Visual Focus of Attention in Dynamic Meeting Scenarios

Michael Voit[1] and Rainer Stiefelhagen[2]

[1] Fraunhofer IITB, Karlsruhe
michael.voit@iitb.fraunhofer.de
[2] Interactive Systems Labs, Universität Karlsruhe (TH)
stiefel@ira.uka.de

Abstract. This paper presents our data collection and first evaluations on estimating visual focus of attention during dynamic meeting scenes. We included moving focus targets and unforeseen interruptions in each meeting, by guiding each meeting along a predefined script of events that three participating actors were instructed to follow. Further meeting attendees were not introduced to upcoming actions or the general purpose of the meeting, hence we were able to capture their natural focus changes within this predefined dynamic scenario with an extensive setup of both visual and acoustical sensors throughout our smart room. We present an adaptive approach to estimate visual focus of attention based on head orientation under these unforeseen conditions and show, that our system achieves an overall recognition rate of 59%, compared to 9% less when choosing the best matching focus target directly from the observed head orientation angles.

1 Introduction

Smart rooms, or smart spaces, proclaim proactive computer services in unobtrusive sensor environments. Knowing at all times, who enters the room, who interacts with whom and where all people reside and look at, allows interfaces to adapt for personal needs and input modalities to relate to context and semantics. Research in this area covers both the fundamental fields of (multiview) visual and acoustical perception, such as face identification [1], gaze recognition [2,3,4], speech detection [5] and speaker localization [6] or audio-visual multi-person tracking and identification [7,8], as well as the combination of all modalities in order to allow higher-level observations and summarizations, as for example in transcribing meetings [9] or analyzing floor control and interaction patterns [9]. One particular cue for modeling (inter-)actions between a group of people or understanding actions and occupations of observed meeting participants and group members, is to understand their visual focus and deduce the respective attentional target they focus on. By means of recognizing objects, colleagues are working on together, or the recognition of a group's joint attention towards a specific speaker during an observed lecture, smart room systems obtain one further cue to modeling a scene's context and individual behavior.

A. Popescu-Belis and R. Stiefelhagen (Eds.): MLMI 2008, LNCS 5237, pp. 1–13, 2008.

To follow eye-gaze and obtain knowledge about one's viewing direction, head orientation usually acts as an approximation to allow non-intrusive sensor setups as applied in our described environment. Due to individual head turning styles, gaze and head orientation tend to differ and a direct interpretation from observed head rotations to a discrete set of focus targets is not always possible as studies and evaluations show [10,11,12]. Measured head rotations are therefore mostly used to describe individually shifted means around predefined focus targets, which, recently in combination with multimodal cues such as presentational slide changes or speech activity [13] both increase recognition rate and allow analysis of group activities or role models during meetings, but still limit the applicational area to a predefined set of non-moving focus targets around a table.

In [14], we extended our system to estimate visual focus of attention from monocular views during recorded meetings [12] to using multi-view head orientation in order to allow for a sensorless work area on the meeting table. Applying the motivation for unrestricted behavior and dynamic scenes to the recorded settings and peoples' focus, we now collected a new dataset, in which a number of scripted events - such as people entering and leaving the room, or phones ringing in the room - were introduced, in order to provoke attention shifts of the meeting participants from their ongoing work. Hence, all meetings contain a varying set of participants and different seating positions as well as the introduction of new objects and moving targets.

2 Dataset

The dataset we recorded consists of 10 meeting videos in total. Each video is approximately 10 min. long and starts with each participant entering the room and finally ends with all persons leaving the room again. For introducing dynamic events and behavior and ensuring the same over all videos, each video consists of three acting participants, that followed a predefined script and a varying number (one or two) of unaware persons, whose attention was to be distracted by different kinds of interruptions, unforeseen persons walking through the room in different trajectories or newly introduced objects.

