camera image to a location in the phone image. This is used to calculate a heat map of fixations at the smart phone user interface. For the heat map calculation, the POR location information is blurred by a Gaussian function and accumulated over a specific time period. Heat map calculation can be at any time instant.

2.3.2 IMAGE BASED MOBILE DEVICE LOCALIZATION

To enable robust tracking the smart phone is used with a colour intensive case in order to identify the device well within the scene video (Figure 5a). Through the detection of the coloured frame lines and the corner points the image segment of the smartphone can be identified and visually tracked.

The tracking algorithm has the following processing steps, use of colour threshold values, which are manually selected in a first frame, and then automatically adapted, so that colour based segmentation is performed in different colour spaces (HSV, YCbCr) - Figure 5b. After noise removal by morphological filter operations, the remaining image regions are reduced to skeletons of the regions by a thinning algorithm [Guo & Hall, 1989]. The resulting lines are then extracted based on Hough Transform [Duda & Hart, 1972]; noisy lines will be removed and intersection points from the remaining lines be found. Thereby corner points of the smart phone are detected, which are validated by previous tracking results and geometrical constrains of the smart phone area. The tracking algorithm is based on a linear prediction which performs well for a random movement of a single object [Yeoh & Abu-Bajar, 2003].

The position of missing corner points can be estimated based on the movement of the detected points and the position prediction of the tracking process. This yield to localization results even if fingers cover parts of the phone area or if only a part of the phone area can be observed within the image. Through the continuous adaptation of the colour thresholds based on actual tracking results, the detection of the smart phone is robust in fast changing lightning Figure 5c). The tracking of the smart phone can fail in case of a change from an indoor to an outdoor scene or vice versa. This can be happen if the phone is not visible during this location change. Then a recalibration is required.

2.3.3 PROOF OF CONCEPT IN A NAVIGATION STUDY

The study targeted towards a proof of concept with a limited number of users, in order to understand which level of accuracy could be achieved. For this purpose, a typical user of mobile applications was involved several times in an outdoor navigation task to find the department of a local hospital. To acquire video and eye-tracking data we used SMI Eye Tracking Glasses, with 30 Hz sampling rate of gaze and a 1280 x 960 pixel resolution scene camera.

Figure 6 depicts (a) the camera view from the Eye Tracking Glasses, overlaid with the bounding box of the smartphone localization estimate from the SMET system; (b) respective heat map from the eye movements mapped onto the mobile display, visualizing the most attended locations in the display.

Table 1 presents most relevant evaluation results from 3 different test runs using the mobile service, under different weather conditions (2 times sunny and 1 time cloudy weather – test number test3 in Table 1). A substantial subset of the number of video frames represents the interaction of the user with the mobile phone. However, the detection method was applied to the total number of frames. The precision rate of the method was sufficiently high to enable a stable and cost-efficient service. We compared the errors that contribute to a final precision estimate. The calibration error was estimated to be $\approx 3\text{mm}$. The human error in the ground truth annotation was $\approx 2\text{mm}$. In this context, the localization error of the SMET system fitted nicely to be $\approx 1.5 \pm 0.9\text{ mm}$, considering the icon button size of $7 \times 7\text{mm}$ in standard (Samsung) displays. Figure 7 depicts the localization error in millimetre (important for mobile HCI) and in pixels (important to compare computer vision tasks). The errors have been determined by difference of estimated to human annotated ground truth corner data, the maximum error has been estimated to be 19 mm (Figure 6c).

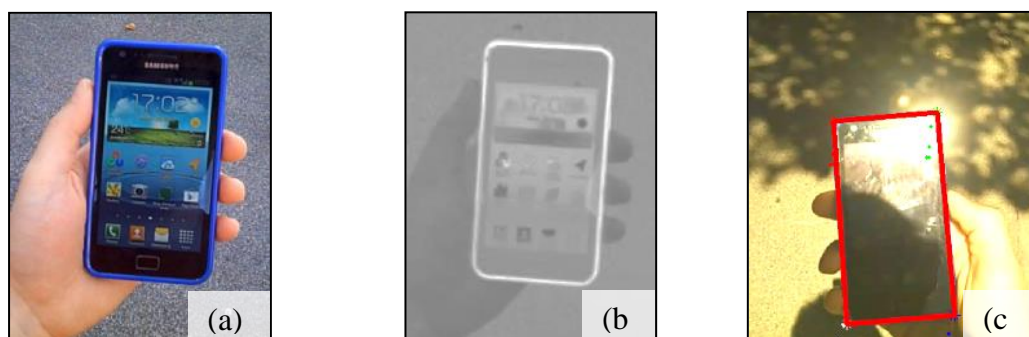


Figure 5 Image of a smart phone with a colour intensive case (a) and the corresponding image in YCbCr colour space (b). Tracking results at different light conditions. (c) The method works robust, even under harsh outdoor illumination conditions.



Figure 6 Automated annotation and heat mapping from an outdoor mobile navigation task.

Table 1 Accuracy of POR mapping on smartphone display.

Benchmark parameter	Test 1	Test 2	Test 3
Total number of frames	4284	6531	3869
Frames with user interaction	1913	1318	3479
False positives on total sequence:	4	0	0
True positives on total sequence	1913	1318	3479
Precision (pos. pred. value) [%]	99.8	100	100
Negative predict. value [%]	95.9	96.3	39.0
Mean [pixel]	3.3	5.2	2.9
Standard deviation [pixel]	±2.2	±8.1	±9.7
Mean [mm]	1.5	1.4	1.4
Standard deviation [mm]	±0.9	±1.4	±2.5

2.3.4 DISCUSSION

We presented a novel and robust approach for marker free tracking of smartphones in the head video of eye tracking glasses. The automated localization will open new avenues for further ideas in the frame of automated processing, to investigate the attention, human factors in general in natural, intuitive interaction, and within the task of interest. The precision of the

approach is sufficient to enable usability analysis in the context of decision choices (buttons) in the display.

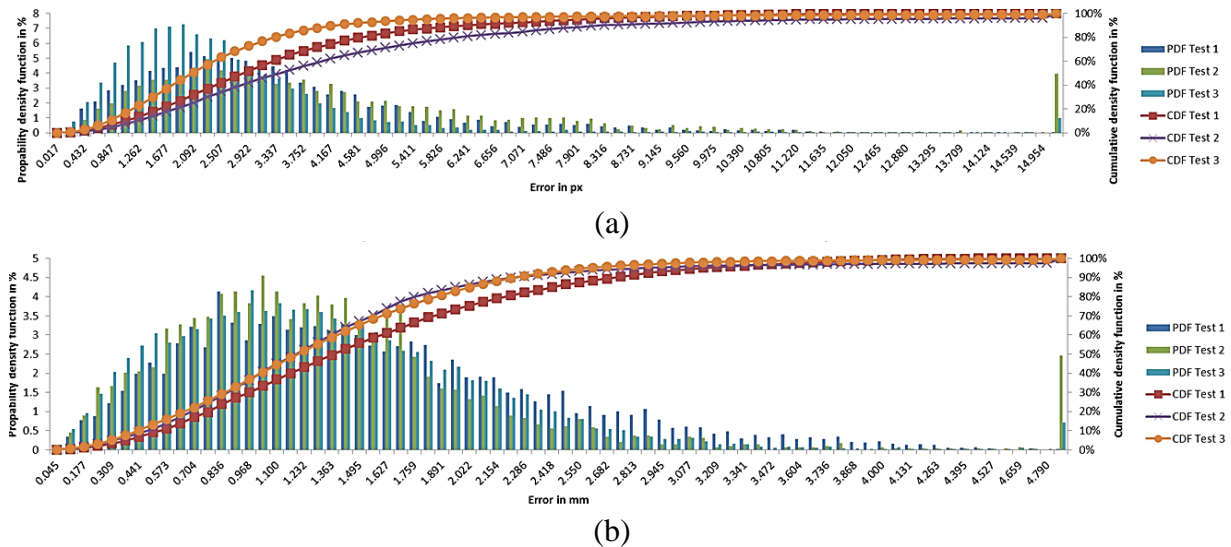


Figure 7 Precision of the localization methodology: (a) Error in [pixel] of the automated localization within the video. (b) Error in [mm] on the display of the mobile device. The accuracy was evaluated in 3 test sequences. Mean error on the display is ≈ 1.5 mm.

2.4 CONCLUSION – MASELTOV SERVICE ASPECTS

The smartphone eye tracking (SMET) toolbox has a high potential for the analysis and design of MASELTOV user interaction components. Human attention processes are directly involved in the selection behaviour and the reaction time, in the accessibility and the usability of the device and the service per se.

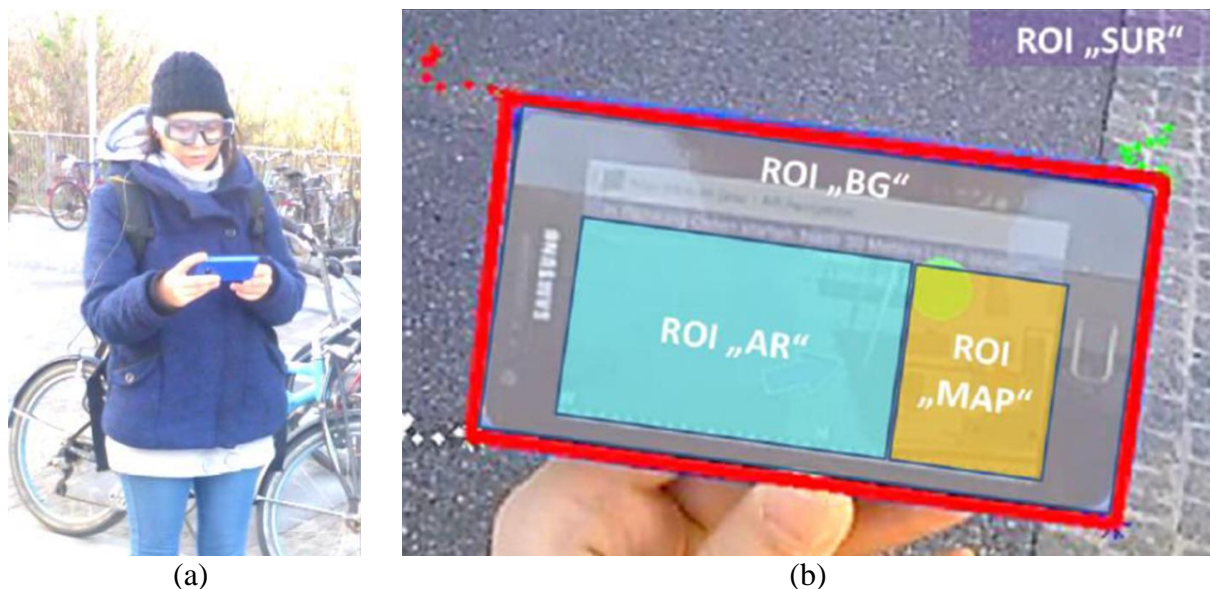


Figure 8 A study using the SMET toolbox on the MASELTOV component 'augmented reality navigation service' is run in Graz. (a) depicts an immigrant user applying the service in the study area, i.e., the hospital "LKH" in Graz (owned by the province of Styria). (b) visualizes the various ROIs as briefly described in Sec. 2.4.

Unlike any previous analysis method, it enables to investigate the human fixations and saccades directly on the display, with a precision that was not achievable before. It enables in principle a fully automated analysis of human attention in interaction with the specific service based user interface, and can be complemented by human annotation, if the illumination conditions or any other condition of the user or the environment would prohibit an accurate matching of the human point-of-regard with the mobile device's display. However, in all experiments and so far implemented applications, the service provided a very well, very continuous service and human intervention for annotation was not necessary.

In Graz, a study is undertaken to test the interaction design of the MASELTOV component “augmented reality navigation”, for navigation in a province owned hospital in Graz, one of the largest by the area it covers in central Europe. The study will be run with 16 immigrants (Turkish women) and 16 local users (female, same age). Figure 8a depicts a typical immigrant smartphone user. Figure 8b visualizes the various ROI that are under investigation using the eye tracking glasses. Table 2 depicts first results that indicate the distribution of transitions – in the sense of transition probabilities that are captured from experience – of eye movements between ROIs. It was experienced from the experiments that, for example, the transition from region “AR” to “MAP” (18%) was substantially lower than the transition from “MAP” to “AR”. From this fact one can conclude that looks on the map based service (“MAP”) involve looks to the surrounding to a higher degree than looks on the augmented reality base service (“AR”) does, which might indicate that a map based service requires more verifications from the surrounding than an augmented reality based service. The study is still on-going and will be finalised until the end of the first field trial, so that any conclusions on the interaction design can be fed into the final technical components update of the MASELTOV service.

Table 2 Matrix indicating the distribution of transition probabilities over various regions of interest, between ROIs “AR” (augmented reality information window), “MAP” (map based information window), “BG” (background within the display, mostly information about time, battery, etc.) and “SUR” (the surrounding apart from the display of the mobile service, i.e., the environment).

ROIs	AR	MAP	BG	SUR
AR	0%	18%	37%	45%
MAP	37%	0%	6%	66%
BG	17%	8%	0%	56%
SUR	46%	13%	58%	0%

3. SUPPORTING ONLINE DIALOGUE EVALUATION

3.1 INTRODUCTION

An important aspect in short person to person dialogues is attention as is manifested by eye-contact between subjects. Thus, to replicate human-like behavior for artificial systems (e. g., humanoid robots or virtual agents) it is believed to be highly important to implement natural patterns of eye contact (Miyauchi et al., 2004). Furthermore, consistency between eye contact and acoustic cues emitted by a system, e. g., by means of speech synthesis, should be ensured. In this study we verify the correlation between visual attention and acoustic cues of a speaker's voice in human to human dialogues. Such information could be used in low resource, or speech only systems which do not have a camera or eye tracking device available, e. g., in voice conversations and chats the other partner could be informed about the eye-gaze behavior of the first partner without actually seeing him/her. Also in forensic analysis these methods could be applied. Further, psychological studies may profit from such knowledge.

3.2 THE GRAS² DATABASE

The Graz Real-Life Affect in the Street & Supermarket (GRAS²) corpus is - to the authors' best knowledge – the first database of visual attention recordings with multiple audio-visual, physiological, and movement sensory cues in real-life conversations. Four subjects took part in the recordings (3 female, 1 male, cf. Figure 9). These were all native Austrian students and they filled a BFI-11 personality questionnaire (Rammstedt and John, 2007). The male subject usually wears glasses and the female subjects did not wear glasses.

