Deducing the Visual Focus of Attention from Head Pose Estimation in Dynamic Multi-View Meeting Scenarios

Michael Voit
Interactive Analysis and Diagnosis
Fraunhofer IITB Karlsruhe
Fraunhoferstr. 1
76131 Karlsruhe, Germany
michael.voit@iitb.fraunhofer.de

Rainer Stiefelhagen
Interactive Systems Laboratories
Universität Karlsruhe
Am Fasanengarten 5
76131 Karlsruhe, Germany
stiefel@ira.uka.de

ABSTRACT

This paper presents our work on recognizing the visual focus of attention during dynamic meeting scenarios. We collected a new dataset of meetings, in which acting participants were to follow a predefined script of events, to enforce focus shifts of the remaining, unaware meeting members. Including the whole room, all in all, a total of 35 potential focus targets were annotated, of which some were moved or introduced spontaneously during the meeting. On this dynamic dataset, we present a new approach to deduce the visual focus by means of head orientation as a first clue and show, that our system recognizes the correct visual target in over 57% of all frames, compared to 47% when mapping head pose to the first-best intersecting focus target directly.

Categories and Subject Descriptors

I.4.8 [**IMAGE PROCESSING AND COMPUTER VISION**]: Scene Analysis

General Terms

Algorithms, Experimentation, Human Factors

1. INTRODUCTION

A main focus of today's research in human-computer interaction involves establishing human-like input and output modalities like speech recognition, computer vision and speech synthesis. The development of *smart spaces* copes with the issue that the operation of such modalities strongly depends on context-awareness. Microphones, loudspeakers and cameras need to be arranged and connected in order to provide an unobtrusive setup of perception to build user and situation models. Particular clues for understanding observed scenarios are the recognition of all involved persons [4] and their positions[6, 8], knowledge about objects introduced and being used as well as about every participant's occupation. To learn about interaction partners, group interest or objects on which someone is working, the looking direction of that respective person is a primary clue to establish knowledge about the observed action or target he or she focuses on. The observation of eye gaze would hereby allow to directly map the looking direction onto the focused target. Difficulties in capturing gaze however, due to the obtrusiveness of head mounted gear or particular camera setups that prevent free movement, make it hardly possible to directly use it as a input modality in real life scenarios. Recognizing the visual focus of attention, the target a person is looking at, without knowing about the true gaze direction is a fairly new task in human-computer interaction and was introduced in 2002 by Stiefelhagen et al. [10]. By means of a panoramic camera on top of a meeting table every participant's head orientation was captured and the meeting partner he or she was orientating to was deductively assigned to be the (visual) focus target. The number of participants was predefined and fixed, the setup monocular and no further objects were allowed to be used. Further work, presented by Ba et al. [1, 2] in this field, continued with a predefined number of meeting participants, but introduced the meeting table and whiteboard as valid targets. Dynamics, as walking people, were solely allowed in a small parallel study with a fixed advertisement being installed on a window and surveying if pedestrians focus upon it while passing by [12]. Despite the new context, the trajectories of the pedestrians were fixed and known in advance, the target to look at predefined and static. In contrast, the embedding of perceiving visual focus of attention in above described smartroom setups demands for coping with undefined scenarios such as people entering and leaving the room, new objects that are introduced during discussions or sudden interrupting sounds that shift people's focus to regions, that previously were not modeled or included.

This paper presents our efforts in coping with these dynamics and describes our initial system to deduce the visual focus of attention of a person based on head orientation as a first clue. We use a setup of multiple camera views in order to achieve unobtrusive captures of meeting participants. This allows us to robustly estimate viewing frustums, even when the corresponding person walks around freely in the room or turns the back of the head towards one or two cameras. From the estimated field of view, we deduce the most likely focus target by using an adaptive scheme of mapping head orientation to its most likely gaze angle counterpart, and hence the target this person is looking at.

ICMI'08, October 20–22, 2008, Chania, Crete, Greece.