2.1 Sensor Setup

The sensor setup we recorded with, consisted of 4 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 a coarse view of the whole room, 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 and a restricted sensor range. The camera array was extended with a panoramic view from a fisheye lens camera that was installed on the ceiling (same specifications). For a complete recording of the scenery and its context, 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), allowing

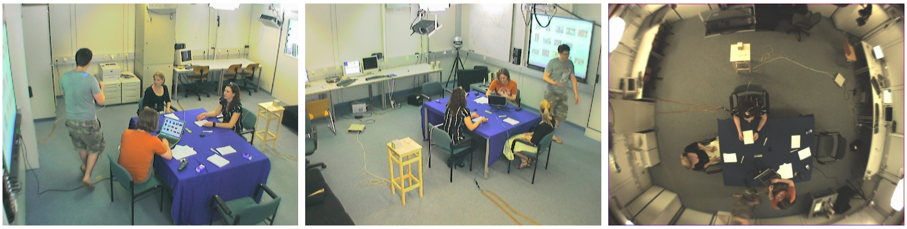

Fig. 1. Example scene of one meeting video. Shown are two out of four camera views from the room's upper corners and the panoramic view, captured from the ceiling. In this scene, interrupting person P04 passes the meeting-table towards the entrance door and walks in between the projection screen and person P00 sitting in front of it, working on his notebook.

for the inclusion of audio source localization, and one table-top microphone for speech recognition and acoustical context modelling.

2.2 Dynamic Meetings

We defined a predefined set of events in a script, that were initiated and followed by all actors in each recorded meeting. The remaining participants were unaware of what was to happen during the recordings, hence, their observed reaction was spontaneous and unplanned. Each meeting consisted of three acting participants and one or two participants that were not aware of the scripted events and the exact purpose of the data collection. To obtain groundtruth information about head orientation and position, one of the unaware persons was wearing a magnetic motion sensor (Flock of Birds, Ascension Technologies) on top of his or her head, calibrated along the room's (hence global) coordinate system. All persons were tagged with respect to their seating position in counter-clockwise order around the meeting table and/or acting role during the meeting: The person sitting at the table's northern edge was named P00, the person to the west P01, the person at the southern edge, always wearing the magnetic sensor was named P02 and the person at the eastern edge P03. The fourth person, called P04 was chosen to interrupt the meeting from time to time, hence entering and leaving the room multiple times and not being bound to one particular seat around the table. The seating positions and roles of all acting persons were changed and rotated during the recordings to prevent repetitive patterns.

In general, the used script followed the particulars given below:

- Person P02 is to be seated beforehand, calibrated along the room's coordinate system. Persons P00, P01 and P03 enter the room successively, meet and greet at the table before sitting down.
- All participants start a discussion about 'Computer Vision' in general and a possible reason for the current meeting.
- The interrupting person P04 enters the room, recognizes the meeting and spontaneously grabs a nearby chair to join it. One of the yet seated participants needs to make room for the additional member, hence his or her

seating position changes - a new person is therefore added around the table, the seating positions disturbed temporarily.
- After a small talk, person P04 stands up, moves his or her chair and leaves the room on either of two possible ways around the table.
- One acting member (P00, P01 or P03) stands up, walks towards the projection screen and starts to give a presentation. Thereby, the presenter gesticulates in front of the screen, changing position in front of it and explains the bullet points listed on the presented slide. All remaining participants were instructed to make notes on the notebooks in front of them on the table and interrupt the presentation with questions and own discussion.
- Person P04 enters the room again, walks towards 'Desktop-Computer 2', sits down and starts working. P04 chooses either way of walking through the room and thus interrupts the presentation for a short amount of time.
- The presenter walks to a nearby placed camera, grabs it, walks back to being in front of the screen and meeting table and introduces the camera before placing it on top of the table for everyone to examine and holding it. The presentation continues.
- The presenter sits down, back onto his or her previous seat. The meeting continues.
- Person P04 starts to play a loud, interrupting sound, initiated from his or her current location in front of 'Desktop-Computer 2'. P04 suddenly stands up, apologizes to the meeting group and rushes to turn the loudspeakers off. P04 then rapidly leaves the room. The meeting continues.
- A cellphone, previously placed inside a cupboard, suddenly starts to ring. Person P04 enters the room, interrupts the meeting by asking if anybody has seen his or her cellphone and follows the ringing sound towards the cupboard. He or she grabs the cellphone, shows it to the meeting participants, turns it off and leaves the room with it. The meeting continues.
- The printer starts to output papers. Person P04 enters the room again, walks to the printer and while shaking and pretending to repair it, P04 complains loudly about a pretended malfunction. P04 grabs papers and leaves the room. The meeting continues.
- All meeting participants, except P02 wearing the sensor, stand up, shake hands and leave the room.

