Motion based situation recognition in group meetings

Julia Moehrmann[a], Xin Wang[b] and Gunther Heidemann[a]

[a]Intelligent Systems Group, Universität Stuttgart, Germany;
[b]High Performance Computing Center, Universität Stuttgart, Germany;

ABSTRACT

We present an unobtrusive vision based system for the recognition of situations in group meetings. The system uses a three-stage architecture, consisting of one video processing stage and two classification stages. The video processing stage detects motion in the videos and extracts up to 12 features from this data. The classification stage uses Hidden Markov Models to first identify the activity of every participant in the meeting and afterwards recognize the situation as a whole. The feature extraction uses position information of both hands and the face to extract motion features like speed, acceleration and motion frequency, as well as distance based features. We investigate the discriminative ability of these features and their applicability to the task of interaction recognition. A two-stage Hidden Markov Model classifier is applied to perform the recognition task. The developed system classifies the situation in 94% of all frames in our video test set correctly, where 3% of the test data is misclassified due to contradictory behavior of the participants. The results show that unimodal data can be sufficient to recognize complex situations.

Keywords: Group meeting recognition, situation recognition, Hidden Markov Model, Motion features

1. INTRODUCTION

The recognition of specific situations like group meetings has been a topic of research for several years. Group meetings are of interest for the automatic creation of meeting transcripts, behavioral analysis in a restricted environment, or simply for organizational purposes. A lot of research concerning meeting recognitions has been performed in smart meeting rooms which are equipped with a variety of physical sensors, for example, cameras, microphones, or sensors for temperature, light, and position. In contrast to most of the mentioned sensors, cameras, i.e. vision based sensors, have the ability to replace a variety of physical sensors by providing the functionality of several physical sensors in one system. For example, vision based sensors may detect whether the light or projector is turned on. Additionally, vision based sensors may provide information about the presence of persons in the room and their actions.

The behavior of persons inside group meetings has been studied in the field of sociology for several decades. These studies led to the development of the "turn taking" mechanism in discussions which describes interactions between participants.[1,2] In detail, participants adopt complementary roles inside such meetings. Therefore, the recognition of the situation as a whole cannot be performed by the recognition of individual activities alone. Instead, the sum of all individual actions is necessary to identify the situation.

The system proposed in this work uses cameras to capture video data of one or more persons. Since cameras are already present in most meeting rooms, no special setup is necessary. Additionally, there is no need to equip the meeting room with a series of physical sensors.

Based on features extracted from the video data, three different individual activities are recognized: *speaking*, *writing* and *idle/listening*. Individual activities indicate the status of one participant. The system is capable of distinguishing the group situations *discussion*, *presentation* and *silent work*, i.e. persons taking notes. Based on the work by Zhang et al.[3] a two-layer framework is employed in this work, since it directly integrates the sociological model of individual behavior and group situations.

The following section gives an overview on the related work. Section 3 describes the developed interaction and situation recognition and the features investigated for this purpose. Section 4 describes the experimental setup and the evaluation of our system. Section 5 concludes this work with aspects to be discussed in the future.

Further author information: Send correspondence to J. Moehrmann
E-mail: Julia.Moehrmann@vis.uni-stuttgart.de, Telephone: +49 (711) 7816-438

2. RELATED WORK

A variety of techniques for the recognition of gestures and actions have been developed. Head pose estimation is an important factor for the recognition of interactions inside groups because the viewing direction can be derived from this information.[4] Similarly to using the head pose, the position of the head can be used to track persons and recognize their actions in a meeting.[5] Zobl et al.[6] discuss an action recognition framework based on global motion features like the center of mass, and motion wideness derived from difference images. Mitra et al.[7] also provide an extensive survey on gesture recognition.

Speech recognition has been used by Yu et al.[8] to automate the creation of meeting transcripts. Morgan et al.[9] discuss the challenges of creating automated transcripts from informal meetings based on speech recognition. Dielmann et al.[10] employed a statistical approach using dynamic Bayesian networks for segmenting multiparty meetings based on audio information. Waibel et al.[11,12] used visual cues in addition to speech recognition to create a meeting browser. A multimodal approach, which uses speech, gesture, handwriting and person identification, was developed by Bett et al.[13] Recognizing motion patterns in a smart room scenario has been studied in the course of the Multi-Modal Meeting Manager project,[14] which tries to analyze meeting situations and create a summary based on audio, video and textual data. The recognition of situations is based on audio-visual data, as well as multimedia information, like mouse movement or PC-based presentation. In the course of this project Zhang et al.[3] developed a framework based on a two-layer Hidden Markov Model (HMM) for the recognition of situations inside group meetings. The framework recognizes individual activities in the first layer and passes the results to the second layer along with group features in order to model the importance of the interaction between the participants discussed in sociological studies.[2] The integration of multimodal sensor data using dynamic Bayesian networks was discussed by Choudhury et al.[15] Shivappa et al.[16] also discussed the integration of multimodal data in an intelligent meeting room. Audio-visual data is used for person tracking, head pose estimation and speaker identification. Each recognition task serves as input data for another recognition task. By combining the individual tasks they support each other and help to improve the overall goal, which is speech recognition.