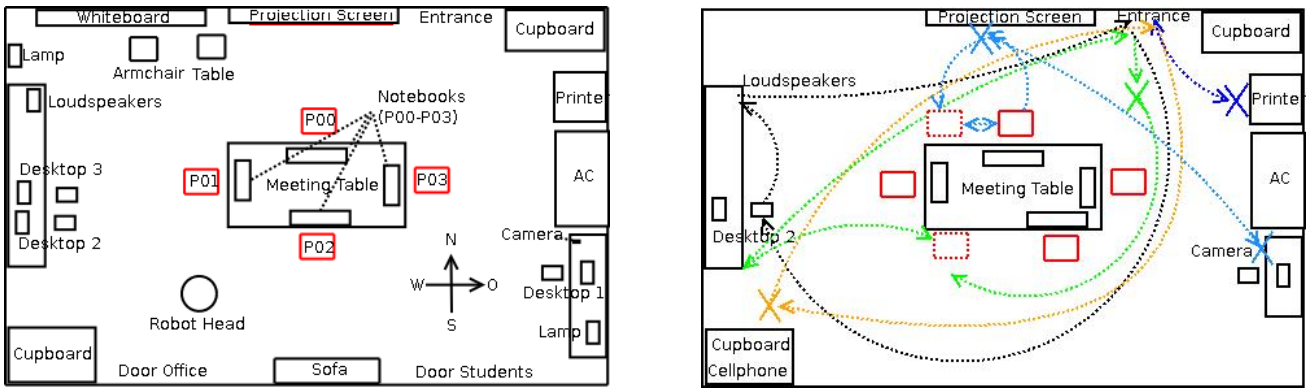

Figure 1: Left: Overview of annotated focus targets within the meeting room. Right: Observed trajectories of all meeting participants. Rather than only gathering meetings with fixed seating positions, participants were advised to walk throughout the entire room, to distract the visual focus of the remaining meeting members and force eventual systems to adapt to new positions.

To have to cope with as many dynamics as possible, we collected a new dataset upon which our system is evaluated. The captures consist of 10 meetings, in which all participants were to enter the room at the very beginning and leave the room when the meeting was finally over. Several times, the captured meetings were interrupted by scripted events to qualify for spontaneous situations. We describe this collection in detail, along with its annotations in Section 2 where we introduce the complete target space and the setup of all sensors and participants. The following Section 3 then describes our approach how to model all potential targets, estimate head orientation over multiple views based on annotations of peoples' head bounding boxes and our implementation to mapping recognized head orientations to likely focus targets. The finalizing segments 4 and 5 present our experimental evaluation of the system on the described dataset and analyze error sources and further enhancements as well as planned integrations and future work in this field.

2. DATASET

We collected a new dataset to gather dynamic meeting scenes, with people entering and leaving the room, walking around or interrupting ongoing discussions. Every video is about 10 minutes long and we collected 10 videos in total. During each meeting, variantly two or three participants were actors and were henceforth involved and taught about predefined events they were to initiate and follow during the recordings. The remaining one or two (variantly) meeting members did not learn about the meeting's topic or what was about to happen. All observations of at least those participants were therefore guaranteed to be spontaneous and unplanned. The predefined set of events included spontaneously joining actors during ongoing discussions, which forced the remaining participants to move with their chairs and make room for the new meeting member, suddenly ringing cellphones, hidden inside cupboards, and actors entering the room looking for them, loud disruptive noises coming from nearby loudspeakers that one of the actors was spontaneously initiating without letting the rest of the participants know what he or she was doing, the sudden use of printers and actors pretending paper jams while talking loud about the printer and forcing the meeting group's focus towards him- or herself and actors walking towards the projection screen and giving a short presentation, including the introduction of an ambiguous object (a camera), that was placed on a table besides the meeting scene beforehand and now grabbed and put on the meeting table for everyone to see it. When acting participants entered the room, they were instructed to pass by the meeting table along every possible edge to ensure they interrupt the meeting. Figure 1 shows some of these observed trajectories and clearly proves, that movement was observed throughout the whole room. The meeting table was placed in the center of the room. The seating positions of all participants varied over the recordings to ensure as much variance as possible and avoid static, locations of all persons. The tagging of all persons happened along their position around the table's edges: P00 was always sitting at the northern edge, P01 at the table's western edge, P02 to the south and P03 at the eastern edge of the table. Person P04 was always chosen to be the interrupting actor that enters the scene from time to time. Figure 1 depicts an overview of the seating locations and the overall layout of the room along with its targets.

2.1 Sensor Setup

The dataset's sensor setup consisted of four fixed cameras in the upper corners of the room, each recording with a resolution of 640×480 pixels and 15 frames per second. The purpose of these cameras is to obtain an instant *coarse* overview of the whole room at all times, for allowing people to move and behave as naturally as possible and walk around and interact with each other without being limited by a predefined setup or a restricted sensor range. The camera array was extended with a panoramic view from a fisheye lens camera that was installed on the room's ceiling (following the same specifications). To further capture the complete context of the meetings, audio was recorded by means of four T-shaped microphone arrays, each installed on every wall of the room (northern, western, southern and eastern side) providing basis for audio source localization, and one table-top microphone for speech recognition and acoustical event modelling.

Figure 2: Example scene of one meeting video. Shown are two out of four camera views from the room's upper corners and the panoramic view, captured from the ceiling. In this scene, the interrupting actor passes the meeting table towards the main entrance and thereby walks in between the projection screen and person P00 sitting in front of the screen, working on his notebook. From the camera views only, it is not unambiguously clear if the visual focus of person P02, sitting opposite and wearing the motion sensor on her head, resides on the screen, person P00 or the passing-by actor.