3.3 EYE CONTACT AND ACOUSTICS

From the eye tracking glasses we can extract the position where the subject is looking at in the coordinate system of the eye tracker frontal camera. From the video of the frontal camera we detect the presence of a face (frontal view) with the openCV face detector (openCV is a publicly available computer vision methods library) based on Local Binary Pattern (LBP) features and try to estimate the eye region within the face with a Haar wavelet based eye detector also available in openCV. If no eye region was detected in the face (e. g., if people wear glasses), we estimate the eye region from the face region as:

$$\begin{aligned} - \quad X_e &= x_f + 0.25w_f & (1) \\ - \quad Y_h &= y_f + 0.25h_f & (2) \\ - \quad W_e &= 0.5w_f & (3) \\ - \quad H_e &= 0.16h_f; & (4) \end{aligned}$$

where the subscript e indicates the eye region bounding box and the subscript f the face region bounding box. X, y, w, and h are the coordinates of the upper left corner, the width, and the height of the bounding box, respectively. By combining the eye tracker coordinates with the detected face and eye region, we can define three classes for where the subject is looking with respect to the partner: Direct eye contact - i. e., looking into the eye region (V_e), looking into the face region (V_b), or looking next to the face region in a corridor with 0.5 width/height to the left, top, right, bottom of the face region (V_a). Additionally, we compute the Euclidean distance between the centre of the detected eye region and the point the subject is looking at. This is referred to as eye-eye distance in the following. If no face is detected in the image, a

maximum value is filled in for this distance. To produce an eye-contact ground truth per speech segment, we apply the following rule in this particular order: If for at least 2 frames there is direct eye contact (V_c), we assign the V_c label to the whole segment. Otherwise, if for at least 2 frames there is case V_b , we assign label V_b , and otherwise the same for V_a . If neither case is present in the segment we assign the label V_n for no eye contact. Detailed statistics on the amount of eye contact in the segments where the subject is talking are found in the paper in the appendix B. There are notable differences between the subjects in terms of eye contact behaviour. Subject A apparently has the most eye contact with his partners, durations of cases V_a and V_b are almost 1 second on average for a two second average segment duration, while for subjects B and C it is only .3 seconds and for D .7 seconds.

3.4 CONCLUSION – TECHNOLOGICAL ASPECTS

We have introduced the GRAS² corpus (Figure 9), a multi-modal and multi-sensory corpus of real-life interactions of people seeking for help and directions from strangers in a public shopping centre. The corpus has been recorded for the purpose of analysing the role of visual attention and dialogue behaviour in such interactions. Using information from multiple audio tracks we were able to automatically label when the subject carrying the recording equipment or his or her dialogue partner is talking. The analysis of correlations between acoustic features of the voice of the subject and the visual attention (eye contact with dialogue partner) has revealed a low, but meaningful correlation between the acoustics and the distance of the point at which the subject is looking and the eye region of the dialogue partner. Yet, the correlations are strong enough, such that an automatic classification of whether a subject is looking at or close by the head of the dialogue partner or somewhere else based only on automatically extracted acoustic speech parameters is feasible with up to 70% unweighted average recall rate (the chance level would be 50 %).

In future work we aim at significantly increasing the size of the corpus by conducting new recordings with the same setup. We will further manually correct the automatic segmentation and conduct experiments on the short interactions to look at the style of the interactions and analyse the reactions and emotions of the dialogue partners. More information in Appendix B.

3.5 CONCLUSION – MASELTOV SERVICE ASPECTS

From the viewpoint of MASELTOV services, this study was able to prove that relevant parameters of a dialogue can be captured and analysed via mobile multisensory technology, in this case, by audio signal analysis only. The basic idea to project - otherwise very complex sensor data – onto a single data stream that can be easily captured in the field, i.e., by microphone, has the consequence that future developments can intrinsically model the eye contact factor into their technology and take advantage of this important feature of a dialogue.

This aspect has certainly been a research part in MASELTOV and has so far no implications on concrete applications and services of the M-app. However, it demonstrates the dimension of audio-visual data analysis, in particular, as it is the innovative part, using the head camera of the mobile eye tracking devices, and perform decision making on the eye movement analyses. The study is a starting point for further research and studies on cultural aspects of communication in dialogues, and performance analysis in second language acquisition, as mentioned in Deliverable D4.2.1 before.

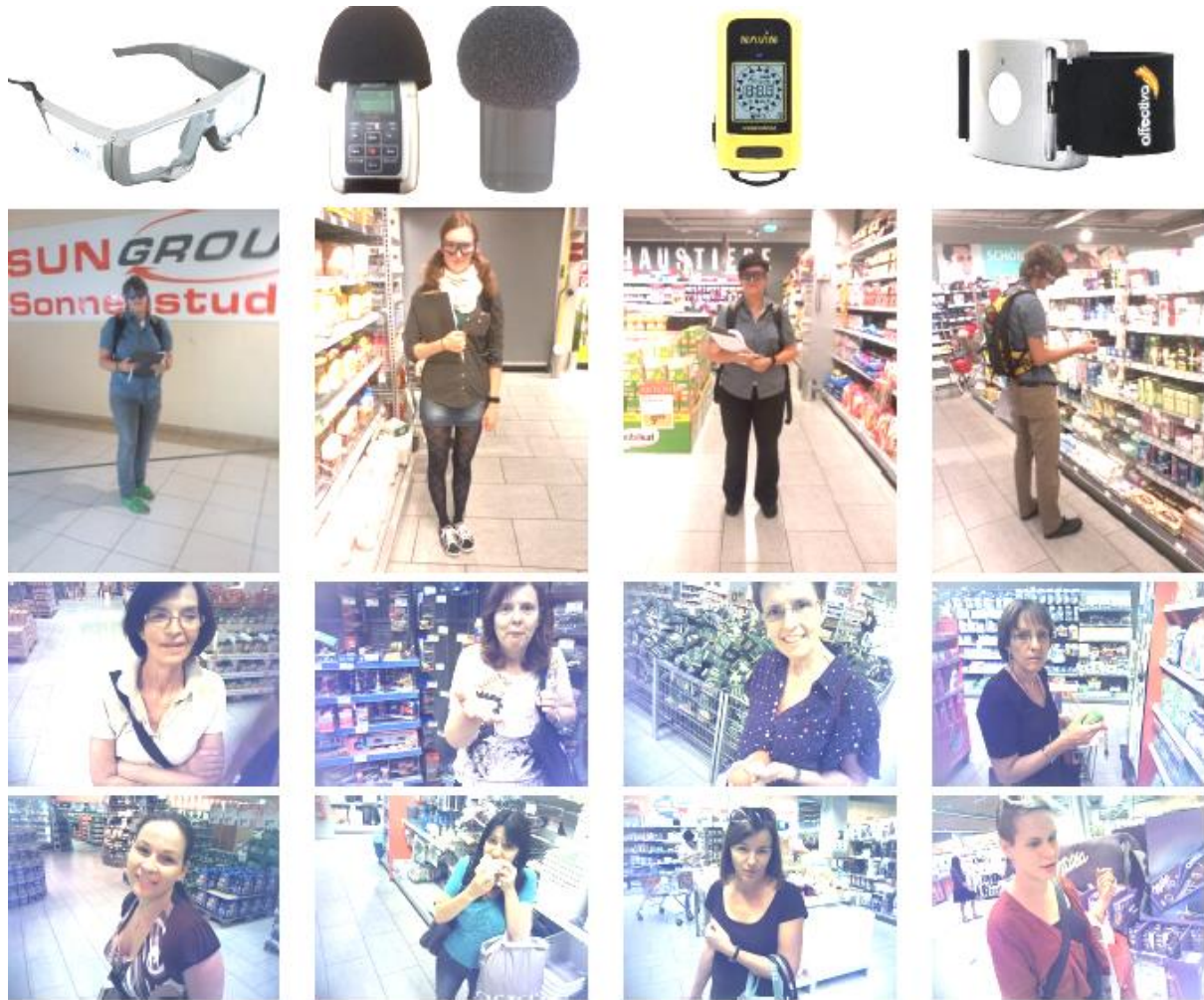


Figure 9 Top-most row: Recording equipment worn by the subjects (from left to right: eye tracking glasses, audio recorder and smart phone, GPS tracker (used for extra accelerometer in backpack), EDA sensor). Second row: The four participating subjects as equipped on site. Bottom two rows: examples of recorded dialogue partners as seen by the four subjects through their worn eye tracking glasses. More details in appendix B.

4. 3D GAZE RECOVERY TO MEASURE EMBODIED ATTENTION

4.1 POSITIONING OF USER ATTENTION IN ARBITRARY ENVIRONMENTS

Within the last couple of years miniaturized mobile eye-tracking systems have become available and been successfully applied in different areas of with the major advantage to evaluate attention in the field where the task of interest is performed. There are some serious limitations that have so far restricted the application of eye-tracking technology to small studies that consequently provide less significant results. A major disadvantage of existing eye-tracking technology is the need to manually analyze the huge amount of collected video data which so far has prevented to perform large and thus more significant user studies. Another important disadvantage of existing systems is that they do not provide any support for alignment of the eye-tracking data, in particular the direction of gaze, with the physical location of the user or with virtual models of the environment.

In our work (Paletta et al., 2012; Paletta et al., 2013; Santner et al., 2013) we seek to attain alignment of the human gaze data within a 3D reconstruction of the environment by recovering the full *6 Degrees of Freedom (DOF)* pose – 3 degrees of freedom of the position in x, y, z world coordinates, and 3 degrees of freedom of the pose of the camera plane in ϕ, θ, ξ variables of angles from tilt, pan and roll rotation - and the sensor should provide three-dimensional information of the environment which has only been shown for small AR workspaces before (Klein & Murray 2007). This means that the human gaze is fully reconstructed and mapped towards the 3D model of the environment and once the environment has been annotated before, the semantics of the gaze trajectory can be reconstructed thereafter as well.

For this purpose we apply the Simultaneous Localization and Mapping (SLAM) framework where one seeks to generate a map using noisy environment information while simultaneously localizing oneself within this map. This is a basic prerequisite for most autonomous robotic applications like path planning or navigation. -based visual SLAM systems perform reliably and fast in medium-sized environments. Currently, their main weaknesses are robustness and scalability in large scenarios. In this work, we propose a hybrid key frame based visual SLAM system, which overcomes these problems. We combine visual features of different strength, add appearance-based loop detection and present a novel method to incorporate non-visual sensor information into standard bundle adjustment frameworks to tackle the problem of weakly textured scenes. On a standardized test dataset, we outperform EKF (Extended Kalman filtering) based solutions (Davison 2005) in terms of localization accuracy by at least a factor of two. On a self-recorded dataset, we achieve a performance comparable to a laser scanner approach.

Figure 10 visualises the effect of the mapping methodology: (a) an avatar is visualized that projects the estimated gaze vector towards the modeled 3D environment. (b) The same situation as it has been actually grabbed into the video frame, with the calibrated fixation point (POR, point of regard) being presented as a blue dot and circle around. It is apparent that the estimation of the 3D model is very accurate since the visualization in the model (a) and the real visual experience in the video frame (b) have a high coincidence where the gaze pointer meets the environment (here: the shelf).



Figure 10. *Reconstruction of human attention in terms of a mapping of attention on infrastructure of the environment.*

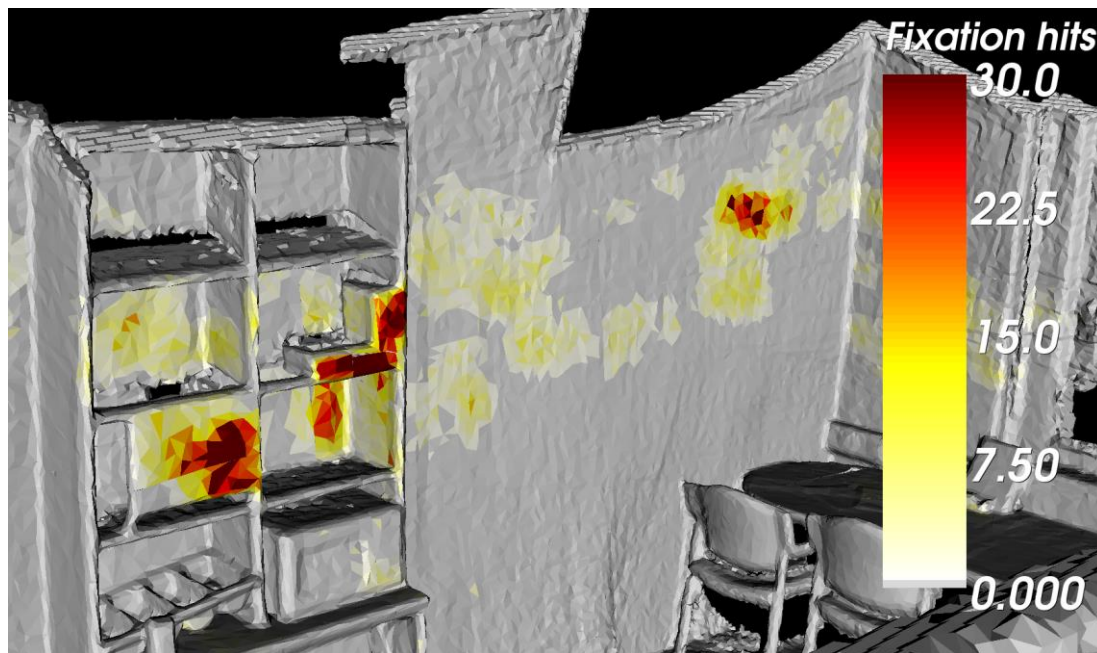
Figure 11a depicts another benefit from the automatic modeling of the environment and the localization of the eye gaze: it becomes possible to collect saliency information over time, i.e., every time the user fixates a location of the environment, that part becomes more salient to the user. In this Figure, it becomes visible that it is possible to derive from these data a heat map (a) directly on the infrastructure of the environment that either reflects the collection of experience of the individual or even from a group of users. This mapping is highly beneficial for conclusions on whether specific signs in the environment will be sufficiently visible to the recent immigrant or not, or whether any parts of the environment are more attended by the user or not. In this figure, parts with specific textures, such as brands, are more frequently fixated than other parts (b).

4.2 CONCLUSION – MASELTOV SERVICE ASPECTS

The use of smartphones and the accessibility of mobile services - in the field, doing the task - are currently mostly ignored. In order to understand the behaviour of immigrants, using the smartphone, in the field, one has to investigate the human factors, underlying and specifying the interaction, in best detail. As argued before, this should be performed using the analysis of the attention process since human attention is distributed on parts of the environment that indicate a context with the interest, with the task context, of the user.