2.3 Annotated Focus Targets

Considering abovely scripted events, a minimum focus target space can be set up with the following:

- Persons P00, P01, P02, P03 and P04 (as available and participating)
- Entrance to the room
- Meeting Table (further, each individual's notebook on top of the table)
- Projection Screen (for during the presentation)
- Camera (that is being introduced during the presenter's talk)

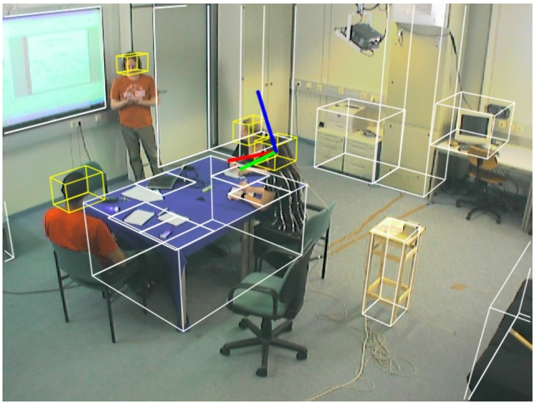

Fig. 2. Camera 1's view of a recorded meeting scene during a short presentation, given by person P00. Person P02, sitting opposite to the presenter, is wearing the magnetic motion sensor to capture her true head orientation, depicted by the red (x), green (y) and blue (z) coordinate axes. All *axis aligned bounding boxes* of focus targets we annotated, visible from this view, are highlighted in white.

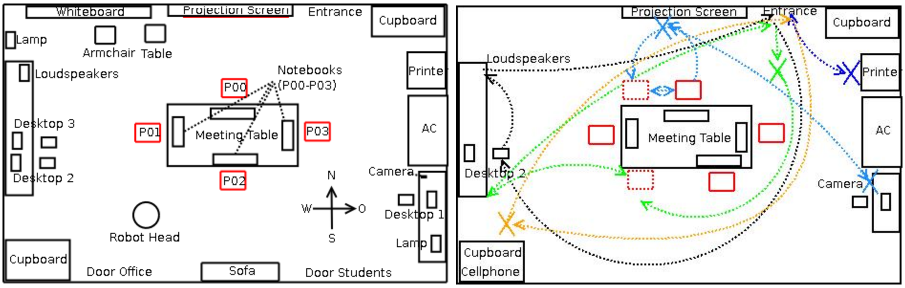

Fig. 3. Left: Overview of annotated focus targets throughout the meeting room. Right: Observed trajectories of all meeting participants. Rather than only gathering meetings with fixed seating positions, participants were advised to walk throughout the entire room, to distract the visual focus of the remaining meeting members.

- Loudspeakers (interrupting the meeting by outputting a disruptive sound)
- Cupboard enclosing the later-on interrupting cellphone
- Printer

For completing the list of potential targets, such as surrounding tables, chairs, working places or cupboards, the room's interieur was completely modeled in 3D: The position and bounding box of each object was measured and head bounding boxes of all meeting participants were annotated for every camera view and every frame recorded. The head bounding boxes were used to triangulate 3D positions of the corresponding heads' centroids in room coordinates and provide basis for future estimation of head orientation for all participants. In addition, the magnetic motion sensor provided groundtruth information about the head orientation for person P02 with approx. 30Hz. Person P02 was always made sure to be one of the unaware meeting participants. All in all, a total of 36 targets were made available for the annotation process, classifying each meeting participant's visual focus. As can be seen in Fig. 3, air conditioning, all chairs and sofas, desktop PCs and cupboards were included as potential targets, too. Even a small robot head we used for different experiments was considered as a potential target due to its position near the meeting table. Fig. 4 depicts a distribution of all annotated objects and persons for how often they were focused throughout the entire dataset by either meeting participant and thus provides a complete overview of all included focus targets. An example of a meeting, with some of the targets being highlighted, can be seen in Fig. 2.