2.2 Annotations

The dataset has been annotated for every participant's visual focus target and head bounding box. The latter labels allow for distinguishing between the different persons and act as stable head locations for head pose estimation we use later on in this paper to approximate the respective gaze direction. The visual focus target was defined to be a corresponding person or object a meeting participant was looking at. To ensure unprejudiced labels, we employed different students, who were not technically involved in the project or research. By only having all camera views available, the students were to decide for each single frame, at whom or what every single participant was looking at. No audio channels were made available for these annotations, the tool the annotators were using only allowed to switch back and forth between the different views for this particular frame. All videos were processed frame-by-frame, manually. We allowed almost every object in the room to be a potential focus target. This included all doors, desktops, tables, lamps, whiteboards and projection screens, notebooks, cameras, air conditioning, cupboards, every person present in the scene and more. Even a camera, placed on top of a tripod, which we use in further research projects as a replacement for dedicated robot heads and was installed for secondary evaluations of other systems, was included in the annotation process. All in all, we gathered a set of over 35 focus targets the annotators had to choose from in each frame, wherein the positions of all participants and various objects were dynamic and changed throughout the meetings. The position and dimensions of each object was measured in 3D. We also varied with the overall number of participants to ensure differences in the distribution of seating positions around the table.

2.2.1 *Annotator Reliability*

The annotations for all participants' visual foci were made by two different annotators with the background to compare our system's overall performance with that of human decisions and specific differences in their annotations. Over 68420 frames in total were annotated by each annotator independently.

Concerning annotation agreement, we computed two common metrics used in analyzing interrator reliability: Cohen's unweighted Kappa and the proportion of agreement [5]. Cohen's Kappa describes the amount of concordant annotations out of all annotations that were actually expected to be non-concordant if both annotators were choosing the target by pure chance and is defined by the following equation:

$$\kappa = \frac{p(a) - p(e)}{1 - p(e)} \tag{1}$$

Here, $p(a)$ symbolizes the observed agreement among the two annotators and $p(e)$ described the agreement to be expected if both annotators where choosing by chance.

The second metric, proportion of agreement, on the other hand depicts the amount of frames $\#Agree$ two annotators selected the same specific class concordantly, in proportion to the overall amount of frames $\#Overall$ in which either annotator selected it:

$$POA = \frac{\#Agree}{\#Overall} \tag{2}$$

As can be seen in Figure 3 Cohen's Kappa shows our annotations to be concordant in $\sim$ 70% of all frames. Most dominantly, the meeting participants themselves, their notebooks, the main entrance, the projection screen, the printer, the camera, the cellphone and one cupboard (inside "Cupboard (SO)" the cellphone was hidden before the meeting, so that although the cellphone's noise could be heard, its source could not be targeted on first sight) are focused by either participant. The camera resembles the object, that the presenter grabs and introduces during his or her presentation.

Confusions mostly happened on the whiteboard, the remaining doors, further furniture or the ceiling. All dominant targets have in common, that they were specifically introduced during the meetings: either by being used or talked about by one participant or by drawing attention themselves, making loud or disturbing noises. The secondary targets all share to be looked at when one participant changed focus and rotated his or her head towards a new direction (or moved his or her field of view generally) and hence included those targets in the shifted viewing frustum, which

Target	Focused (%)	κ	POA
P00	16.0	0.665	0.554
P01	13.7	0.743	0.636
P02	9.8	0.724	0.603
P03	10.5	0.736	0.617
P04	7.9	0.752	0.629
Notebook P00	3.3	0.687	0.537
Notebook P01	2.2	0.709	0.557
Notebook P02	2.3	0.800	0.674
Notebook P03	2.2	0.702	0.549
Notebook P04	0.0	0.000	0.000
Chair P04	0.6	0.475	0.314
Meeting Table	2.8	0.116	0.066
Whiteboard	0.1	0.029	0.015
Projection Screen	11.5	0.748	0.641
Printer	1.3	0.868	0.770
A/C	0.1	0.133	0.072
Cupboard (NO)	0.1	0.198	0.110
Cupboard (SO)	0.8	0.678	0.515
Main Entrance	1.3	0.627	0.461
Door Students	0.1	0.225	0.127
Door Office	0.1	0.207	0.116
Loudspeakers	0.5	0.648	0.481
Small Table	0.1	0.347	0.211
Desktop 1	0.1	0.305	0.181
Desktop 2	4.3	0.935	0.882
Desktop 3	0.5	0.479	0.317
Camera	5.0	0.685	0.537
Cellphone	0.5	0.682	0.519
Robot Head	0.1	0.108	0.058
Sofa	0.2	0.222	0.126
Armchair	0.0	0.129	0.069
Lamp 1 (SE)	0.0	0.105	0.056
Lamp 2 (NW)	0.0	0.019	0.010
Floor	1.8	0.518	0.355
Sensor Stool	0.1	0.298	0.176
Ceiling	0.1	0.030	0.015
Total	100.0	0.707	0.735