Mobile location based services are per se in principle designed to have a relation to the immediate environment, either by a typical “nearby point of interest” service, or even by relating to details of the environment as is the case with the text lens service component. The text lens grabs pictures of the environment and automatically detects text parts in it, and is then capable to automatically translate that text into a target language as selected by the user. How the user selects and grabs text from the environment is still an open issue, in particular, the open research question is what is of maximum interest to the immigrant in a certain context, and whether that could be different to a local user. These differences would be pivotal in order to understand the perception, the behaviour and finally, the concrete assistance of immigrants in their daily life.

The 3D gaze recovery in arbitrary environments can therefore be a strong support to understand the interaction of immigrant mobile users not only with the mobile device but also with respect to its immediate environment. This will impact the work flow, the user interface and also potential updates in terms of new service functionality of MASELTOV components.



(a)



(b)

Figure 11 Construction of a heat map from fixation frequency data with the methodology described in Section 3. (a) Heat map visualisation. (b) Corresponding model of the environment with photorealistic textured that allows to conclude that highly texture parts of the environment such as brands and text are more often fixated than other parts.

5. APPLICATION IN USER STUDIES

5.1 USER STUDY ON MOBILE INTERACTION IN NAVIGATION APPLICATIONS

For the study of specific requirements in immigrants' mobile interaction patterns, we are planning to run a study on the use of a mobile navigation application. Specifically, 32 young women will take part in the study of which will be 16 of Turkish origin, and 16 of local origin in the city of Graz. The purpose is to determine in a sequence of 3 navigation trials for each proband the characteristics in the mobile interaction, including profound analysis of the attention process and the user trajectories towards the targeted navigation path. We intend to extract from the statistics of the interaction specific anomalies between the local and the Turkish behavior pattern that would allow us to understand a strategy that is more settled in the statistics of the environment (locals) in comparison to a less settled one in the case of the immigrants.

The study will be performed as a joint endeavor of applying human factors technologies and at the same time investigating with usability research methodology.

5.2 USER STUDY ON MOBILE DIALOGUE EVALUATION

As described in Section 3, the studies have already been performed on the basis of the GRAS² database. See Appendix B for more details on the study.

5.3 USER STUDY ON INTERACTION WITH ENVIRONMENTS

The plan is to perform user studies in the context of the “Text Lens” application which is directed towards the environment and has to make sense about the view of the user being attracted towards the display but at the same time also towards the stimuli (text occurrence) in the environment.

6. CONCLUSIONS

The deliverable sketches an overview on the innovative technology developments performed in work package WP04, Task 4.2. Three major technologies have been presented that can be used to shape future user study measurements in the frame of cultural diversity and mobile interaction analytics.

Central to this task is the development of the Smartphone Eye Tracking Toolbox (SMET) that enables to perform large user studies without manual interaction, in a fully automated manner. The work will be presented at the CHI 2014 (Paletta et al., 2014) and at the Symposium for Eye Tracking Research and Applications, ETRA 2014 (Paletta et al., 2014).

At least one larger user study will be performed in the last project year, i.e., the navigation study with 32 users from Austria and abroad, i.e., of Turkish origin.

7. 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 Provincial Government of Styria (NeoAttrakt). We kindly thank INTERSPAR Graz and CITYPARK GmbH for the permission to capture the data.

8. REFERENCES

- B. Rammstedt and O. John. Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German. *Journal of Research in Personality*, 41:203-212, 2007.
- Barnard, L., Yi, J.S., Jacko, J.A., & Sears, A. (2007). Capturing the effects of context on human performance in mobile computing systems. *PUC*, 11(46), pp. 81-96.
- Cauchard, J.R., Löchtefeld, M., Irani, P., Schoening, J., Krüger, A., Fraser, M., & Subramanian, S. (2011), Visual separation in mobile multi-display environments, *Proc. UIST 2011*.
- Chynał, P., Szymański, J. M., Sobiecki, J. (2012) Using eyetracking in a mobile applications usability testing, *Proc. 4th ACIIDS 2012*, Taiwan, Part III, J.-S. Pan, S.-M. Chen, N. T. Nguyen, Eds., 2012, vol. 7198, pp. 178-186.
- D. Miyauchi, A. Sakurai, A. Nakamura, and Y. Kuno. Active eye contact for human-robot communication. In *Proc. CHI 2004*, pages 1099-1102. ACM, 2004.
- Duda, R.O. & Hart, P.E. (1972). Use of the Hough Transformation to Detect Lines and Curves in Pictures, *Communications of Assoc. for Computing Machinery*, 15(1):11-15, 1972.
- Fritz, G. And Paletta, L. (2010), Semantic Analysis of Human Visual Attention in Mobile Eye Tracking Applications, *Proc. ICIP 2010*, pp. 4565 - 4568.
- Gehring, S., Daiber, F., & Lander, C. (2012). Towards Universal, Direct Remote Interaction with Distant Public Displays, *Proc. 3rd Workshop on Infrastructure and Design Challenges of Coupled Display Visual Interfaces*, 2012.
- Giannopoulos, Ioannis, Peter Kiefer, And Martin Raubal (2012), GeoGazemarks: Providing gaze history for the orientation on small display maps, *Proc. ICMI 2012*.
- Guo, Z. & Hall, R.W. (1989). Parallel Thinning with Two-Subiteration Algorithm, *Communic. ACM*, 32(3) 1989.
- Jacob, R.J.K. & Karn, K.S. (2003). Eye Tracking in Human-Computer , Interaction and Usability Research: Ready to Deliver the Promises, in *The Mind's Eye: Cognitive and Applied Aspects of Eye Movement Research*.
- Kluge, M. & Asche, H. (2012). Validating a smartphone-based pedestrian navigation system prototype - An informal eye-tracking pilot test. *Proc. ICCSA 2012*, Springer, pp. 386-396.
- Mardanbegi, D. & Hansen, D.W. (2011), Mobile gaze-based screen interaction in 3D environments, *NGCA 2011*.
- Miluzzo, M., Wang, T., & Campbell, A.T. (2010). EyePhone: Activating Mobile Phones With Your Eyes; *Proc. MobiHeld*, pp. 15-20.
- Nagata, S (2003) Multitasking and interruptions during mobile web tasks. *Proc. Human Factors and Ergonomics Society*, 47th annual meeting, pp. 1341-1346.
- Neerincx, M.A., Streefkerk, J.W. (2003). Interacting in desktop and mobile context: emotion, trust and task performance. *Proc. EUSAI*, Eindhoven, Netherlands.

- Paletta, L., Almer, A., Thallinger, G., Mayer, H., Pirker, K., Bischof, H., Behmel, A., Scheitz, W., Schrammel, J., Tscheligi, M. (2012). Semantic Mapping of Embodied Attention, *Proc. 5th International Conference on Cognitive Systems, COGSYS 2012*, Vienna, 22-23 February, 2012.
- Paletta, L., Santner, K., Fritz, G., and Heinz Mayer (2013). 6 DOF Reconstruction of Human Gaze in Uncalibrated Environments, *Proc. Intelligent Robots and Computer Vision XXX: Algorithms and Techniques*, Conference EI118, SPIE Electronic Imaging, San Francisco, January, 2013.
- Paletta, L., Neuschmied, H., Schwarz, M., Lodron, H., Pszeida, M., Luley, P., Ladstätter, S., Deutsch, S., Bobeth, J., Tscheligi, M. (2014). Attention in Mobile Interactions: Gaze Recovery for Large Scale Studies, *Proc. ACM SIGCHI Conference on Human Factors in Computing Systems, CHI'14 extended abstracts*, Toronto, Canada.
- Paletta, L., Neuschmied, H., Schwarz, M., Lodron, H., Pszeida, M., Ladstätter, S., Luley, P. (2014), Smartphone Eye Tracking Toolbox: Accurate Gaze Recovery on Mobile Displays, *Proc. Symposium on Eye Tracking Research and Applications, ETRA 2014*, Safety Harbor, Florida.
- Rohs, M., Schöning, J., Raubal, M., Essl, G., & Krüger, A. (2007). Map Navigation with Mobile Devices: Virtual versus Physical Movement with and without Visual Context, *Proc. ICMI 2007*, pp. 146-153.
- Santner, K., Fritz, G., Paletta, L., and Heinz Mayer (2013). Recovery of Saliency Maps from Human Attention in 3D environments. *Proc. IEEE International Conference on Robotics and Automation*, Karlsruhe, Germany, May, 2013.
- Schrammel, J., Mattheiss, E., Döbelt, S., Paletta, L., Almer, A., Tscheligi, M. (2011), Attentional Behavior of Users on the Move Towards Pervasive Advertising Elements, in eds., Müller, Alt, Michelis, *Pervasive Advertising*, Springer.
- Yeoh, P.Y. & Abu-Bajar, S. A. R. (2003), Accurate real-time object tracking with linear prediction method, *Proc. ICIP 2003*, Vol. 3, pp. 941-944.

Appendix A

Paletta, L., Neuschmied, H., Schwarz, M., Lodron, H., Pszeida, M., Luley, P., Ladstätter, S., Deutsch, S., Bobeth, J., Tscheligi, M. (2014). Attention in Mobile Interactions: Gaze Recovery for Large Scale Studies, *Proc. ACM SIGCHI Conference on Human Factors in Computing Systems*, CHI'14 extended abstracts, Toronto, Canada

Attention in Mobile Interactions: Gaze Recovery for Large Scale Studies

Lucas Paletta

DIGITAL – Inst. Information & Comm. Techn.
JOANNEUM RESEARCH FgesmbH, Austria
lucas.paletta@joanneum.at

Helmut Neuschmied

DIGITAL – Inst. Information & Comm. Techn.
JOANNEUM RESEARCH FgesmbH, Austria
helmut.neuschmied@joanneum.at

Michael Schwarz

DIGITAL – Inst. Information & Comm. Techn.
JOANNEUM RESEARCH FgesmbH, Austria
michael.schwarz@joanneum.at

Gerald Lodron

DIGITAL – Inst. Information & Comm. Techn.
JOANNEUM RESEARCH FgesmbH, Austria
gerald.lodron@joanneum.at

Martin Pszeida

DIGITAL – Inst. Information & Comm. Techn.
JOANNEUM RESEARCH FgesmbH, Austria
martin.pszeida@joanneum.at

Patrick Luley

DIGITAL – Inst. Information & Comm. Techn.
JOANNEUM RESEARCH FgesmbH, Austria
patrick.luley@joanneum.at

Stefan Ladstätter

DIGITAL – Inst. Information & Comm. Techn.
JOANNEUM RESEARCH FgesmbH, Austria
stefan.ladstätter@joanneum.at

Stephanie Deutsch

CURE – Center for Usability Research & Eng.
Modecenterstr. 17, Obj. 2, Vienna, Austria
deutsch@cure.at

Jan Bobeth

CURE – Center for Usability Research & Eng.
Modecenterstr. 17, Obj. 2, Vienna, Austria
bobeth@cure.at

Manfred Tscheligi

ICT&S Center, University of Salzburg;
AIT Austrian Institute of Technology GmbH
manfred.tscheligi@ait.ac.at

Abstract

Understanding human attention in mobile interaction is a relevant part of human computer interaction, indicating focus of task, emotion and communication. Lack of large scale studies enabling statistically significant results is due to high costs of manual penetration in eye tracking analysis. With high quality wearable cameras for eye-tracking and Google glasses, video analysis for visual attention analysis will become ubiquitous for automated large scale annotation. We describe for the first time precise gaze estimation on mobile displays and surrounding, its performance and without markers. We demonstrate accurate POR (point of regard) recovery on the mobile device and enable heat mapping of visual tasks. In a benchmark test we achieve a mean accuracy in the POR localization on the display by ≈ 1.5 mm, and the method is very robust to illumination changes. We conclude from these results that this system may open new avenues in eye tracking research for behavior analysis in mobile applications.

Author Keywords

Human attention; gaze recovery; mobile interaction heat maps.

ACM Classification Keywords

H.1.2 User/Machine Systems - Human information processing.

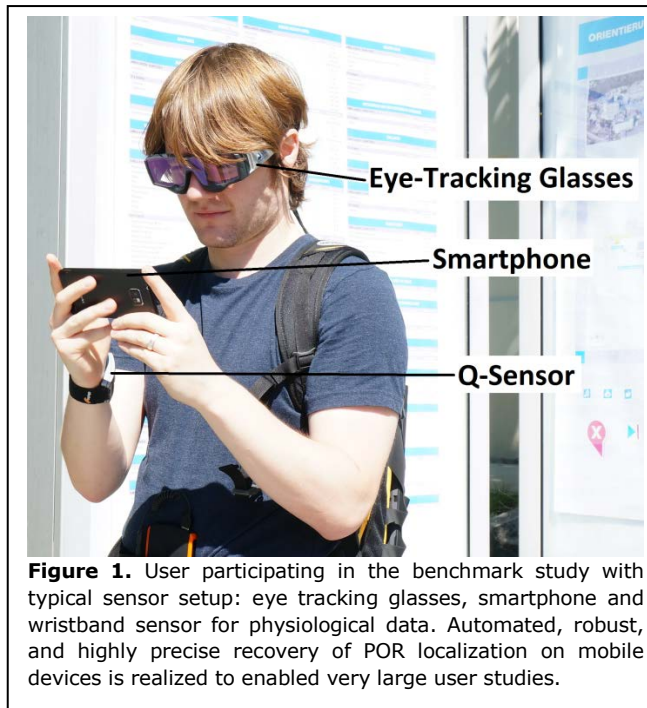
Copyright is held by the author/owner(s).

CHI 2014, Apr 26 - May 01 2014, Toronto, ON, Canada

ACM 978-1-4503-2474-8/14/04.

Introduction

The evaluation of interaction designs in mobile applications requires in general the investigation of human attention by means of eye movements, in analogy to usability analysis for the iterative development of desktop technologies [9], such as for websites.



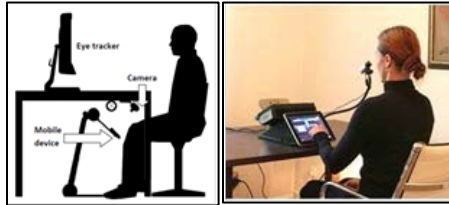
interruption are of high relevance, since in a mobile context the frequency of distracting events is much higher than for a desktop application and tasks with interruptions take longer to complete on a mobile device than with a desktop application [4]. Special interest is therefore on increased competition with regard to attracting the users' attention and on

interaction as a non-primary task in a certain context [5]. However, current technologies for the mapping of point-of-regards (PORs) to mobile displays do not enable natural interaction with the mobile device: users of the mobile device are either conditioned to act with tightly mounted displays and hence prevented from performing in the typical mobile user's environment or they are distracted by markers in the view (Figure 3).