3 Estimating Visual Focus of Attention

3.1 Target Modeling

Due to targets moving, we decided to describe each object and person by its *axis aligned bounding box* in 3D space (see Fig. 2). In order for targets to be able to be focused, their box must overlap or intersect with the respective person's viewing frustum. This viewing cone was defined to open up 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 Baseline: Choosing the Nearest Target

A comparative baseline is established by classifying for target F_i, who seems to be nearest to the observed head orientation. Hence, we distinguished targets by their euclidean distance, computed with their respective representative points P_i and the head pose vector.

3.3 An Adaptive Focus Model

We adopted the described visual focus model presented in [15], which summarizes a linear correlation between the corresponding gaze angle α_G towards the target and the observable head turning angle α_H when focusing on it:

$$\alpha_H = k_\alpha \cdot \alpha_G \tag{1}$$

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 / Persons	P00	P01	P02	P03	P04	Mean
P00	0.06	20.66	24.39	18.22	8.05	15.82
P01	19.54	0.00	14.32	26.38	8.87	13.91
P02	18.79	12.08	0.07	11.59	5.72	9.62
P03	16.47	20.32	13.55	0.00	6.44	12.09
P04	9.37	10.55	6.52	7.72	0.00	7.63
Notebook P00	8.07	1.45	0.29	0.52	0.11	2.09
Notebook P01	0.36	6.37	0.41	0.52	0.18	1.84
Notebook P02	0.14	0.22	8.75	0.67	0.00	2.40
Notebook P03	0.35	0.56	0.82	10.61	0.01	2.70
Notebook P04	0.00	0.00	0.00	0.02	0.09	0.01
Printer	1.48	0.24	0.31	0.09	8.38	1.30
Ceiling	0.23	0.13	0.00	0.08	0.00	0.09
Whiteboard	0.02	0.01	0.10	0.04	0.24	0.06
Robot Head	0.25	0.10	0.11	0.08	0.16	0.13
Desktop 3	0.37	0.03	0.45	0.16	1.88	0.42
Desktop 2	3.52	0.11	0.34	0.19	34.79	4.40
Desktop 1	0.24	0.10	0.11	0.07	0.03	0.12
Meeting-Table	2.80	3.70	2.54	2.01	2.89	2.79
Entrance	0.76	1.11	1.15	0.98	3.89	1.31
Cupboard (North)	0.20	0.08	0.04	0.02	0.21	0.09
Cellphone	0.36	0.52	0.11	0.13	1.87	0.44
Floor	2.79	1.53	0.19	0.99	5.63	1.74
Door Students	0.17	0.02	0.00	0.16	0.20	0.09
Sofa	0.16	0.25	0.00	0.14	0.45	0.17
Lamp (South-East)	0.00	0.01	0.03	0.00	0.00	0.01
Armchair	0.00	0.03	0.01	0.00	0.13	0.02
Chair P04	1.14	0.57	0.04	0.45	1.36	0.60
Lamp (North-West)	0.05	0.00	0.00	0.06	0.00	0.02
Camera	4.10	6.17	7.73	5.04	0.00	5.29
Cupboard (South)	0.91	0.49	0.21	0.31	3.16	0.74
Door Office	0.22	0.15	0.00	0.03	0.03	0.09
Sensor Stool	0.02	0.18	0.00	0.07	0.34	0.10
Projection Screen	6.36	11.83	17.20	12.30	2.52	11.27
Loudspeakers	0.46	0.26	0.18	0.32	1.80	0.45
Small Table	0.09	0.08	0.03	0.00	0.49	0.09
Air Conditioning	0.14	0.06	0.02	0.01	0.05	0.05
Sum	100.00	100.00	100.00	100.00	100.00	100.00