Figure 3: Overview of all annotated targets in the room, how often these targets were focused on in percent by either participant and average annotator reliability for this target over all participants, depicted with both Cohen's Kappa and the Proportion of Agreement.

often happens when someone was staring into emptiness. In these cases it seems, that the two annotators variantly decided between secondary targets and the primary involved in the meeting and could not achieve concordant selections. As for the remaining confusions, when specifically looking at the confusion matrices in detail, the remaining cases where the annotations do not agree, often show situations where primary targets nearby are confused, as for example the presenter standing right in front of the projection screen and the screen itself. A good example can be given with meeting member P02 (who is sitting opposite to the projection screen at the southern edge of the table): in 10898 frames, both annotators decided for him or her to focus on person P00 sitting right in front of P02 (at the table's northern side, in front of the projection screen). In 701 frames however, annotator 2 decided for person P00 to be the target, while annotator 1 recognized the screen to be focused. In 3928 frames, it was the other way round. Compared to concordant frames, the latter case happened in 36% of the time and shows that these situations are problematic, even for humans. These *misclassifications* happened especially during both presentations, while the presenter (P00) was walking around in front of the screen, explaining slides and bullet points upon it and times when person P00 was sitting calmly in front of the screen, crossing P02's looking direction towards the slides. Since the pupils of all participants are hardly visible, there is not always the possibility to unambiguously decide between two or more targets. At least not in only such a frame-by-frame decision task where only the captures of the corner cameras are available as in our task. Further information seems to be necessary to properly distinguish between nearby targets: was the presenter pointing at one specific item on the slide and *forced* the audience's focus towards it? Is the presenter currently speaking towards person P02 so that the interaction between the two clearly shifts P02's focus onto the presenter? Were any of the remaining participants discussing items on the projection screen and P02 was likely to follow the discussion? Or a well common case in our meeting videos: Was the presenter only passing by the projection screen, while P02 actually focused upon the projected slides? The latter cases, where moving targets distract the focus of one person are quite often, when the interrupting actor enters the room and passes by the meeting table. Suddenly, a second target, a moving person behind a sitting one, appears. The only possibility to decide for the correct one, is to either include previous context or zoom in on eye gaze directly: In over 1072 frames, the annotators were not concordant when choosing either the interrupting person P04 (who enters the room from time to time and passes by the meeting table on his trajectory) or the person P00. These were clearly the cases, when the actor was walking behind P00's seating position or crossing between P00 giving a presentation and the meeting table. Figure 2 depicts such an example. Another strong example can be found with the cupboard, right next to the main entrance of the room was concordantly chosen to be focused in approx. 20% of its selected frames. Confusions were made with persons standing next to it, the floor, or either the printer or entrance right beside it. The same observations can be made with all secondary targets from the table depicted in Figure 3. All these targets have in common, that they only began to be part of scene, once a person was passing its respective position or started to cope with it (as for example the printer, which was part of the scripted events, in that one acting person began to print papers but ran out of ink or encountered a paper jam). Then again, all these targets happened to be difficult to being distinguished, which shows that the looking direction alone only dictates a frustum of objects to be likely focused at within, but does not allow for a clear separation of this cluster into individual subparts. For the example of the actor encountering the paper jam when printing, both the actor and the printer (as well as the shelf the printer is placed on) merge into one likely target cluster. Therefore, one question to ask certainly is, how and when to generalize targets into joint clusters of interest?

3. ESTIMATING THE VISUAL FOCUS OF ATTENTION

As stated in section 2.2, we define the visual focus of attention of a person to be the target he or she is visually looking at and focusing on. For this purpose, the respective looking direction is most likely determined by tracking

Figure 4: Camera 1's view of a recorded meeting scene during a short presentation, given by person P00. Person P02, sitting opposite to the presenter, is wearing the magnetic motion sensor to capture her true head orientation, depicted by the red (x), green (y) and blue (z) coordinate axes. All *axis aligned bounding boxes* of focus targets we annotated, visible from this view, are highlighted in white.

the person's eye gaze. However, due to the obtrusiveness of head mounted gears and restrictions in freedom of movement when using respective camera setups instead, tracking the person's head orientation and hence viewing frustum is often used as an approximation of the actual gaze.