We propose a novel and successful approach that enables natural interaction with the mobile device and its application (Figure 1). We use eye tracking glasses (ETG) to capture the user's fixations on its environment and apply computer vision for highly accurate recovery of human gaze on the mobile display. In a benchmark study we achieve a mean accuracy of POR localization on the display of $\approx 1.5 \pm 0.9$ mm. Furthermore, we demonstrate that the approach enables the analysis of large user studies in contrast to using state-of-the-art annotation tools. In contrast to state-of-the-art methodologies, our approach enables natural interactions, from these one can draw more valid conclusions and better interaction designs, in the frame of usability research in both outdoors and indoors studies. We conclude that this will open new avenues for understanding mobile interaction.

Related Work

Eye tracking devices are frequently used to analyze mobile interaction. In indoor studies, mapping of PORs to mobile displays is usually performed by measuring with tightly mounted phones (Figure 3a,b). For example, [6] screw the mobile device under a table, and enable evaluation of mobile interaction only when sitting (Figure 3a). [7] investigated and proved the significance of eye tracking in mobile applications



(a) Tobii (b) Mindfacts GmbH



(c) Annotation [10] (d) Dikablis GmbH

Figure 3 State-of-the-art in usability oriented eye tracking research on mobile interaction. (a) Tobii setup of phone mounted underneath the table. (b) Mindfacts setup on top of the table. (c) Manual Annotation with BeGaze (SMI). (d) Automated annotation with markers.

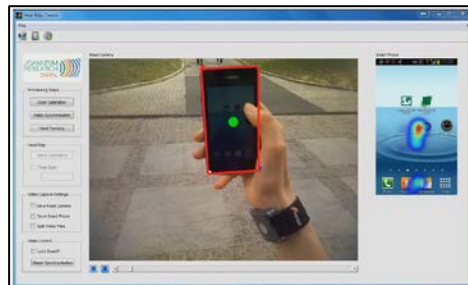


Figure 4 Graphical User Interface of the proposed Smartphone eye tracking (SMET) system application. ETG video (center) and display cast (right) with actual heat map are

usability testing, using a head mounted eye tracker. [8] developed an eye tracking pilot test on validating a smartphone based pedestrian navigation system, describing the relationship between reality and navigation instructions, using a standard annotation toolbox. [10,16] presented static display localization in eye tracking tasks. However, their approach is specific, localization of mobile devices involves more degrees of freedom. [11] referred to the concept of mobile device detection but lacked quantitative information on experimental results. [12] mentioned Dikablis based marker tracking (Figure 3d) which principally distracts the user's attention.

Alternative approaches use the smartphone camera directly with its rear view on the user to estimate the eye gaze on the mobile phone [9]. This method achieves accuracies only in centimeters range, is vulnerable to illumination conditions, and cannot provide information about the surrounding that contains relevant information [15].

Smartphone Eye Tracking Toolbox

We developed a software toolbox package, i.e., the 'smartphone eye tracking toolbox' (SMET) that will assist usability analysis towards fully automated analysis of human attention processes in user studies. Figure 2 depicts a schematic sketch of the

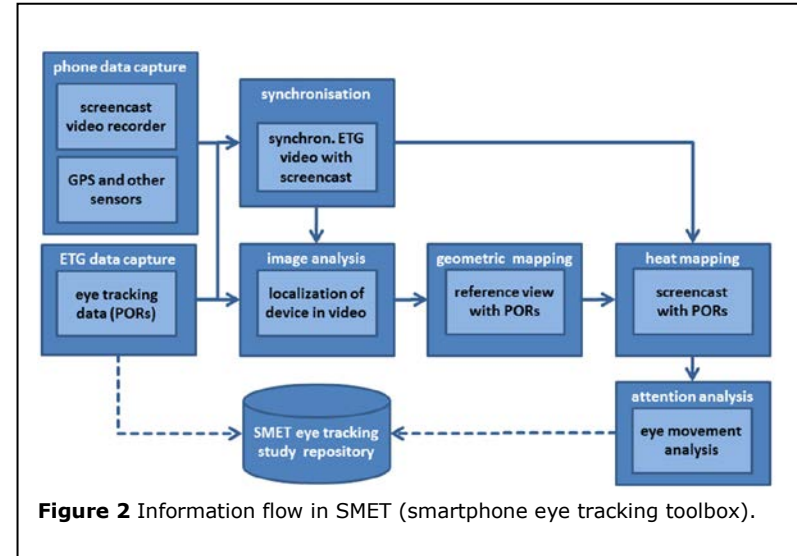


Figure 2 Information flow in SMET (smartphone eye tracking toolbox).

information flow in the toolbox: data capture is applied with the ETG and the screencast recorder; synchronization and image analysis provide a data stream with synchronized POR and smartphone display events. Finally, geometric transformation and heat mapping provide the basis for attention analysis.

The Graphical User Interface (GUI) of the SMET application (Figure 4) requires three kinds of input data: the scene camera based video, the smartphone video, and the captured POR data. Firstly, points that lie on image areas of the colored phone case are manually selected for initialization, but further then the color thresholds which are required by the tracking algorithm are automatically determined. The two videos are then synchronized in a semi-automated manner and the automated smartphone tracking process is applied. Once the smartphone has been detected in the image,

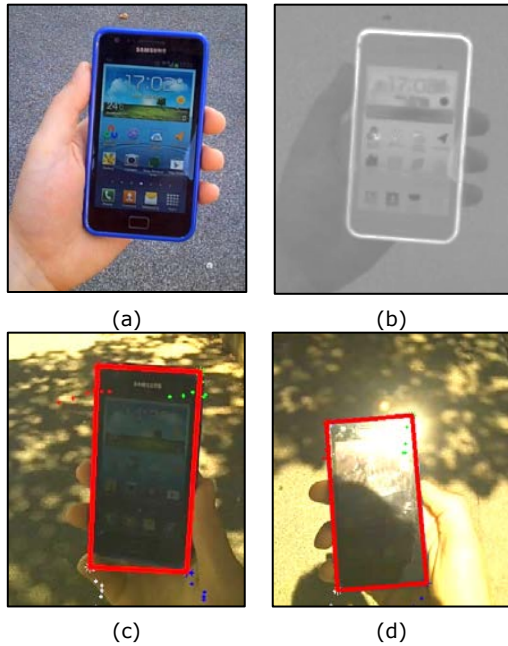


Figure 5 (a) Mobile device with color case, (b) corresponding image in YCbCr color space, (c,d) resulting detections with difficult light conditions.

benchmarks	test 1	test 2	test 3
no. of frames	4284	6531	3869
detections	1913	1318	3479
false positives (total)	4	0	0
true positives (total)	1913	1318	3479
precision [%]	99.8	100	100
error [pixel] mean	3.3±2.2	5.2±8.1	2.9±9.7
error [mm] mean	1.5±0.9	1.4±1.4	1.4±2.5

Table 1. Accuracy of POR mapping on displays.

an affine transformation is computed to assign the POR location in the scene camera image to a location in the phone image. This is used to calculate a standard Gaussian function type heat map of PORs at the smartphone user interface, with an arbitrary time window to accumulate PORs for heat mapping

To enable robust tracking the smart phone is used with a color intensive case (Figure 5a) in order to identify the device well within the scene video. Through the detection of the colored frame lines and the corner points the image segment of the smartphone can be identified and visually tracked. The tracking algorithm makes adaptive use of color threshold values, which are manually selected in a first frame, and then automatically adjusted, so that color based segmentation is performed in appropriate color spaces (HSV, YCbCr) (Figure 5b). After noise removal by morphological filter operations, the remaining image regions are reduced to skeletons of the regions by a thinning algorithm. The resulting lines are then extracted based on Hough Transform [13]; noisy lines are removed and intersection points from the remaining lines are robustly found. Corner points of the smart phone are detected, are validated by previous tracking results and geometrical

constrains of the smart phone area. The tracking algorithm is based on a linear prediction which

performs well for a random movement of a single object [14]. Positions of missing corner points are estimated based on the movement of detected points and position predictions from tracking, resulting in localizations even if hands occlude parts of the area of interest or only small parts of the phone are observed within the image. Through the continuous adaptation of the color thresholds based on actual tracking results, the detection of the smart phone is robust in challenging illumination conditions (Figure 5c,d).

Proof of Concept - Navigation Study

The study targeted a proof of concept, investigating which level of accuracy in POR mapping can be achieved. Typical users of mobile applications were involved several times in an outdoor navigation task, including map based and augmented reality (AR) guidance, to find locations in a local hospital. To acquire video and eye-tracking data we used SMI Eye Tracking Glasses, with 30 Hz sampling rate of gaze and a 1280 x 960 pixel resolution scene camera. Figure 6a,b depict the camera view from the Eye Tracking Glasses, overlaid with the bounding box of the smartphone localization; Figure 6c presents a heat map from PORs mapped onto the display of the mobile device.

Precision of Smartphone Localization is crucial for the usability of automated annotation in ETG video frames. Table 1 presents most relevant evaluation results from 3 different test runs using the mobile service, under different weather conditions (2x sunny, 1x cloudy weather). A substantial subset of the number of video frames ('detections') represents the interaction of the user with the mobile phone, however, the detection method was applied to all video frames. The precision rate was sufficiently high to enable stable, cost-efficient

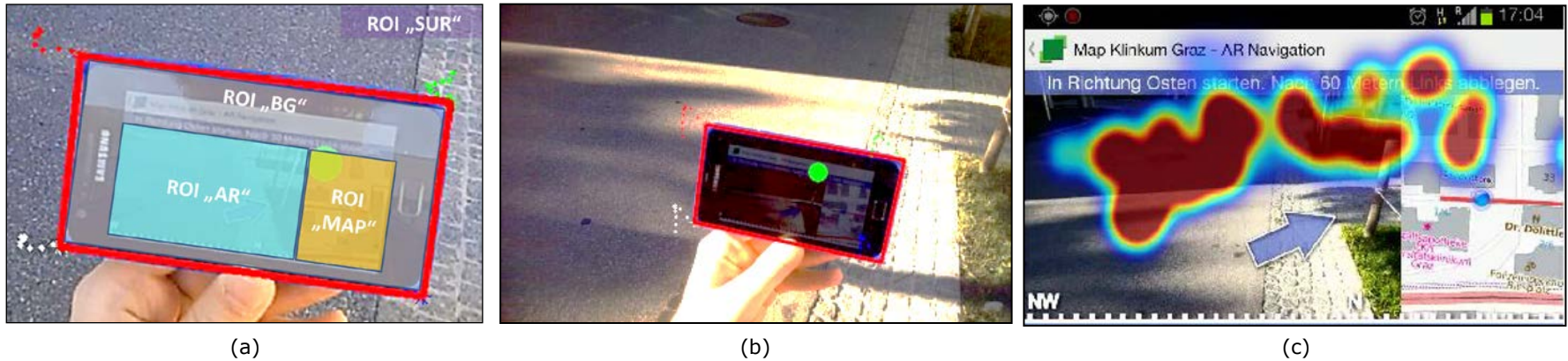


Figure 6 Attention analysis of mobile interaction from eye tracking and computer vision. (a) ROIs (regions of interest) on the display. (a,b) Automated annotation (red) and (c) heat mapping (using a time window, including fixations within past 5 sec.) on the screen cast display that is synchronized with the ETG video (see Figure 4).

processing. The calibration error was estimated $\approx 3\text{mm}$. The human error in the ground truth annotation was $\approx 2\text{mm}$. The localization error fitted nicely to be $\approx 1.5 \pm 0.9\text{ mm}$, in reference to index to icon buttons of, e.g., size $7 \times 7\text{ mm}$ in standard Samsung displays.

Semantic Annotation. We are interested to measure the distribution of attention on regions of interest (ROIs) of the smartphone display and the surrounding in a navigation application (Figure 6a): “AR” depicts the augmented reality view, i.e., the orientation arrow overlaid on the live camera view; “MAP” refers to the map view on the environment and a red line representing the planned trajectory of a user; “BG” the information panel of the smartphone display; “SUR” the surrounding that is viewed beyond the smartphone. We analyzed the ETG video as described before and determined the PORs’ recovery on the device’s display. We then associated the POR to ROIs’ segments as depicted in Figure 6a.

Performance Analysis. From the statistics of 24.000 video frames from 2 users, among those 12.884 frames with automated smartphone detections, we deduce the fact that 62% of PORs were localized within “AR”, 9% on “MAP”, 9% on “BG” and 20% on “SUR”. The avg. dwell time on “AR” has been calculated to be 471 (± 540) ms, for “MAP” 128 (± 115) ms, for “BG” 88 (± 96) ms and for “SUR” 176 (± 198) ms. All quantities were extracted and calculated in a fully automated manner, and will be input for more profound analysis.

Automated annotation of ROIs enables to perform large user studies without substantial efforts, which has not been applicable so far. In a user study with 100 persons, and a study with similar quantities as above (100 times 12884 frames to annotate), counting 3 sec./frame using standard (such as BeGaze) interactive annotation software, assuming 8 hours daily work –one operator would need 134 days (i.e., 26.8 weeks) to annotate - this in contrast to the presented automated annotation process with no intervention needs.

Discussion and Conclusion

We presented a novel and robust approach for marker free tracking of smartphones in the head video of eye tracking glasses. The localization precision is sufficient to enable analysis of button choices in the display. This enables natural interaction, especially in large user studies, using automated instead manual annotation, and carries a huge potential to be extended with computer vision methodology. Limitations are in extreme illumination, such as direct sunlight and dawn.

Potential applications are mobile interaction studies both outdoors (navigation, image/video based user interfaces, urban points-of-interest, mixed reality games, cell based interaction with ticket machines or urban displays) and indoors (multiple displays, mobile-TV interaction). Future work will focus on the consideration of attended semantics in the surrounding (detection of persons, objects, landmarks), and on the fusion with information from psychophysiological and activity sensors.

Acknowledgements

This work has been partly funded by FP7(2007-2013) (grant n°288587, MASELTOV) and by the Austrian Research Prom. Agency (grant n°832045) FACTS.