Fig. 4. Distribution of visually focused targets. Each column depicts the focus distribution for the person sitting at the respective position (P00, P01, P02, P03, P04). Each row describes one single focus target.

targets representational points P_i described in 3.2. A measured mapping coefficient k_α could thus be obtained with

$$k_\alpha = \frac{\alpha_H}{\alpha_G} \tag{2}$$

As depicted in Fig. 6 and intuitively assumed, k_α's value does not stay fixed throughout the observations, but rather changes, depending on the dynamics in the observed scene. Its variance can be described as rather high, changing from positive to negative values, adapting to focus changes that happen over time. The presented camera view in Fig. 6 shows the recorded scene during the highlighted time range in k_α's plot. Its values were computed for person P02 (left person in the images), who is positioned at the table's southern edge, wearing the magnetic motion sensor on her head and being one of the unaware meeting participants. The scene shows all participants meeting at the table, greeting each other. P02's focus changes from Person P03 (standing to the right at the table's western edge, with an approximate horizontal gaze angle of +60°) to person P00 (standing right in front of her to the north, at approximately −10° horizontally). While looking at P03, head orientation was measured to intersect with the target, hence the mapping factor of 1.0. During focus change to P00, head pose slowly adapted to the observable gaze change, but stopped at approximately +3°: the mapping coefficient k_α changed to a value of -3 to shift the head vector onto the target's real position at −10°. Fig. 5 shows an exemplary depiction of this process: Depicted are three targets to focus on. In the top row of the image, focus changes quickly between targets 2 and 3. Due to the rapid interaction, head orientation slows down right between the two persons and eye gaze is used to overcome the difference to focus on the particular target. A fixed mapping coefficient would map target 2's position towards target 3 and target 3's position even further away. If only this kind of interaction is given, a static model interprets the shifted position of target 2 successively and classifies correctly for person 2, even though its position seems rather shifted. The bottom row shows a successive focus change between targets 1 and 2. Here, using the same fixed mapping coefficient would map target 2's position towards target 3 again (but not as far as in the top row) and target 1's position towards target 2. A static model trained with these head observations, would assume target 2's gaze angle to lie nearer in front than the static model trained with observations from the top image. Further, the fixed mapping coefficient clearly shows, that head orientation, when focusing on target 1, is clearly mapping into the wrong direction. It needs to adapt to a lower value. The example shows, how the region of interaction and interest influences the necessary transformation value and should be due to adapt.

To better model the dynamics in mapping, we defined a discrete set of possible coefficients (k_α, k_β) for mapping horizontal and vertical head orientation α_h and β_H, and reweighed them by means of the most likely focus target F_i's a-posteriori probability, given the corresponding mapping:

$$\pi_{(k_\alpha,k_\beta),t} = \gamma \cdot \pi_{(k_\alpha,k_\beta),t-1} + (1-\gamma) \cdot \arg\max_{F_i} p(F_i|\Phi_{k_\alpha,k_\beta}) \tag{3}$$

The mapping coefficient pair (k_α, k_β) with highest weight 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.

In general, the a-posteriori likelihood is defined by

$$p(F_i|\Phi_{k_\alpha,k_\beta}) = \frac{p(\Phi_{k_\alpha,k_\beta}|F_i) \cdot P(F_i)}{p(\Phi_{k_\alpha,k_\beta})} \tag{4}$$