3.1 Estimating Head Orientation

The use of single camera setups for recognizing the head's pose was already subject of a lot of research and system implementations. We, too, use head orientation to deduce the looking direction from the estimated viewing frustum. For this, we established a single-view system based on a Neural Network approach that easily adopts to our smart-room's multi-view setup [9]: one single Neural Network is trained to output a likelihood distribution over a specified range of head pose angles, relative to the camera's line of sight. For horizontal rotations, we trained the network from $-180°$ to $+180°$, with $10°$ wide classes and use the cropped and grayscaled head region along with its edge magnitude as the network's input. The network's estimation in relative angles allows to apply the same classifier on further, different views, easily extensible without retraining the whole setup. We use this advantage to gather the distributions from multiple views and merge them in a Bayesian filter scheme, to obtain a joint and more robust estimate based on the information from different angles and existing redundancy of overlapping captures. This multi-view integration no longer depends on (near-)frontal or profile captures but allows all person to rotate their head freely within the whole room.

Applying tracking of head orientation in multi-sensor installations is a fairly new topic in establishing human-oriented input modalities. In 2006, the CLEAR Workshop [11] introduced first international evaluations for this task, and provides until now the main databases for comparing different, recently built systems [9, 3, 7]. We evaluated our approach on the provided dataset and observed results as low as $8°$ for horizontal and $13°$ for vertical pose recognition.

3.2 Deriving Visual Focus

Estimating head orientation keeps track of a person's field of view, but does not allow to gather detailed information about that person's looking direction within the estimated viewing frustum. To recognize for a target, that lies within that field of view, to be looked at, individual head turning styles and gaze patterns have to be coped with [10].

3.2.1 Target Modeling and Field of View

We describe a focus target with its *axis aligned bounding box* in the room's global coordinate system (see Fig. 4). In order for a target to be directly looked at, the gaze vector must intersect with the respective box. A more generic definition would be, that the nearer a target's box lies towards the viewing frustum's center, the more likely that target would be looked at. We defined this cone to open with $60°$ horizontally and $50°$ vertically. A potential target F_i thus lies within the viewing frustum, if its axis aligned bounding box contained at least one point $P_i = (x, y, z)$ on its shell within that cone. For gaining that *representational point* P_i, we computed the nearest point (by its euclidean distance) on the box, relative to the head orientation vector. P_i either resembles a true intersection or a point on the box' edges. P_i is verified to reside within the viewing cone - targets outside the viewing frustum are ignored, their likelihood to be focused was set to 0.

3.2.2 Mapping Pose to Targets

Our focus model builds on [1], which concludes a linear correlation between corresponding gaze angles α_G towards targets and observable head turning angles α_H when focusing on them:

$$\alpha_H = k_\alpha \cdot \alpha_G \tag{3}$$

We analyzed this relation for dynamic and moving persons and objects by computing α_H based on the annotations we made upon our dataset and all targets representational points P_i described in 3.2.1. A measured mapping coefficient k_α could thus be obtained with

$$k_\alpha = \frac{\alpha_H}{\alpha_G} \tag{4}$$

and was computed over our dataset. As it showed, the mean value of k_α was found to be 0.72, which we used for a

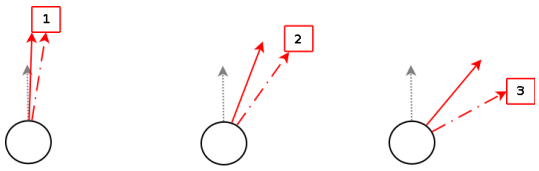

Figure 5: Example for using a fixed mapping factor k_α as in equation 3: Three static objects are presented successively. The spotted grey arrow defines the person's upper body resting orientation, thus his or her initial head rotation. The observed head orientation (solid red arrow) shows a smaller angle than the actual angle towards the focused target (dashed red arrow). The distance is coped with eye movement.

fixed mapping of head rotation to likely focus target. However, computing the mapping coefficient over the dataset, we could see, that k_α does not stay constant over time, but rather changes, depending on the scene focus is shifted and disrupted in. Figure 6 depicts k_α's changes during a 30 seconds long scene.

We computed k_α's values for those participants, that were wearing the motion sensor and computed the groundtruth mapping coefficient given the measured head orientation. Especially during focus changes between two dedicated targets, head orientation clearly points right between those two targets and only shifts towards either one, but does not exceed over the target's position. Here, a constant mapping factor would not classify for those two targets correctly, since the person or object positioned with the lower gaze angle needed for a much lower coefficient value in order to shift head pose backwards instead of even further away. In particular, equation 3 assumes that head orientation always tends to be lower than the real gaze angle. In our case, for one of the two targets, gaze is actually lower than head pose, due to constant switching forward to a target positioned further away. We therefore define a discrete set of possible coefficients (k_α, k_β) for mapping horizontal and vertical head rotation angles α_h and β_H, and reweigh them by means of the most likely focus target F_i's a-posteriori probability, given a corresponding mapping:

$$\pi_{(k_\alpha,k_\beta),t} = \gamma \cdot \pi_{(k_\alpha,k_\beta),t-1} + (1-\gamma) \cdot \arg\max_{F_i} p(F_i|\Phi_{k_\alpha,k_\beta}) \tag{5}$$

The mapping coefficient pair (k_α, k_β) with highest weight $\pi_{(k_\alpha,k_\beta),t}$ is chosen for mapping head pose and finally classifying for the target, that shows maximum a-posteriori probability.