References

- [1] Jacob, R.J.K. & Karn, K.S. (2003). Eye Tracking in Human-Comp. Interaction and Usab. Research, *The Mind's Eye: Cognit. and Appl. Aspects of Eye Mov. Res.*
- [2] Barnard, L., Yi, J.S., Jacko, J.A., & Sears, A. (2007). Capturing the effects of context on human perform. in mobile comp. syst. *PUC*, 11(46), pp. 81-96.
- [3] Rohs, M., Schöning, J., Raubal, M., Essl, G., & Krüger, A. (2007). Map Navigation with Mobile Devices: Virtual versus Physical Movement with and without Visual Context, *Proc. ICMI*, pp. 146-153.
- [4] Nagata, S (2003) Multitasking and interruptions during mobile web tasks. *Proc. HFES, 47th Ann. Meet.*
- [5] Schrammel, J., Mattheiss, E., Döbelt, S., Paletta, L., Almer, A., Tscheligi, M. (2011), Attentional Behavior of Users on the Move Towards Pervasive Advertising Elements, *Pervasive Advertising*, Springer.
- [6] Cauchard, J.R., Löchtefeld, M., Irani, P., Schöning, J., Krüger, A., Fraser, M., & Subra-Manian, S. (2011), Visual separ. in mobile multi-display envir., *Proc. UIST*.
- [7] Chynał, P., Szymański, J. M., Sobiecki, J. (2012) Using eyetracking in a mobile applications usability testing, *Proc. ACIIDS*, vol. 7198, pp. 178-186.
- [8] Kluge, M. & Asche, H. (2012). Validating a smart-phone-based pedestrian navigation system prototype. *Proc. ICCSA*, pp. 386-396.
- [9] Miluzzo, M., Wang, T., & Campbell, A.T. (2010). EyePhone: Activating Mobile Phones With Your Eyes; *Proc. MobiHeld*, pp. 15-20.
- [10] Fritz, G. And Paletta, L. (2010), Semantic Analysis of Human Visual Attention in Mobile Eye Tracking Applications, *Proc. ICIP*, pp. 4565 - 4568.
- [11] Gehring, S., Daiber, F., & Lander, C. (2012). Towards Universal, Direct Remote Interaction with Distant Public Displays, *Proc. PPD*.
- [12] Giannopoulos, Ioannis, Peter Kiefer, And Mar-Tin Raubal (2012), GeoGazemarks: Providing gaze history for the orientation on small display maps, *Proc. ICMI*.
- [13] Duda, R.O. & Hart, P.E. (1972). Use of the Hough Transformation to Detect Lines and Curves in Pictures, *Comm. of Assoc. for Comp. Machinery*, 15(1):11-15.
- [14] Yeoh, P.Y. & Abu-Bajar, S. A. R. (2003). Accurate real-time object tracking with linear prediction method, *Proc. ICIP*, Vol. 3, pp. 941-944.
- [15] Paletta, L., Santner, K., Fritz, G., Mayer, H., Schrammel, J. (2013). 3D Attention: Measur. of Visual Saliency Using Eye Track. Glasses, *Proc. CHI ext. abstr.*
- [16] Mardanbegi, D. & Hansen, D.W. (2011), Mobile gaze-based screen interact. in 3D envir., *Proc. NGCA*.

Appendix B

Eyben, F., Weninger, F., Schuller, B., and Paletta, L., The acoustics of eye contact - Detecting visual attention from conversational audio cues, *Proc. 6th Workshop on Eye Gaze in Intelligent Human Machine Interaction*, (GAZE-IN 2013), held in conjunction with the ACM ICMI 2013, Sydney, Australia, 13 December, 2013

The acoustics of eye contact – Detecting visual attention from conversational audio cues.

Florian Eyben
Maschine Intelligence and
Signal Processing (MISP)
Technische Universität
München
Munich, GERMANY
eyben@tum.de

Felix Weninger
Maschine Intelligence and
Signal Processing (MISP)
Technische Universität
München
Munich, GERMANY
weninger@tum.de

Lucas Paletta
Joanneum Research
Graz, AUSTRIA
lucas.paletta@joanneum.at

Björn Schuller^{*}
Maschine Intelligence and
Signal Processing (MISP)
Technische Universität
München
Munich, GERMANY
schuller@tum.de

ABSTRACT

An important aspect in short dialogues is attention as is manifested by eye-contact between subjects. In this study we provide a first analysis whether such visual attention is evident in the acoustic properties of a speaker's voice. We thereby introduce the multi-modal GRAS² corpus, which was recorded for analyzing attention in human-to-human interactions of short daily-life interactions with strangers in public places in Graz, Austria. Recordings of four test subjects equipped with eye tracking glasses, three audio recording devices, and motion sensors are contained in the corpus. We describe how we robustly identify speech segments from the subjects and other people in an unsupervised manner from multi-channel recordings. We then discuss correlations between the acoustics of the voice in these segments and the point of visual attention of the subjects. A significant relation between the acoustic features and the distance between the point of view and the eye region of the dialogue partner is found. Further, we show that automatic classification of binary decision eye-contact vs. no eye-contact from acoustic features alone is feasible with an Unweighted Average Recall of up to 70%.

Categories and Subject Descriptors

H.4 [Information Systems Applications]: Miscellaneous;

^{*}Björn Schuller is also affiliated with Joanneum Research, Graz, Austria

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

ICMI 2013 Sydney, Australia

Copyright 20XX ACM X-XXXXX-XX-X/XX/XX ...\$15.00.

D.2.8 [Software Engineering]: Metrics—*complexity measures, performance measures*

General Terms

Eye-gaze, attention, visual, acoustic

Keywords

Eye-gaze, attention, visual, acoustic

1. INTRODUCTION

An important aspect in short person to person dialogues is attention as is manifested by eye-contact between subjects. Thus, to replicate human-like behavior for artificial systems (e.g., humanoid robots or virtual agents) it is believed to be highly important to implement natural patterns of eye contact [4]. Furthermore, consistency between eye contact and acoustic cues emitted by the system, e.g., by means of speech synthesis, should be ensured. In this study we verify the correlation between visual attention and acoustic cues of a speaker's voice in human to human dialogues. Such information could be used in low-resource, or speech only systems which do not have a camera or eye-tracking device available, e.g., in Voice conversations and chats the other partner could be informed about the eye-gaze behaviour of the first partner without actually seeing him/her. Also in forensic analysis these methods could be applied.

In this paper, we introduce and use the GRAS² database. The database contains multi-sensor recordings of subjects engaged in a real-life short dialogue (typically one short question and a short answer). The paper is structured as follows: first, we introduce and describe the GRAS² database in Section 2, and the automatic identification of speech segments in the continuous recordings and labelling of them as speech from the test subject (referred to as subject in the ongoing) and speech from dialogue partners (referred to as partner in the ongoing) in Section 4. Next, we present an analysis of the correlations of acoustic features with the eye

contact between subject and partner in Section 5 as well as results of experiments where we try to predict whether the subject is looking at the partner's eyes/face or somewhere else just from the acoustics of his/her voice in Section 6. We summarise our findings in Section 7.

2. THE GRAS² DATABASE

The Graz Real-Life Affect in the Street & Supermarket (GRAS²) corpus is – to the authors' best knowledge – the first database of visual attention recordings with multiple audiovisual, physiological, and movement sensory cues in real-life conversations. Four subjects took part in the recordings (3 female, 1 male, cf. Figure 1).

These were all native Austrian students and they filled a BFI-11 personality questionnaire [7]. The male subject usually wears glasses, the female subjects did not wear glasses. All four subjects were equipped with SMI Eye Tracking Glasses able to record both the eyes of the person wearing these and what the person is looking at (static frontal camera, not affected by eye tracker result). They allow for precise measurement of visual attention focus (30 Hz binocular with automatic parallax compensation; pupil/CR by dark pupil tracking, spatial resolution 0.1°, gaze position accuracy 0.5° over all distances from 40 cm to infinity with a gaze tracking range of 80° horizontal and 60° vertical) in the simultaneously recorded field of vision (recorded in HD 1280 x 960 pixels at 24 fps compressed with the H.264 codec; viewing field of 60° horizontal and 46° vertical). They also feature a monophonic microphone on the left earpiece of the glasses that records in 16 kHz, 16 bit.

The recording of the data from these glasses was carried out on an SMI-ETG laptop (1.3 kg) worn in a backpack. The USB-Cable connection was hidden under hair and clothing as much as possible. Further, subjects were equipped with the Affectiva Q Sensor 2.0, a wearable sensor that measures Electro-Dermal-Activity (EDA) and skin temperature [6] to capture indication on arousal during attention. It was worn on the left hand to resemble a watch in appearance.

To record additional audio data without particularly demanding hardware conditions, an Android smart phone Samsung Galaxy Nexus was used similar as in [8]: The phone was loosely located in a front-pocket of a shirt worn by the subjects. The standard media recording APIs of Android use an AMR codec with poor quality. Therefore, the audio stream was accessed directly and saved uncompressed at 44.1 kHz, 16 bit. The recording component was implemented as a service and thus could run in the background with the phone locked and the screen off. In addition to this, limited motion sensing on the phone is available through an InvenSense MPU-3050 accelerometer unit. This sensor contains a MEMS accelerometer and a gyroscope. Linear and angular accelerations can be captured at a sampling rate of up to 100 Hz. Since the audio recording already puts the processor under considerable load, the actual achievable sampling rate was 10–20 Hz. The MARIA application [5] for public transport guidance was adjusted in a way to log the audio alongside acceleration data from the motion sensor. In addition, a Zoom H2 four-channel recording device recording at 48 kHz, 16 bit was worn on the for high quality audio. Finally, a secondary accelerometer sensor was worn in the backpack: It was contained in the NAVIN Mini Homer GPS tracker that was operated at a sample rate of 0.2 Hz. With this setup, 2-way video, 6 channel audio, EDA, tempera-

ture, and twice 3D motion is measured from the subjects. However, in the ongoing only video and audio from the eye tracker glasses and the smartphone will be used.

The subjects were accompanied by a supervisor (27 years, male) who helped with the setup and monitored the progress from a distance. The equipment setup and 3-point calibration was carried out on a parking lot of the Citypark in Graz/Austria. Three-times hand-clapping looking at the hands is used as anchor point for synchronization between those units that are not directly connected. The subjects had to search three stores as a first “warm-up” task (Le Clou Jewelry, Oxyd fashion store, and the Golden Sun Solarium) to familiarize with the worn equipment that was hidden as much as possible (cf. Figure 1).

The recordings of interest then took place in the Inter-SPAR supermarket, where the subjects engaged in dyadic discourse exclusively with female persons shopping in this supermarket. These are referred to as (dialogue) “partners” in the ongoing as opposed to the knowingly involved and equipped four “subjects”. These had no knowledge at first that they were part of the study – 28 persons agreed to provide their recorded audiovisual-footage for scientific purpose (cf. examples in Figure 1, two bottom rows) – data of subjects not agreeing was deleted by the recording subject. The limitation to female subjects was decided upon to reduce gender effects. Further, permission from the site-holders was given to carry out the recordings and use the material.

The subjects followed a study protocol as follows to engage in discussion in German language (Graz-region Styrian dialect) with subjects: They needed to search for *Sauerkraut* and a Swiss chocolate drink (*Ovomaltine*, or US: *Ovaltine*), ask for a SPAR chocolate, a specific Calculator available in the supermarket, a “typical Austrian product”, Turkish Ayran, denture adhesive for third teeth, and anti-athlete's foot cream. Thereby, they stuck with one dialogue partner as long as he/she was willing to help. Subsequently, they immediately asked for written consent explaining the experiment which was also recorded and usually consumed the larger partition of the time. This included a questionnaire on the demographics of the dialogue partners.

The choice of items to ask for and the sequential asking for continued help as well as the surprising revealing of them being recorded in an experiment are intended to elicit a range of affect including besides neutral also joyful, uncertain, surprised, confused, and negative emotional behavior in diverse real-life blend. This was further benefited by the condition that the subjects addressed their dialogue partners with the second personal pronoun “Du” (you) as usually used with friends and familiar persons as opposed to the formal and polite German “Sie” (also translates as you in English—but usually equivalent to addressing a person with the last name only). In the questionnaire, five dialogue partners would usually prefer the casual form “Du”, four the formal “Sie”, two would not have cared, and 17 made no statement. The age range of the dialogue partners that agreed is as follows (in years): 18–25 (3x), 26–35 (2x), 36–45 (4x), 46–55 (6x), 56–65 (4x), and no mention (9x). The four subjects which carried out the recordings are referred to as subjects A, B, C, and D in the ongoing, where A is the male subject and the other three are female subjects.

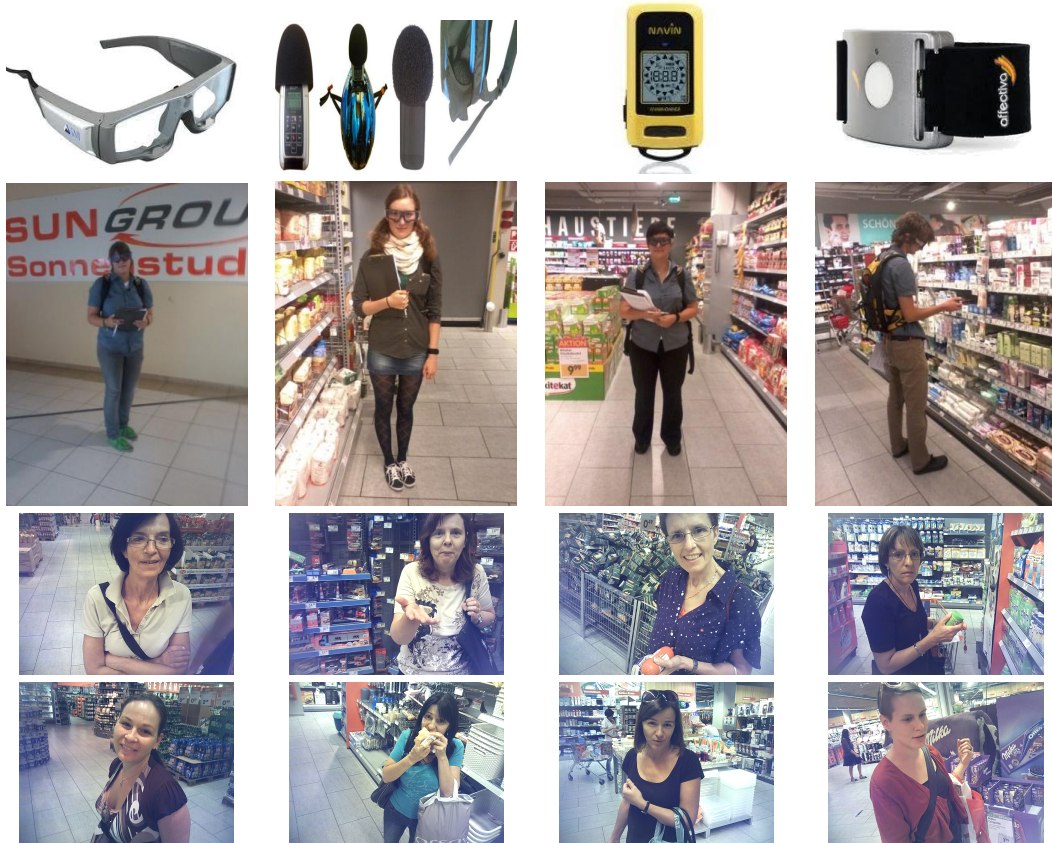


Figure 1: Top-most row: Recording equipment worn by the subjects (from left to right: eye tracking glasses, audio recorder, GPS tracker (used for extra accelerometer in backpack), EDA sensor (Affectiva) – details in the text). Second row: The four participating subjects as equipped on site. Bottom two rows: Examples of recorded dialogue partners as seen by the four subjects through their worn eye tracking glasses.