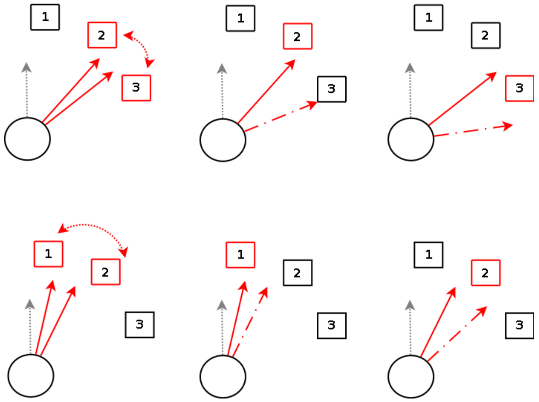

Fig. 5. Top row: Focus changes back and forth between target 2 and target 3 (highlighted red): when focusing target 2 (middle image), a fixed mapping coefficient $0 < k_\alpha < 1$ maps head orientation (solid arrow) onto target 3 (dashed arrow). The gaze angle to target 3's is put even further away than its real position (right image). Bottom row: successively happening focus change between target 1 and target 2: while in the top row, when focusing target 2, head pose tends to cluster towards target 3, here, its corresponding head orientation can be observed between target 1 and 2. Furthermore, target 1's mapped gaze position is shifted towards the second target, instead of backwards to its real origin.

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:

$$P(F_i) = \frac{1}{n} \sum_{t=T-n}^{T-1} (p_t(F_i|(\Phi_t))) \cdot \varphi(\frac{\partial(\angle(\Phi, F_i)))}{\partial t}) \qquad (5)$$

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 $\frac{\partial(\angle(\Phi,F_i))}{\partial t}$ over time shows peaks of negative values and implies a more likely focus change towards that particular target.

3.4 Experimental Evaluation

We reduced the target space to meeting participants, meeting table and projection screen only. This included 88% of all focused objects as annotated in Table 4 and reduces complexity both for these first evaluations and annotations,

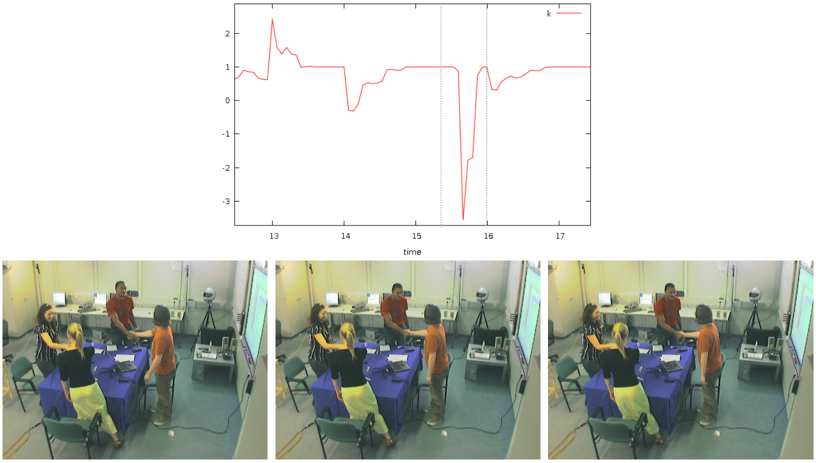

Fig. 6. Respective plot of person P02's mapping coefficient k_α and the corresponding scene in the meeting to the highlighted time window in the plot: Person P02 is standing to the left in the images. Her mapping coefficient k to project the horizontal head orientation to its respective gaze-angle does not stay fixed over time (as visible in the plot). Values of 1 depict, that head orientation points directly to the target and intersected its axis aligned bounding box. The strong variance in the highlighted time window depicts a focus change to person P00 standing in front of her and shows that head orientation not always points behind gaze angles, but depending on the direction focus changes are happening from, might also point ahead of the targets' true positions. Thus, k needs be adapt to a much lower value to map head pose to a lower gaze angle.

Table 1. Recognition rates on the described dataset for person P02. Four different approaches are compared (three direct mappings with a fixed mapping coefficient k_α respectively and our adaptive approach with a variable k_α). The constant mapping factor $k_\alpha = 0.72$ was computed as being the mean mapping coefficient when mapping the observed head orientation to the annotated targets. Head orientation was measured with a magnetic motion sensor.

Mapping	Meeting 1	Meeting 2	Meeting 3	Meeting 4	Mean
Direct Mapping ($k_\alpha = 1$)	54%	49%	**57%**	51%	53.5%
Direct Mapping ($k_\alpha = 0.5$)	53%	49%	43%	55%	49.5%
Direct Mapping ($k_\alpha = 0.72$)	53%	52%	48%	51%	50%
Adaptive Mapping ($\gamma = 0.95$)	**58%**	**59%**	55%	**61%**	**59%**

which are still happening and take a lot of time to define all object and person positions. At the current time writing this paper, all of the videos are annotated for all persons' corresponding visual focus, but only four videos provide the positions and bounding boxes of above mentioned targets. 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.