Since most coefficients might intersect with a target, hence return a high likelihood for the given transformation, each target includes an a-priori factor for stating the probability of actually focusing it or changing focus towards it.

The a-posteriori likelihood is defined by

$$p(F_i|\Phi_{k_\alpha,k_\beta}) = \frac{p(\Phi_{k_\alpha,k_\beta}|F_i) \cdot P(F_i)}{p(\Phi_{k_\alpha,k_\beta})} \tag{6}$$

with $\Phi_{k_\alpha,k_\beta} = (\frac{\alpha_H}{k_\alpha}, \frac{\beta_H}{k_\beta})$ being the adapted head orientation with the horizontal rotation α_H, transformed with the mapping factor k_α and β_H being the vertical head rotation transformed with k_β.

The a-posteriori probability of a target F_i is composed of different factors that describe possible models of the scene's context. By now, we simply include the likelihood of looking at this target in the last n frames and secondly a change of pose to the target in the current frame T:

$$P(F_i) = \varphi(\frac{\partial(\angle(\Phi, F_i)))}{\partial t}) \cdot \frac{1}{n} \sum_{t=T-n}^{T-1} (p_t(F_i|(\Phi_t))) \tag{7}$$

The angular difference $\angle(\Phi, F_i)$ describes the distance between the real head orientation and target F_i's representational point P_i. If the head is rotated towards a target F_i, the angular difference decreases, hence its derivation $\frac{\partial(\angle(\Phi,F_i)))}{\partial t}$ shows peaks of negative values and implies a more likely focus change towards that particular target. φ was implemented to be a simple, linear likelihood distribution over the possible derivation values: the lower the derivation value, the stronger and faster a person rotates his or her head towards the target, the more likely the focus changes.

Mapping / Gaze	Measurements	Pose Estimates
$k_\alpha = 1$	47 (54)	30 (35)
$k_\alpha = adapt.$	57 (60)	46 (49)

Table 1: Recognition results in percentage for estimating visual focus targets. Depicted are the results when using either head pose estimates or measured head orientation angles, both for our adaptive approach and a direct mapping from looking direction to first-best target that intersects with the observed looking vector. The values in brackets depict results when evaluating only on targets on which both annotators agreed concordantly.

4. EXPERIMENTAL EVALUATION

The dataset we collected included an annotated set of over 35 potential targets throughout the room. During our experimentation, we experienced, that the fine-granulated target-definition did not match the complexity of mapping head orientation onto likely targets. For this first evaluations, we therefore reduced the target space to meeting participants, meeting table and projection screen only, targets that are included in current state of the art systems that only copy with static scenes. This included 86% of all focused objects for person P02 (against who we are evaluating our system, due to the available groundtruth measurement of his or her head orientation and the fixed upper body orientation) and reduces complexity both for these first evaluations and annotations, which are still happening and take a lot of time to define all object and person positions. Due to missing upper body annotations for all remaining participants, our evaluations only included estimating focus for person P02, wearing the magnetic motion sensor, whose body orientation was always made sure to show towards the projection screen because of the fixed setup with the transmitter and cable.

4.1 Recognizing the Visual Focus

Obviously, the results are still rather low, although the complexity of the task was drastically reduced. Clearly visible is, that our adaptive mapping of head pose to the respective focus target increases the recognition rate in average by over 10%. The results showed, that especially rapid focus changes between two targets were difficult to detect: slight head rotations towards respective targets were observable in our videos but mostly gaze was used to switch back and forth between those two - hence, our system did not recognize a focus switch and stayed fixed upon the same target the whole time.