3. AUDIO TRACK ALIGNMENT

The audio tracks from the eye tracker and the smartphone (and also the Zoom H2 recorder – however, this is not used here) needed to be aligned for three reasons: a) the start time of the recordings differs, as recording on the devices was started sequentially by hand, b) the sampling clocks of the devices were not synchronized and thus drifted significantly over the course of a one hour recording, and c) the recording app on the mobile phone occasionally dropped audio frames at random locations – presumably due to high system load – of more than 1 second.

To be able to process as much audio data as possible and ideally have all audio tracks aligned as perfectly as possible at every time instant, we used an automatic alignment algorithm. This algorithm is capable of aligning the audio tracks completely unsupervised. We consider two tracks, where one is referred to as the master track, and the other as slave track. The goal is to align the slave track to the master track. The master track is not modified in any way. The algorithm consists of three steps:

1. Finding the initial displacement of the tracks at the beginning of the recording
2. Finding frame drops and sudden misalignments within the recording

3. Estimating sample-wise displacements (compensating drift of sampling clocks).

Once the displacement values for each sample are known, spline interpolation is applied on a sample level to align the slave track to the master track. The initial displacement is estimated via a window based cross-correlation search, which finds the position of the first 8 seconds of master audio in the slave signal. The slave audio before this position is truncated before the other two steps are executed. In step (2) a large sliding window (8 s) is used for cross-correlations.

The windows of the master signal are sampled at a constant rate of 8 s, while the corresponding windows in the slave signal are dynamically shifted by the current displacement, which is initially 0 for the first window, and for the second window equal to the displacement found by cross-correlation of the first window of master and slave, etc. Due to the large window, discontinuities caused by frame drops of up to 2 seconds in both directions can be robustly detected, which is sufficient for the GRAS² corpus. As an example, the result of the displacement analysis between the eye tracker's audio (master) and the Phone's audio (slave) is shown in Figure 2 for subject A. In step (3) the locations of the frame drops causing jumps in the track displacement function (Figure 2) are estimated with a better temporal resolution by a smaller search window (0.25 s). Next, the

Subject	A (m)	B (f)	C (f)	D (f)
Recording duration [min]	85	61	81	67
# Subject speech segments	611	480	566	329
Subject speech segments duration [min]	20	14	20	13
% segments with face present	91.2	92.5	95.1	95.1
Duration face detected [min]	9.5	7.1	11.1	7.9
% segs. w. eye – eye view (V_c)	10.3	5.6	3.0	13.7
% segs. w. eye – face view (V_{b+c})	47.5	27.3	24.9	37.1
% segs. w. eye – near face (V_{a+b+c})	64.0	56.9	48.6	58.4
Per turn mean length of case V_c [s]	.39	.11	.11	.16
Per turn mean length of case V_b [s]	.57	.21	.20	.48
Per turn mean length of case V_a [s]	.30	.23	.24	.34
Mean speech segment length [s]	2.0	1.8	2.1	2.3
Max speech segment length [s]	15.6	12.5	14.5	19.0
Std. dev. of speech segment length [s]	1.8	1.7	1.9	2.4
Energy difference (M-S) bias	.006	-.012	.007	.022

Table 1: Data statistics for the four subjects A-D (1 male (m), 3 female (f)). Last row gives energy difference threshold between two recording microphones which was used for the automatic segmentation of subject utterances (see text on automatic segmentation for details).

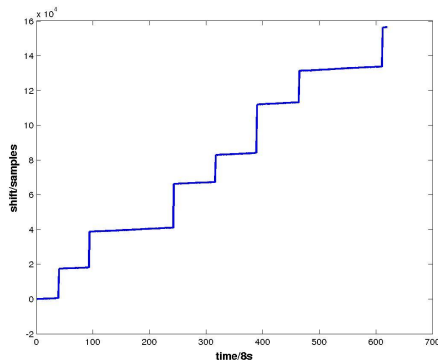


Figure 2: Displacement between eye tracker and Phone audio signals. The amount of samples by which the Phone audio signal needs to be shifted to match the eye tracker audio signal is shown on the y -axis. The x -axis shows the time in units of 8 s windows.

accuracy of the small drift occurring by the sample clock de-synchronization is refined with the same 0.25 s search windows in regions where no jump occurs. The lag of the cross-correlation thereby is constrained by the upper and lower bounds estimated for the previous and the following frame in step (2).

4. AUTOMATED SEGMENTATION

As can be seen in Table 1, the time the subject talks is much less than the total recording time. Therefore annotations are needed for speech and non-speech segments as well as whether speech comes from the subject (wearing the eye tracking glasses) or the partner or some other person close by. To be able to annotate large amounts of data in a short amount of time, we used an automated annotation method: To robustly detect speech segments in a high level of background noise (supermarket) we used our highly accurate Long Short-Term Memory Recurrent Neu-

ral Network (LSTM-RNN) multi-condition Voice Activity Detector (VAD), pre-trained as described in [1]. The background noise contains babble from other people in the store, announcements, background music, children playing, and shopping carts moving around. For increased robustness, we apply the VAD to both, the phone and eye tracker audio track.

To detect whether the subject or someone else is speaking, we rely on the relative energy differences in voice segments between the phone and the eye tracker audio tracks. For this, both audio tracks were normalized to 0dB peak amplitude before frame-wise (25ms, sampled every 10ms) root-quadratic energy was computed. In the cases where the subject is speaking, the energy level in a 500ms sliding window is generally larger for the eye tracker recordings than for the phone recordings. A small adjustment of a bias of the level differences was needed independently for each subject. These biases were found by empirically looking at the number of detected segments and the balance between subject and other segments for energy level biases in the range from -0.05 to +0.05. The threshold yielding the maximum number of segments and at the same time yielding a higher number of subject segments than partner segments was used (cf. Table 1, last row). As the segmentation is automatic and unsupervised, there are errors in the detected segments. However, a manual inspection of a subset of the detected segments confirmed that the automatic segmentation has a high accuracy and the segments can be used in further experiments.

5. EYE CONTACT AND ACOUSTICS

From the eye tracking glasses we can extract the position where the subject is looking at in the coordinate system of the eye tracker frontal camera. From the video of the frontal camera we detect the presence of a face (frontal view) with the openCV face detector based on Local Binary Pattern (LBP) features and try to estimate the eye region within the face with a Haar-wavelet based eye detector also available in openCV. If no eye region was detected in the face (e.g., if people wear glasses), we estimate the eye region from the

face region as:

$$Xe = xf + 0.25wf \quad (1)$$

$$Yh = yf + 0.25hf \quad (2)$$

$$We = 0.5wf \quad (3)$$

$$He = 0.16hf, \quad (4)$$

where the subscript e indicates the eye region bounding box and the subscript f the face region bounding box. X , y , w , and h are the coordinates of the upper left corner, the width, and the height of the bounding box, respectively.

By combining the eye tracker coordinates with the detected face and eye region, we can define three classes for where the subject is looking with respect to the partner: Direct eye contact – i.e., looking into the eye region (V_c), looking into the face region (V_b), or looking next to the face region in a corridor with 0.5 width/height to the left, top, right, bottom of the face region (V_a). Additionally, we compute the Euclidean distance between the center of the detected eye region and the point the subject is looking at. This is referred to as eye-eye distance in the following. If no face is detected in the image, a maximum value is filled in for this distance.

To produce an eye-contact ground truth per speech segment, we apply the following rule in this particular order: If for at least 2 frames there is direct eye contact (V_c), we assign the V_c label to the whole segment. Otherwise, if for at least 2 frames there is case V_b , we assign label V_b , and otherwise the same for V_a . If neither case is present in the segment we assign the label V_n for no eye contact. Detailed statistics on the amount of eye contact in the segments where the subject is talking are found in Table 1. There are notable differences between the subjects in terms of eye contact behavior. Subject A apparently has the most eye contact with his partners, durations of cases V_a and V_b are almost 1 second on average for a two second average segment duration, while for subjects B and C it is only .3 seconds and for D .7 seconds.

6. EXPERIMENTS AND RESULTS

In order to explore correlates between the acoustic and vocal properties of speech with the location of where the subject is looking at in a conversation with another person, we present an analysis of the correlations between acoustic features and the eye-eye distance. The audio recorded via the eye tracker microphone is used for this purpose. As acoustic feature set, we use a large standard set of acoustic features, as used for the baseline results in the Interspeech 2013 ComParE Challenge [9].

The features were extracted with our open-source feature extraction and affect recognition toolkit openSMILE [2]. The ComParE feature set contains 6373 features, which are functionals of acoustic low-level descriptors (LLDs). The LLDs include prosodic features (signal energy, perceptual loudness, fundamental frequency), voice quality features (jitter and shimmer of the fundamental frequency, voicing probability, and the harmonics-to-noise ratio), spectral features (spectrum statistics such as variance and entropy and energies in relevant frequency bands), and cepstral features (Mel-Frequency cepstral coefficients – MFCC). From these LLDs, the first order delta coefficients are computed and both LLDs and delta coefficients are smoothed with a 3 tap moving average filter over time. Then, functionals are applied to the

LLDs and their delta coefficients over a complete speech segment resulting in one final 6373 dimensional feature vector for the particular segment. The functionals include statistical measures such as moments (means, variances, etc.), statistics of peaks (mean amplitude of peaks, mean distance between peaks, etc.), distribution statistics such as percentiles (esp. quartiles and inter-quartile ranges), regression coefficients obtained by approximating the LLD over time as linear or quadratic function and the errors between the approximation and the actual LLD, temporal characteristics such as positions of maxima and the percentage of values above a certain threshold, and modulation characteristics expressed as linear predictor (autoregressive) coefficients of a predictor of five frames length.

In this study, rather than simply computing the Pearson correlation coefficients (CC) across all subjects and taking those with the highest absolute correlation, we use a selection criterion that rewards consistent correlation across subjects and penalizes inconsistencies such as a feature being correlated for one person yet inversely correlated for another. This leads to the following criterion for feature f :

$$CC'_f = \frac{\sum_{s=1}^S \sum_{t=s+1}^S (|CC_f^{(s)} + CC_f^{(t)}| - |CC_f^{(s)} - CC_f^{(t)}|)}{S(S-1)} \quad (5)$$

where S is the number of subjects. This criterion ranges from -1 to +1, with -1 indicating strong inconsistency, zero indicating low correlation or medium inconsistency, and one indicating perfect and consistent correlation across all subjects. One of the best correlated acoustic features from the ComParE set ($CC' = 0.21$; max. $CC = 0.37$ for subject B) is the gain of the linear prediction on the voicing probability. In speech analysis this gain resembles the energy of the “predictable” (i.e., correlated and generated by the human vocal tract) signal parts. As we are applying linear predictive coding to the contour of the voicing probability, the gain has a different meaning. The gain thus resembles the energy of predictable modulation of the voicing probability and is therefore related to speech rhythm caused by the sequence of voiced and unvoiced phonemes. The more regular the rhythm, the higher the gain is. The best negatively correlated feature ($CC' = -0.22$) is the range of the peak amplitudes relative to the arithmetic mean for the 6-th critical band (approx. 500–620 Hz) of an auditory filter bank after applying a RASTA-style band-pass filter to emphasize speech-rate modulations in the range from 4–8 Hz. This frequency range corresponds to a frequency relevant for the first formant of vowels. Thus, if the range of peaks in this frequency band is high, there is a high variation of articulatory strength of individual vowels, which corresponds to a sloppy style of articulation, or might resemble general level variations due to quickly changing acoustic conditions (e.g., a person moving relative to the microphones). Altogether the results indicate that modulation descriptors (functionals) are the most relevant. This might indicate that if we have eye contact with a person, we articulate clearer and with a different rhythm than if we do not have eye contact.

Let us now turn to the feasibility of fully automatic attention recognition based on selected acoustic features. In preliminary experiments, we found regression on the actual eye-eye distance too challenging, and four-way classification of V_a , V_b , V_c and V_n to suffer from data sparsity in the V_b and V_c classes. Hence, we unified the V_a , V_b , and V_c classes and

Subject	A	B	C	D	Mean
UAR [%]	69.6	67.0	64.8	68.2	67.4 ± 2.0
AUC	.765	.707	.679	.775	$.732 \pm .046$

Table 2: Results of automatic classification of V_{abc} (Looking at eyes, head, or near head) vs. V_n (looking somewhere else) on the GRAS² corpus, using leave-one-subject-out cross-validation and SVM classifiers. Evaluation in terms of unweighted average recall (UAR) and area under the receiver operating curve (AUC). Mean and standard deviation across four subjects (A–D). Chance level for AUC and UAR is .5 and 50%, respectively.

considered their discrimination from V_n as a binary classification task. For choosing the most relevant features for the attention recognition task at hand, we perform a straightforward ranking based selection, taking into account the CC’ criterion with the minimum eye-eye distance as in the feature relevance analysis described above, but applying feature selection in a cross-validation scheme (leave-one-person out) to reduce the danger of over-fitting to the four test subjects. In particular, for testing on each of the four subjects in the database, we use the remaining three subjects as training data. In this way, we select the 200 most relevant features by CC’ on the training data, and train a support vector machine (SVM) classifier using the Sequential Minimal Optimization (SMO) algorithm implemented in the Weka toolkit [3]. SVMs are particularly suited to learn from large feature sets with probably inter-correlated features. After classification, the unweighted average recall (UAR) of the classes is computed, as well as the area under the receiver operating curve (AUC). We obtain the results shown in Table 2. Both, UAR and AUC are significantly above chance level (.5) according to a one-tailed z -test ($p < .001$), indicating that the selected features generalize across recordings from different subjects. The low inter-subject deviations of the UAR and UAC further indicate the robustness of the obtained classification results.