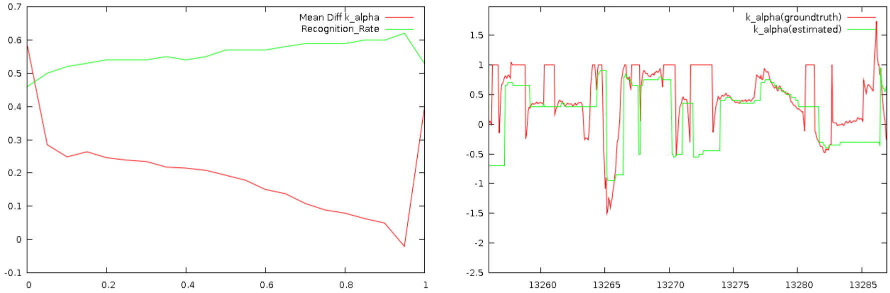

Fig. 7. Left image: Recognition Rate (upper green plot) and mean difference of estimated k_α to groundtruth k_α (lower red plot), with respect to increasing adaption factor γ. A value of $\gamma = 1.0$ describes that the scores π_{k_α} are not adapted at all. In this case, the constant mapping coefficient $k_\alpha = 0.72$ 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.

The low numbers clearly show the difficulty of the task, especially of this particular setting we chose for meetings: Person P00's seat is right in front of the projection screen. Reliantly distinguishing between the two targets is only possible, if either of them is ignored for any reason (possibly due to person P00 sitting a lot nearer and thus overlapping too much of the viewing frustum towards the screen) or context is further taken into account for understanding whether the interest relies on a person sitting in front of the screen or the screen directly behind.

Clearly visible from the results however is, that an adaptive mapping of head pose to the respective focus target increases the recognition rate in almost every case. The only exception shows to be video 3, where a direct interpretation of head pose seems to perform slightly better than a variable (or even fixed with different values) mapping. This might be due to the fact, that this person mostly used its eye gaze to focus on targets - head orientation stayed fixed for most of the time. During the video, our system kept the mapping coefficient relatively constant due to the missing head movements. Especially, rapid focus changes between two targets were more or less completely ignored by our system: Where in the remaining videos slightly head rotations towards the respective targets were observable, here, only gaze was used to switch back and forth - hence, our approach only recognized one target focused during this time; due to the mapped head orientation often the wrong one during these interactions. Further, especially moving targets, for example person P04 passing by in between the meeting table and the projection screen as depicted in Fig. 1, only distracted person P02's visual focus by quick eye movements, instead of letting head orientation follow that respective trajectory. The focus did not change for more than fractions of frames, it was kept on the previous target all the time. However, the a-priori likelihood described in equation 5 includes the derivation of the difference between head pose 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 only. In the example of person P02 during meeting video 3, the interrupting person shows high likelihoods for being focused at when walking only through the room, even though head pose stayed fixed. The focus change here, is enforced to be recognized, even though in this case it does not happen at all.

Hence, general questions that are to be answered in future work (especially as soon as the complete dataset annotation process is finished) 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 distinguishing between every single item. Future work also includes the 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, also should 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.