Secondly, moving targets, for example the already briefly described person P04, passing by between the meeting table and the projection screen (depicted in Fig. 2), distracted person P02's visual focus by quick eye movements only, instead of letting his or her head rotate to follow that respective trajectory. The a-priori likelihood described in equation 7 includes the derivation of the difference between head pose

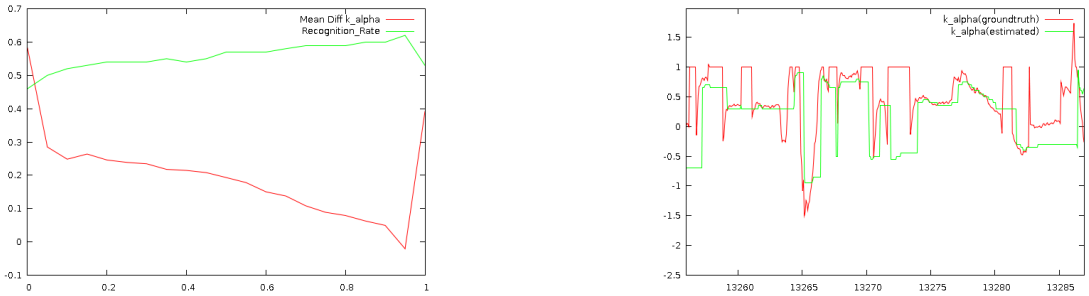

Figure 6: Left image: Recognition Rate (upper green plot) and mean difference of estimated k_α to groundtruth k_α (lower red plot), with respect to increasing adaption factor γ. A value of $\gamma = 1.0$ describes that the scores π_{k_α} are not adapted at all. In this case, the constant mapping coefficient $k_\alpha = 0.72$ was used, which showed to be the measured mean mapping factor over all videos for person P02. Right image: Groundtruth (red plot) versus estimated (green plot) mapping values k_α in a 30sec. long scene.

and a target's gaze angle. This derivation even shows peaks, if head orientation stays fixed and the target passes by, since then, the angular distance decreases down to the point where head pose and the trajectory intersect. This factor seems to provide a possible basis for recognizing focus changes, but does not allow to distinguish between *real* focus changes and moving objects or persons. And the third source for false focus targets is the merging of two targets into one joint cluster of interest: as described above, a presenter giving a presentation is hard to distinguish from the screen he or she is pointing at. Eye gaze follows the pointing gesture, the respective head remains still, expect when the item the presenter is pointing at is either far away or seems to be of interest for a longer time. The slight focus change towards the screen and away from the presenter, or even the other way round from the screen towards the presenter is only recognizable if speech and gestures are taken into account as well for detecting the focus shift that was initiated and enforced by the presenter himself. The same happens with other objects as people are occupied with them: a malfunctioning printer a person is working with, only can be detected as the primary focus target if either the looking person's eye gaze can be observed or the origin of the viewing cone is near enough to the target, so that the cone's edges only include the target itself (which would mean for the person to stand directly in front of the printer and not sit at the table). Since both are hardly the case, this clearly leaves our system to include corresponding generalizations to merge both printer and the person working on it into one joint region of interest and one single target. This is not yet included in our approach.

4.2 Head Orientation Errors

We compared our head orientation estimates with the measurements of the attached motion sensor Person P02 was wearing. In total we experienced a mean error of 8° for horizontal and 17° for vertical estimates. This clearly shows, that distinguishing targets in vertical direction is very ambiguous for our system. Targets involved are mostly the meeting table, the person sitting opposite and the projection screen. Often, these three targets are all successively in one vertical line and can only be distinguished from P02's tilting. Scenes, that show extreme tilt estimation errors are, when P02 is occupied writing on his or her notebook right in front of the table (hence show a high tilting). For some participants, also hair falls into their face, further occluding the camera's sight onto the (already low) facial details. In these cases, our system estimates a false tilt angle and hence detects the visual focus to lie on either the screen or the opposite sitting person. Since we instructed all participants to use their notebook and keep notes, this happens quite often and thus shows to be the main error source when using estimated head poses. Another problem, we observed over our dataset, are individual head rotating styles, by which everybody uses different tilt angles in his or her resting position. We encountered participants to tilt their head slightly downwards when actually focusing straight ahead. Of course, this leads to the table being recognized as the primary focus target when the person really is looking at the screen. Further, head is turned very slowly, often dragging behind the actual gaze and focus change. Considering equation 7, the rapid rotation towards a target is then missing and a change is not detectable.

General questions that are to be answered are in our next work are, how head orientation correlates to moving targets and if a fitting user model for this perception can be found during meetings (do people tend to follow behind the target's trajectory or do they rather estimate the trajectory in advance and adapt to movement changes?) as well as how several focus targets merge into one single group of interest for particular meeting members or objects instead of necessarily distinguishing between every single item. Future work also includes fast estimation of upper body orientation to easily recognize every meeting member's resting position and initial head orientation when looking straight forward. This cue should also show strong correlation to group behavior and allow focus target abstractions by separating persons into groups, analyzing group roles and including multi-person focus of attention and region of interests with respect to individual groups and their interactions. Further, the looking direction of the remaining participants might play a strong role in distinguishing between nearby targets, since their respective focus might be less ambiguous and thus helps in increasing the likelihood for either target in ambiguous situations (if more people are actually focusing on the same target, its likelihood should increase).