3. SITUATION RECOGNITION

The recognition of situations inside group meetings is performed as a two-stage process. In the first classification stage individual activities are recognized, i.e. whether a participant is *speaking*, *writing* or *idle/listening*. Since our recognition system uses image data only, the speaker is recognized on the basis of his or her motion. The second classification stage combines the resulting individual activities in order to recognize whether a discussion or presentation is taking place or whether persons are working independently of each other.

3.1 Interaction recognition

Standard face detection[18] is employed to identify the number of participants and initialize a skin color model for the purpose of tracking head and hand motion. The motion data is used as the basis for recognizing individual activities. Since our system concentrates on the recognition of situations inside group meetings, actions like coming in, sitting down or leaving are not identified. However, an identification of these actions, for example, based on the work by Charif et al.[5] could be used to extend our system to a more general situation recognition. The skin color is modeled by a gaussian distribution in hue-saturation space. Due to our test scenario containing static lighting conditions only, the use of mixture models for modeling skin color did not increase the segmentation performance significantly. The results of the skin color based tracking are shown in Figure 1. A skin color model is initialized for each person as soon as a frontal face is detected. The face is usually detected when a person is sitting down in front of a camera. As a result of the face detection, the initial position of the face is known. This information is used to define a search area for the face in the consecutive frame and thereby improve the tracking by eliminating regions which were detected by mistake, due to similar color information. The same approach is adopted for the tracking of the hands. Initially hands are identified as one or two skin color regions in a search area below the face. If only one skin color region is detected, we assume that the hands are folded and adopt the same position for both hands. If hands cannot be detected it is very probable that the hands are occluded by a table or the persons arms. If a previous position of the hands exists, this position is maintained, otherwise no position can be extracted.

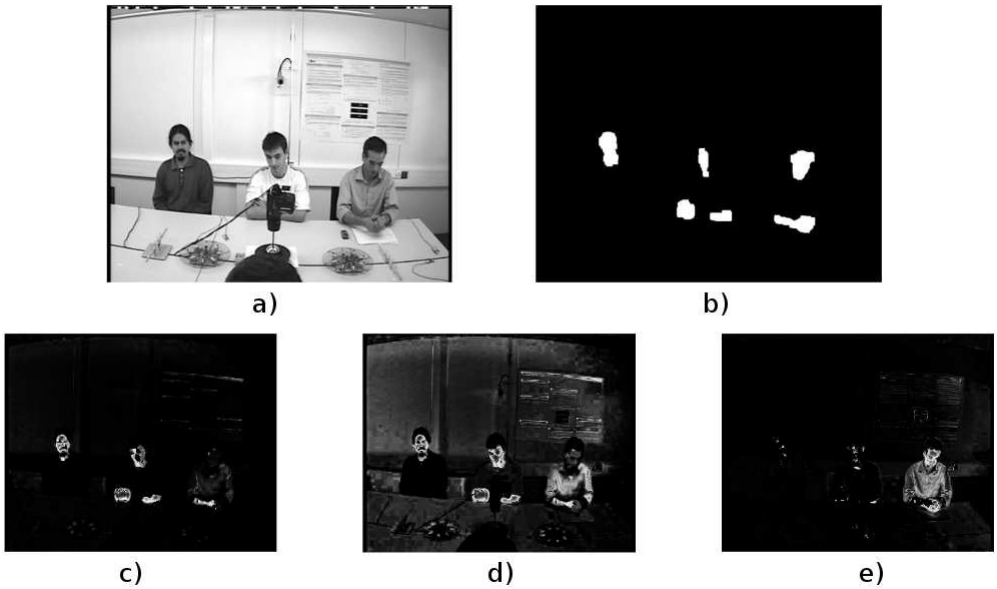

Figure 1. Results of the skin color segmentation. a) shows the original image taken from the PETS-ICVS 2003 Smart meeting room data,[17] b) shows the result of the combined and thresholded segmentations of all three skin color models. The segmentation results for the individual color models are shown in figure c) to e) with c) showing the segmentation result of the skin color model belonging to the person on the left, d) belonging to the person in the middle and e) belonging to the person on the right.

3.1.1 Motion feature extraction

As mentioned previously, we extract the motion trajectories by identifying the position information of the head and both hands. Figure 2 shows a plot of the positions of head and both hands, split into x- and y coordinates. It is obvious that the motion patterns do not provide any hints about the current activity of the person. However, one can see that the x- and y-dimension do not significantly differ from each other. Additionally, one can see that the motion of the left hand is highly correlated to the motion of the right hand.