4 Conclusion

In this paper we presented our work on enhancing the estimation of visual focus of attention in group meetings: We collected a new dataset to include dynamic scenes and moving persons and objects. The dataset contains recordings of meetings from the beginning where all participants enter the room and follows a predefined script of events that three acting meeting members in the recordings were to follow and suprise further attending and unaware participants with. The sensor setup both contains visual recordings from wideangle cameras in the room's upper corners and a panoramic camera on the ceiling as well as audio recordings from T-shaped microphone arrays and one table-top microphone on top of the meeting table. All recordings were annotated for the participants' head bounding boxes, everybodies' visual focus of attention and the complete room's interieur in 3D by means of bounding boxes of each object and allowed target that was annotated. Secondly, we described and evaluated our first system to estimate visual focus of attention for one person on moving targets and achieved an overall mean recognition rate of 59%. We compared our approach to interpreting head orientation as the actual gaze direction and mapping its vector onto the first-best matching, nearest corresponding focus target and our enhancements showed an overall increase in recognition rate by almost 9%. 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. Further, in order to adopt our approach on every meeting participant, independent of his or her movement, research on estimating upper body orientation is due to be done and combined with estimating head orientation and a fully automate multi-person tracking and identification.

References

1. Ekenel, H., Fischer, M., Stiefelhagen, R.: Face recognition in smart rooms. In: Popescu-Belis, A., Renals, S., Bourlard, H. (eds.) MLMI 2007. LNCS, vol. 4892, Springer, Heidelberg (2008)
2. Voit, M., Nickel, K., Stiefelhagen, R.: Head pose estimation in single- and multi-view environments - Results on the CLEAR 2007 benchmarks. In: Stiefelhagen, R., Bowers, R., Fiscus, J.G. (eds.) CLEAR 2007 and RT 2007. LNCS, vol. 4625. Springer, Heidelberg (2007)
3. Head orientation estimation using particle filtering in multiview scenarios. In: Stiefelhagen, R., Bowers, R., Fiscus, J.G. (eds.) CLEAR 2007 and RT 2007. LNCS, vol. 4625. Springer, Heidelberg (2007)
4. Lanz, O., Brunelli, R.: Joint bayesian tracking of head location and pose from low-resolution video. In: Stiefelhagen, R., Bowers, R., Fiscus, J.G. (eds.) CLEAR 2007 and RT 2007. LNCS, vol. 4625. Springer, Heidelberg (2007)
5. Maganti, H.K., Motlicek, P., Gatica-Perez, D.: Unsupervised speech/non-speech detection for automatic speech recognition in meeting rooms. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) (2007)
6. Maganti, H.K., Gatica-Perez, D.: Speaker localization for microphone-array-based asr: the effects of accuracy on overlapping speech. In: Proceedings of IEEE International Conference on Multimodal Interfaces (ICMI) (2006)
7. Bernardin, K., Stiefelhagen, R.: Audio-visual multi-person tracking and identification for smart environments. In: Proceedings of ACM Multimedia (2007)
8. Lanz, O., P.C., Brunelli, R.: An appearance-based particle filter for visual tracking in smart rooms. In: Stiefelhagen, R., Bowers, R., Fiscus, J.G. (eds.) CLEAR 2007 and RT 2007. LNCS, vol. 4625. Springer, Heidelberg (2007)
9. Chen, L., Harper, M., Franklin, A., Rose, R.T., Kimbara, I.: A multimodal analysis of floor control in meetings. In: Renals, S., Bengio, S., Fiscus, J.G. (eds.) MLMI 2006. LNCS, vol. 4299. Springer, Heidelberg (2006)
10. Freedman, E.G., Sparks, D.L.: Eye-head coordination during head-unrestrained gaze shifts in rhesus monkeys. Journal of Neurophysiology 77, 2328 (1997)
11. Wang, X., Jin, J.: A quantitative analysis for decomposing visual signal of the gaze displacement. In: Proceedings of the Pan-Sydney area workshop on Visual information processing, p. 153 (2001)
12. Stiefelhagen, R.: Tracking focus of attention in meetings. In: Proceedings of IEEE International Conference on Multimodal Interfaces (ICMI), p. 273 (2002)
13. Ba, S., Odobez, J.: Multi-party focus of attention recognition in meetings from head pose and multimodal contextual cues. In: International Conference on on Acoustics, Speech, and Signal Processing (ICASSP) (2008)
14. Voit, M., Stiefelhagen, R.: Tracking head pose and focus of attention with multiple far-field cameras. In: International Conference on Multimodal Interfaces (ICMI) (2006)
15. Ba, S., Odobez, J.: 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 (ICME) (2007)