5. CONCLUSION

This paper presents a new data collection of dynamic meeting scenarios and our effort on recognizing the visual focus of attention by means of head orientation from the meeting participants. The dataset contains 10 videos, each approximately 10 minutes long, providing highly dynamic scenes, in which people enter and leave the room, give presentations and introduce new objects into ongoing discussions. A predefined script of events ensured that all videos contained the same amount of events, a subpart of the meeting participants being actors, were advised to initiate every event as instructed. The remaining meeting members were unaware of the happenings to make sure that their reaction was spontaneous and unplanned. This dataset was annotated for every participant's visual focus of attention in every frame by two different students. All in all, a total of 35 targets (all persons and objects in the room) were allowed to be looked at. The annotations are compared with two different interrator reliability metrics (Cohen's Kappa and Proportion of Agreement) and analyzed for their differences. We further describe our efforts on estimating head orientation to recognize the direction in which participants are looking and the deduction of the most likely person or object they visually focus on. Due to the complexity of distinguishing between all 35 targets, for these first evaluations, we reduced the target space to the primary objects and the participants in each meeting to be allowed foci ($\geq$ 85% of all targets). In 57% of all frames, our system recognizes the correct object or person being looked at, despite the included dynamics of moving targets and sudden interruptions. Compared to a direct mapping of the participant's looking direction onto the first-best intersecting target, our approach performes 10% better, which shows that head pose not always directs towards the target a person is focusing on. Slight eye gaze movements within the viewing cone often overcome the distance to targets nearby. Current and ongoing work and research include the analysis of the targets' movements, adding a correlation model to moving focus targets and extending the target space to all annotated objects in the room. In order to adopt our approach to every meeting participant, independent of his or her movement, research on estimating upper body orientation is due to be done. Since this approach only relied on visual features for deducing focus, the recorded audio context is subject for further research, too: speaker diarization, knowledge about interrupting noises and explicitly introduced objects during discussions intuitively provide further clues to building an overall situation model and help enhancing the recognition of one's visual focus of attention.

6. ACKNOWLEDGMENTS

This work was supported by the FhG Internal Programs under Grant No. 692 026.

7. REFERENCES

[1] S. Ba and J. Odobez. A Cognitive and Unsupervised Map Adaptation Approach to the Recognition of Focus of Attention from Head Pose. In *Proceedings of International Conference on Multimedia and Expo*, 2007.

[2] S. Ba and J. Odobez. Multi-party Focus of Attention Recognition in Meetings from Head Pose and Multimodal Contextual Cues. In *Proceedings of International Conference on Acoustics, Speech and Signal Processing*, 2008.

[3] C. Canton Ferrer, J. Casas, and M. Pardàs. Head Orientation Estimation using Particle Filtering in Multiview Scenarios. In *Proceedings of the Second International Evaluation Workshop on Classification of Events, Activities and Relationships*, 2007.

[4] H. Ekenel, M. Fischer, and R. Stiefelhagen. Face Recognition in Smart Rooms. In *Proceedings of 4th Joint Workshop on Multimodal Interaction and Related Machine Learning Algorithms*, 2007.

[5] Harold L. Kundel and Marcia Polansky. Measurement of Observer Agreement. *Radiology*, 228:303.

[6] Keni Bernardin and Rainer Stiefelhagen. Audio-Visual Multi-Person Tracking and Identification for Smart Environments. In *Proceedings of ACM Multimedia*, 2007.

[7] O. Lanz and R. Brunelli. Joint Bayesian Tracking of Head Location and Pose from Low-resolution Video. In *Proceedings of the Second International Evaluation Workshop on Classification of Events, Activities and Relationships*, 2007.

[8] O. Lanz, P. Chippendale, and R. Brunelli. An Appearance-based Particle Filter for Visual Tracking in Smart Rooms. In *Proceedings of the Second International Evaluation Workshop on Classification of Events, Activities and Relationships*, 2007.

[9] Michael Voit, Kai Nickel, and Rainer Stiefelhagen. Head Pose Estimation in Single- and Multi-view Environments - Results on the CLEAR'07 Benchmarks. In *Proceedings of the Second International Evaluation Workshop on Classification of Events, Activities and Relationships*, 2007.

[10] Rainer Stiefelhagen. Tracking Focus of Attention in Meetings. In *Proceedings of IEEE International Conference on Multimodal Interfaces*, page 273, 2002.

[11] Rainer Stiefelhagen, Rachel Bowers, and John Garofolo. Classification of Events, Activities and Relationships - Evaluation and Workshop. `http://www.clear-evaluation.org/`.

[12] K. C. Smith, S. O. Ba, D. Gatica Perez, and J.-M. Odobez. Tracking the Multi-Person Wandering Visual Focus of Attention. In *Proceedings of International Conference on Multimodal Interfaces*, 2006.