7. CONCLUSIONS

We have introduced the GRAS² corpus, a multi-modal and multi-sensory corpus of real-life interactions of people seeking for help and directions from strangers in a public shopping center. The corpus has been recorded for the purpose of analyzing the role of visual attention and dialogue behavior in such interactions. Using information from multiple audio tracks we were able to automatically label when the subject carrying the recording equipment or his or her dialogue partner is talking. The analysis of correlations between acoustic features of the voice of the subject and the visual attention (eye contact with dialogue partner) has revealed a low, but meaningful correlation between the acoustics and the distance of the point at which the subject is looking and the eye region of the dialogue partner. Yet, the correlations are strong enough, such that an automatic classification of whether a subject is looking at or close by the head of the dialogue partner or somewhere else based only on automatically extracted acoustic speech parameters is feasible with up to 70% unweighted average recall rate (the chance level would be 50%).

In future work we aim at significantly increasing the size

of the corpus by conducting new recordings with the same setup. We will further manually correct the automatic segmentation and conduct experiments on the short interactions to look at the style of the interactions and analyze the reactions and emotions of the dialogue partners.

8. ACKNOWLEDGMENTS

The research leading to these results has received funding from the European Community’s Seventh Framework Programme (FP7/2007-2013) under grant agreements No. 289021 (ASC-Inclusion) and No. 288587 (MASELTOV).

9. REFERENCES

- [1] F. Eyben, F. Wenginger, S. Squartini, and B. Schuller. Real-life voice activity detection with lstm recurrent neural networks and an application to hollywood movies. In *Proc. ICASSP, Vancouver, Canada*. IEEE, 2013.
- [2] F. Eyben, M. Wöllmer, and B. Schuller. openSMILE – The Munich Versatile and Fast Open-Source Audio Feature Extractor. In *Proc. of the 9th ACM International Conference on Multimedia, MM, Florence, Italy*, pages 1459–1462. ACM, 2010.
- [3] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. Witten. The WEKA Data Mining Software: An Update. *SIGKDD Explorations*, 11(1), 2009.
- [4] D. Miyauchi, A. Sakurai, A. Nakamura, and Y. Kuno. Active eye contact for human-robot communication. In *Proc. of CHI 2004*, pages 1099–1102. ACM, 2004.
- [5] L. Paletta, R. Sefelin, J. Ortner, J. Manninger, R. Wallner, M. Hammani-Birnstingl, V. Radoczky, P. Luley, P. Scheitz, O. Rath, M. Tscheligi, B. Moser, K. Amlacher, and A. Almer. MARIA – mobile assistance for barrier-free mobility in public transportation. In *Proc. CORP, Vienna, Austria*, pages 1151–1155, 2010.
- [6] M. Z. Poh, N. C. Swenson, and R. W. Picard. A wearable sensor for unobtrusive, longterm assessment of electrodermal activity. *IEEE Transactions on Biomedical Engineering*, 57(5):1243–1252, May 2010.
- [7] B. Rammstedt and O. John. Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German. *Journal of Research in Personality*, 41:203–212, 2007.
- [8] B. Schuller, F. Pokorny, S. Ladstätter, M. Fellner, F. Graf, and L. Paletta. Acoustic geo-sensing: Recognising cyclists’ route, route direction, and route progress from cell-phone audio. In *Proc. ICASSP, Vancouver, Canada*, May 2013.
- [9] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Wenginger, F. Eyben, E. Marchi, H. Salamin, A. Polychroniou, F. Valente, and S. Kim. The interspeech 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism. In *Proc. Interspeech 2013, Lyon, France*. ISCA, 2013.

Appendix C

Paletta, L., Santner, K., Fritz, G., Hofmann, A., Lodron, G., Thallinger, G., and Mayer, H. (2013) FACTS - A Computer Vision System for 3D Recovery and Semantic Mapping of Human Factors, *Proc. 9th International Conference on Computer Vision Systems*, LNCS 7963, pp. 62-72, Springer-Verlag Berlin Heidelberg, ICVS 2013, Sankt Petersburg, Russia, July 16-18, 2013.

FACTS - A Computer Vision System for 3D Recovery and Semantic Mapping of Human Factors

Lucas Paletta, Katrin Santner, Gerald Fritz, Albert Hofmann,
Gerald Lodron, Georg Thallinger, Heinz Mayer

JOANNEUM RESEARCH Forschungsgesellschaft mbH,
DIGITAL - Institute for Information and Communication Technologies
Steyrergasse 17, 8010 Graz, Austria

{lucas.paletta, katrin.santner, gerald.fritz, albert.hofmann,
gerald.lodron, georg.thallinger, heinz.mayer}@joanneum.at

Abstract. The study of human attention in the frame of interaction studies has been relevant for usability engineering and ergonomics for decades. Today, with the advent of wearable eye-tracking and Google glasses, monitoring of human attention will soon become ubiquitous. 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 brings new potential into automated processing of human factors, opening new avenues for attention studies.

Keywords. Visual attention, 3D information, SLAM, human factors.

1 Introduction

The study of human attention in the frame of interaction studies has been relevant in usability engineering and ergonomics for decades [1]. Today, with the advent of wearable eye-tracking and Google glasses, monitoring of human visual attention and the measuring of human factors will soon become ubiquitous. This work describes a multi-component vision system that enables pervasive mapping and monitoring of human attention. The key contribution is that our methodology enables full 3D recovery of the gaze pointer, human view frustum and correlated human centered measurements directly into an automatically computed 3D model. It applies RGB-D SLAM and descriptor matching methodologies for 3D modeling, localization and automated annotation of ROIs (regions of interest) within the acquired 3D model. This innovative methodology will open new opportunities for attention studies in real world environments, bringing new potential into automated processing for human factors technologies.

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 and from this enables to precisely analyze human attention in the context of the semantics

of the local environment (objects [18], signs, scenes, etc.). Figure 1 visualizes how accurately human gaze is mapped into the 3D model for further analysis. *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 will provide a scaling up of nowadays still small sketched attention studies – such as, in shop floors, the analysis of navigation, and human-robot interaction – with ca. 10-15 users and thus enable for the first time large scale, statistically significant usability studies.

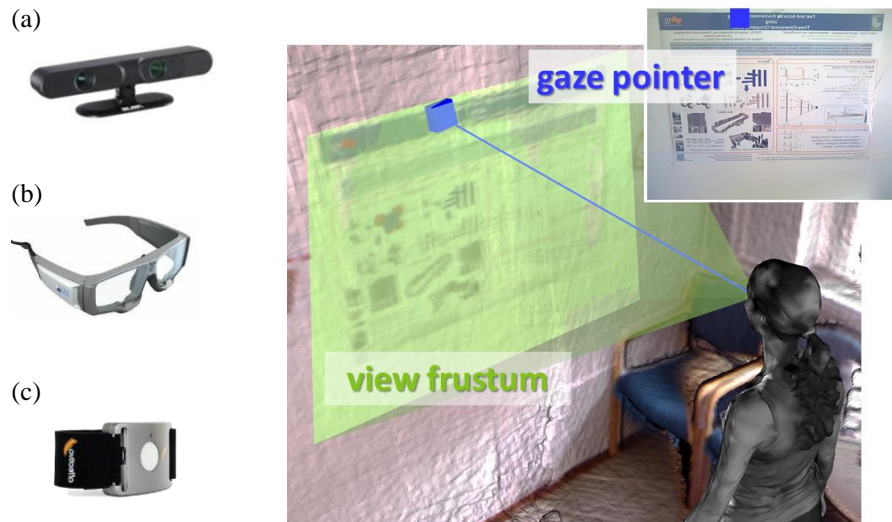


Figure 1. Sketch of sensors used in the study (left) and typical gaze recovery (right). A full 6D recovery of the view frustum and gaze (right) is continuously mapped into the 3D model. (a) RGB-D scanning device, (b) eye tracking glasses (ETG) and (c) bio-electrical signal device.

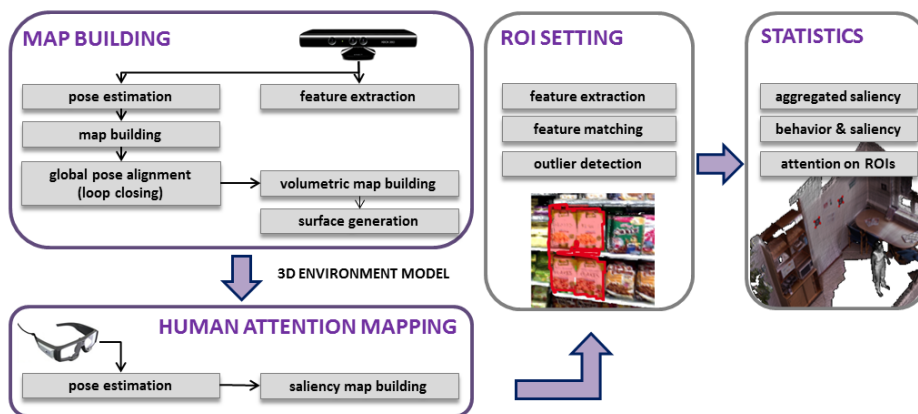


Figure 2. Sketch of workflow for 3D gaze recovery and semantic ROI analytics.

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 [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 then 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, for example, being further evaluated in the frame of user behaviors of interest. The performance evaluation of the presented 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).

2 Related Work

Human Attention Analysis in 3D. 3D information recovery of human gaze has recently been targeted in various contexts. Munn et al. [5] introduced monocular eye-tracking and triangulation of 2D gaze positions of subsequent key video frames, obtaining observer position and gaze pointer in 3D. However, they reconstructed only single 3D points without reference to a 3D model with angular errors of $\approx 3.8^\circ$ (compared to our $\approx 0.6^\circ$). Voßkühler et al. [6] analyzed 3D gaze movements with the use of a special head tracking unit, necessary for their intersection of the gaze ray with a digitized model of the surrounding. Pirri et al. [7] used for this purpose a mass marketed stereo rig that is required in addition to a commercial eye-tracking device, and attention cannot be mapped and tracked. The achieved accuracy indoor is ≈ 3.6 cm at 2 m distance to the target compared to our ≈ 0.9 cm (Paletta et al. [4]). Waizenegger et al. [19] tracked 3D models of conferees for the provision of virtual eye contact; Park et al. [20] presented 3D ‘social saliency fields’ to be established in human communication from head mounted camera views – however, they refer to dynamic in contrast to static parts of the 3D environment as in this work. In general, we present a straight forward solution of mapping fixation distributions onto a 3D model of the environment. The presented work extends through the automated annotation process and discusses the complete system description with shop floor scenario.

Vision Based Dense Reconstruction. Vision based Simultaneous Localization and Mapping (SLAM) aims at building a map of a previously unknown environment while simultaneously estimating the sensors pose within this map. In the last years SLAM has been performed using a huge variety of visual sensor such as single cameras, stereo or trinocular systems [8]. With the launch of range image devices, large scale dense reconstruction of indoor environments has been proposed [9], with real-time dense tracking and mapping system of small desktop scenes (KinectFusion). Dense reconstruction of large cyclic indoor environments has been presented via bundle adjustment techniques and fusion of probabilistic occupancy grid maps with loop closing (Pirker et al. [2]).

Vision Based Localization. Recently, several authors proposed a least-squares optimization routine minimizing the re-projection errors, others perform a perspective n-Point pose estimation algorithm. Both groups rely on correspondences established between 3D model points and 2D image points. In case of a large scale map consisting of thousands of model points, correspondence estimation becomes computationally too expensive. Therefore, image retrieval techniques [3] have been proposed to reduce the number of possible matching candidates.

Logo Detection. Logo detection is done on more general reference material in our work. The base for reference logos is packaging similar to those contained in the Surrey Object Image Library. An evaluation of state-of-the-art algorithms in machine vision object recognition on image databases shows that SIFT performs best on comparative image databases (SOIL-47 dataset [15]).

3 Gaze Localization in 3D Models

Visual Map Building and Camera Pose Estimation. For realistic environment modeling we make use of an RGB-D sensor (e.g. ASUS Xtion ProLive¹) providing per pixel color and depth information at high frame rates. After intrinsic and extrinsic camera calibration, each RGB pixel can be assigned a depth value. Since we are interested in constructing a 3D environment in reasonable time, we perform feature based visual SLAM relying on the approach of [2]. Our environment consists of a sparse pointcloud, where each landmark is attached with a SIFT descriptor for data association during pose tracking and for vision based localization of any visual device within this reconstructed environment. Estimated camera poses (keyframes) 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 error of 2D-3D correspondences. We decided to use the current frame for map expansion if the camera has moved for a certain amount or if the number of positive matches falls below a certain threshold. New landmarks are established using the previously estimated camera pose and the depth information stemming from the RGB-D device. Finally, sliding-window bundle adjustment is performed to refine both camera and landmark estimates. To detect loop closures we use a bag-of-words approach [3]. To close the loop we minimize the discrepancy between relative pose constraints through a pose graph optimization routine [10] followed by natural landmark transformation using corrected camera poses.

Densely Textured Surface Generation. For realistic environment visualization, user interaction and subsequent human attention analysis, a dense, textured model of the environment is constructed. Therefore, depth images are integrated into a 3D occupancy grid [2] using the previously corrected camera pose estimates. Hereby, we follow the pyramidal mapping approach implemented on the GPU. In contrast to existing approaches, we are able to reconstruct environments of arbitrary size. Space is not limited by GPU memory but only by the computer's memory resources. The whole volume is divided into subvolumina, whose sizes depend on the memory architecture

¹ http://www.asus.com/Multimedia/Xtion_PRO_LIVE/

of the GPU (typically 512^3 voxels). Unused subvolumina (e.g. already mapped or not visible) are cached in CPU memory (or any arbitrary storage devices) and reloaded on demand. Realistic surface construction is done by a marching cubes algorithm [12], where overlapping subvolumina guarantee a watertight surface. To apply realistic texture, we use a simple per vertex coloring approach. Hereby, the visible subset of points for each pose is determined using the z-Buffer together with a color buffer based selection technique. Each vertex' RGB color value is computed by projecting it onto the color image plane and taking the running average over all possible values resulting in a smooth, colored mesh (see Figure 3).

3D Gaze Recovery from Monocular Localization. To estimate the proband's pose, SIFT keypoints are extracted from ETG video frames and then matched landmarks from the prebuilt environment and a *full 6DOF pose* is estimated using the perspective n-Point algorithm [13]. Given the proband's pose together with the image gaze position, we are interested in its fixation point within the 3D map. Therefore, we compute the intersection of the viewing ray through the gaze position with the triangle mesh of the model. For rapid interference detection we make use of an object oriented bounding box tree [14] reporting the surface triangle and penetration point hit by the ray. Fixation hits are integrated over time resulting in a *saliency map* used to study and visualize each user's attention in the 3D environment (see Figure 4). For a smoother visualization of saliency and to account for uncertainties in localization and gaze, we use a Gaussian weighted kernel with nearby surface triangulation.

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. We apply the SIFT descriptor to find the appropriate logo in each input frame. Visual tracking has been omitted so far in order not to introduce tracking errors into the 3D mapping step (see Figure 5). For robustness, ROI polygons are filtered if the geometric transformation of nearby frames significantly differs from the identity transform. For ROI identification in 3D model space, we use the keyframe poses estimated as described above together with the ROIs automatically detected in the associated image. Each surface point inside the camera's view frustum is projected onto the image plane and checked for being inside the ROI polygon (resulting ROIs in image and 3D domain are depicted in Figure 5).

4 Experimental results

Eye Tracking Device. The mass marketed SMI™ eye-tracking glasses (Figure 1b) - a 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 $80^\circ/60^\circ$ horizontal/vertical assure a precise localization of the human's gaze in the HD 1280x960 scene video with 24 fps. An accurate three point calibration (less than 0.5° validation error) was performed and the gaze positions within the HD scene video frames were used for further processing. To evaluate our system in a realistic environment we recorded data on a shop floor cover-

ing an area of about 8x20m². We captured 2366 RGB-D images and reconstructed the environment consisting of 41700 natural visual landmarks and 608 keyframes. The resulting textured surface is shown in Figure 3c.

Recovery of 3D gaze. In the study on human attention, 3 proband's were wearing eye-tracking glasses and the Affectiva Q sensor for measuring electrodermal activity (EDA) and accelerometer data. Proband's had the task to search for three specific products which define the ROI in 2D and 3D. The ratio of successfully localized versus acquired frames is described in in Table 1: user 1 and 2 are efficiently localized. In general, blurred imagery, images depicting less modeled area in the test environment and too close views of the scenery cause localization outages that will be improved with appropriate tracking methodology as future work. However, the system allows a fully automatic computation of each proband's path within the environment, the full recovery of gaze and aggregation of saliency over time within the 3D model (Figure 3). The accuracy estimation of the proposed 3D gaze recovery has been reported in [4] with an angular projection error of $\approx 0.6^\circ$ within the chosen 3D model which is therefore smaller than the calibration error of the eye-tracking glasses ($\approx 1^\circ$). The Euclidean projection error was only ≈ 1.1 cm on average and thus enables to capture attention on daily objects and activities (packages, cups, books, pencils).



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

Table 1. Performance of localization in two typical user tracks.

proband	# frames in total	# frames successfully localized
1	1903	1512 (79.45%)
2	1306	1088 (83.31%)

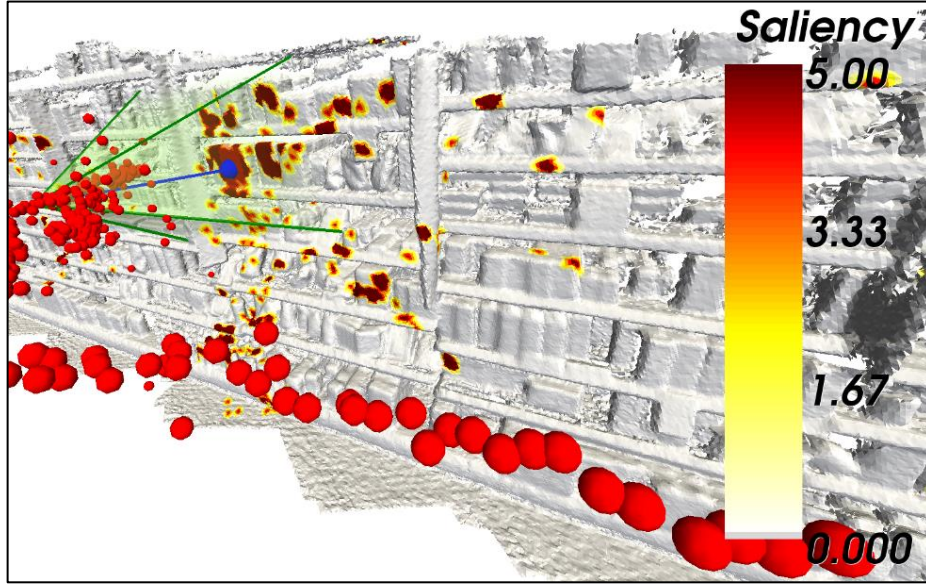


Figure 4. Mapping of saliency on the acquired 3D model and automated recovery of the trajectory of ETG camera positions (spheres). Recovery of frustum (view) and gaze (line with blob).

Table 2. Accuracy evaluation of the ROI detection algorithm.

performance \ ROI	# 1	# 2	# 3	total
# ground truth annot.	21	87	95	203
# logo detections	19	184	82	285
# true positives	19	86	70	175
precision	1.00	0.47	0.85	0.61
recall	0.90	0.99	0.74	0.86
avrg. spatial overlap (σ)	0.87 (0.02)	0.90 (0.06)	0.86 (0.05)	0.88 (0.06)

Table 3. Evaluation results of the three-dimensional ROI computation.

R	O(R) 2D automatic ROI detection	O(R) 2D manual ROI annotation
# 1	0.728369	0.731168
# 2	0.327668	0.848437
# 3	0.611904	0.671895

ROI detection in 2D and 3D. To evaluate the performance of automated 3D ROI association from HD video, we generated ground truth data (Figure 5). For accuracy evaluation of ROI detection, we are only interested in the detection performance, since tracking is not relevant for 3D mapping and robust detections are preferred instead of continuous tracks containing imprecise regions. ROI may be composed of multiple parts, hence the *temporal coverage* by detections of each gives an indication

whether the complete ROI is covered. We have 3 ROIs consisting of 8 parts with temporal coverage between 37% and 98%. The precision and recall values act as a common quality measure of detection performance. Here, we employ an overlap criterion $O(R)$ [18] of a ROI R and its ground truth (R) $O(R) = \frac{A(R \cap G(R))}{A(R \cup G(R))}$, where $A(\cdot)$ denotes of the area of the polygonal region detected in the RGB images. The results are presented in Table 2. While having high precision rates at ROI #1 and #3, very similar products cause the algorithm to produce false positives resulting in a low precision rate for ROI #2 (see Figure 5). Given the high recall and spatial overlap values, the logo detection algorithm provides suitable input for the 3D mapping procedure. To evaluate the reliability, correctness and accuracy of the automatic ROI computation in 3D domain, we manually segmented ROIs in 3D space as ground truth. As an accuracy measure we again employ the overlap criterion defined above, where $A(\cdot)$ denotes the area of the region formed by surface triangles. To show the influence of the accuracy of the ROI detection in the image domain, we compare the 3D ROIs computed out of the *fully automatic* 2D ROI detection algorithm against the *manually annotated* ones (results in Table 3). Clearly, false, missing or imprecise detections in the image domain produce a high error in the 3D space, since the overlap criterion for manually annotated ROIs is higher in each case. This is also visualized in Figure 5, where a single 2D detection outlier results in erroneous 3D ROIs.



Figure 5. Automated ROI detection in the 2D and 3D domain. (left) 3D ground truth annotation. (mid) ROIs computed in 3D out of automatic 2D detections. (right) 3D ROIs computed out of the 2D ground truth data.

Semantic Mapping of Attention. The proposed system allows a fully automated workflow to evaluate human attention performance entirely within a 3D model. The automatic detection of ROIs in three-dimensional space enables the system to provide the user with statistical evaluation without any manual annotation, which is known to be time consuming and error prone. One of the basic indicators when dealing with ROIs is called AOI hit, which states for a raw sample or a fixation that its coordinate value is inside the ROI [16]. ROI #1 received the maximum hits (287) by all users, with the maximum hits counted for user #1 (112 hits). Another example is the dwell – often known as ‘glance’ in human factor analysis – and defined as one visit in an ROI, from entry to exit. The maximum mean dwell time was measured for ROI #1 but by user #3 (133.3 ms). Figure 6 plots the distribution of the dwell times for ROI #1 over all participants. Notice that dwells shorter than 35ms are excluded from the plot to enhance the readability. There are in total 287 hits for ROI #1, 45 visits of the region took at least 35 ms and the longest visit lasted 733.3 ms. From these data we conclude

that only a minority of the captured fixations is related to human object recognition since this is known to trigger from 100 ms of observation / fixation [17]. However, the investigation of human attention behavior is dedicated to future work and we believe that we developed a most promising technology, in particular, for the purpose of studying mobile eye tracking in the field, in the real world, and for computational modeling of attention modeling.

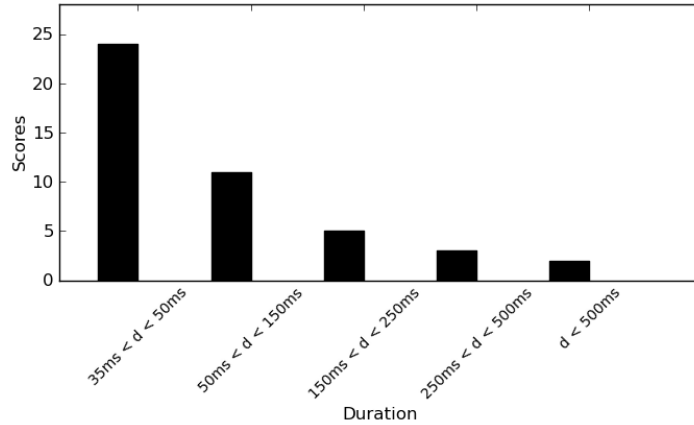


Figure 6. Distribution of dwell times according to their duration for ROI #1. Human object recognition is known to trigger from 100 ms of observation / fixation [17].

5 Conclusion and Future Work

We presented 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 examples from a first relevant user study demonstrate the potential of our computer vision system to perform automated analysis and/or evaluation of the human factors, such as attention, using the acquired 3D model as a reference frame for gaze and semantic mapping, and with satisfying accuracy in the mapping from eye tracking glasses based video onto the automatically acquired 3D model. The presented system represents a significant first step towards an ever improving mapping framework for quantitative analysis of human factors in environments that are natural in the frame of investigated tasks. Future work will focus on improved tracking of the human pose across image blur and uncharted areas as well as on studying human factors in the frame of stress and emotion in the context of the 3D space.

Acknowledgments

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

References

1. Salvendy, G., ed., Handbook of Human Factors and Ergonomics, John Wiley, 2012.
2. Pirker, K., Schweighofer, G., Rüther, M., Bischof, H.: GPSlam: Marrying Sparse Geometric and Dense Probabilistic Visual Mapping. *Proc. 22nd BMVC*, 2011.
3. Nistér, D. and Stewénius, H.: Scalable Recognition with a Vocabulary Tree, *Proc. Conference on Computer Vision and Pattern Recognition (CVPR)*, 2006.
4. Paletta, L., Santner, K., Fritz, G., Mayer, H., Schrammel, J.: 3D Attention: Measurement of Visual Saliency Using Eye Tracking Glasses, *Proc. ACM SIGCHI Conference on Human Factors in Computing Systems (CHI'13)*, extended abstracts, p. 199-204.
5. Munn, S. M., and Pelz J. B.: 3D point-of-regard, position and head orientation from a portable monocular video-based eye tracker. *Proc. ETRA 2008*, p. 181-188, 2008.
6. Voßkühler A., Nordmeier V. and Herholz S.: Gaze3D - Measuring gaze movements during experimentation of real physical experiments. *Proc. Eur. Conf. Eye Mov. (ECM)*, 2009.
7. Pirri, F., Pizzoli, M., Rudi, A.: A general method for the point of regard estimation in 3D space. *Proc. Conf. Computer Vision and Pattern Recognition (CVPR)*, pp. 921-928, 2011
8. Marks, T. K. and Howard, A. and Bajracharya, M. and Cottrell, G. W. and Matthies, L.: Gamma-SLAM: Using stereo vision and variance grid maps for SLAM in unstructured environments. *Proc. IEEE International Conf. Robotics and Automation (ICRA)*, 2008.
9. Izadi, S. and Kim, D., Hilliges, O., Molyneaux, D., Newcombe, R., Kohli, P., Shotton, J., Hodges, S., Freeman, D., Davison, A. and Fitzgibbon, A.: KinectFusion: real-time 3D reconstruction and interaction using a moving depth camera. *Proc. 24th annual ACM symposium on user interface software and technology*, 2011.
10. H. Strasdat, J. M. M. Montiel and A. Davison: Scale Drift-Aware Large Scale Monocular SLAM, in *Proceedings of Robotics: Science and Systems*, 2010.
11. Pirker, K., Schweighofer, G., Rüther, M. and Bischof, H.: Fast and Accurate Environment Modeling using Three-Dimensional Occupancy Grids, *Proc. 1st IEEE/ICCV Workshop on Consumer Depth Cameras for Computer Vision*, 2011.
12. Lorensen, W. E., Cline, H. E.: Marching Cubes: A high resolution 3D Surface Construction Algorithm, in *Computer Graphics*, vol. 21, 1987, pp. 163-169.
13. 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.
14. Gottschalk S. & Lin M. C. & Manocha D.: OBB-Tree: A Hierarchical Structure for Rapid Interference Detection, *Proc. Annual Conf. Comp. Graphics & Interact. Techniques*, 1996.
15. Everingham, M. , Van Gool, L. , Williams, C. K. I. , Winn, J. and Zisserman, A.: The PASCAL Visual Object Classes (VOC) Challenge. *Intl. Journal of Computer Vision*, 2010.
16. Holmqvist, K., Nyström, M., Andersson, R., Dewhurst, R., Jarodzka, H., and van de Weijler, J.: Eye Tracking – A Comprehensive Guide to Methods and Measures, Oxford University Press, 2011, pp. 187.
17. Grill-Spector, K. and Sayres, R. (2008). Object Recognition: Insights From Advances in fMRI Methods, *Current Directions in Psychological Science*, Vol. 17, No. 2. (April 2008), pp. 73-79.
18. Fritz, G., Seifert, C., Paletta, L., and Bischof, H. (2005). Attentive object detection using an information theoretic saliency measure, *Proc. 2nd International Workshop on Attention and Performance in Computational Vision*, Springer-Verlag, LNCS 3368, p. 29-41.
19. Waizenegger, W., Atzpadin, N., Schreer, O., Feldmann, I., Eisert, P. (2012). Model based 3D gaze estimation for provision of virtual eye contact, *Proc. ICIP 2012*.
20. Park, H.S., Jain, E., and Sheikh, Y. (2012). 3D gaze concurrences from head-mounted cameras. *Proc. NIPS 2012*.