Although the position features are not suitable for the recognition of individual activities, they form the basis for the extraction of other features. We extracted velocity and acceleration of head and both hands, as well as the motion perimeter and the autocorrelation of hand motion to investigate suitable features. The motion perimeter mp_t for one participant in frame t is given as

$$mp_t(\boldsymbol{h}(t), \boldsymbol{l}(t), \boldsymbol{r}(t)) = d(\boldsymbol{l}(t), \boldsymbol{r}(t)) + d(\boldsymbol{h}(t), \boldsymbol{l}(t)) + d(\boldsymbol{h}(t), r(t)), \tag{1}$$

with $\boldsymbol{h}$, $\boldsymbol{l}$ and $\boldsymbol{r}$ being the position vector of head, right- and left hand respectively. T is the sum of all frames t. As distance $d(\boldsymbol{v}, \boldsymbol{w})$ the Manhattan distance is used which is defined as

$$d(\boldsymbol{v}, \boldsymbol{w}) = \sum_{i=1}^{n} |\boldsymbol{v}_i - \boldsymbol{w}_i|, \tag{2}$$

with n being the number of dimensions. The motion perimeter depends on the physique of the individual person. In order to provide a person independent metric, the motion perimeter is normalized for every person. The normalized motion perimeter $\widetilde{mp}_t$ is given by

$$\widetilde{mp}_t = \frac{mp_t}{max_{t' \in T}(mp_{t'})}. \tag{3}$$

The calculation of the motion perimeter does not depend on other time steps, therefore it is not sufficient for modeling temporal relations. In signal analysis tasks the autocorrelation is used to detect repeating patterns

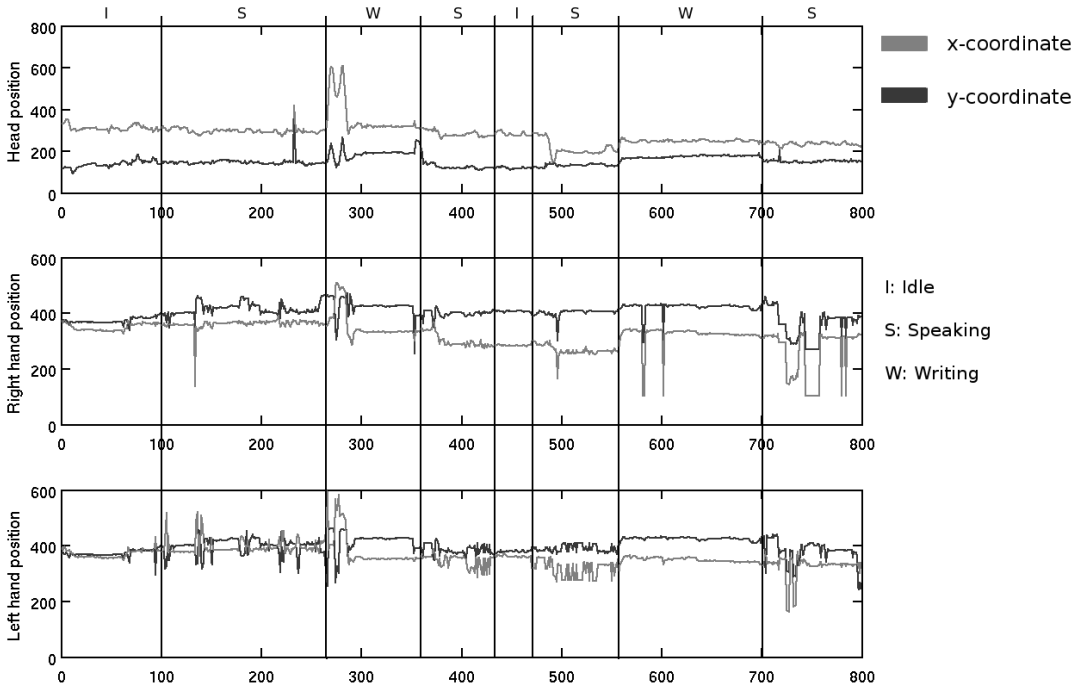

Figure 2. These plots show the absolute position of the head and both hands of one participant for a duration of 800 frames. The actual activity of the participant as identified on the video data is given at the top of the plot. The position information is split into x- and y-coordinate data. $\boldsymbol{I}$, $\boldsymbol{S}$ and $\boldsymbol{W}$ denote sections of idle, speaking and writing respectively.

in a sequence. Similarly, the autocorrelation function can be used to detect recurring hand motions. As can be seen in the top row of Figure 3, the motion perimeter intuitively provides a better description of the three individual activities than the absolute positions. Based on this observation the distance of both hands, given as $d(\boldsymbol{l}(t), \boldsymbol{r}(t))$, was chosen as the basis for the autocorrelation calculation. Since our goal is not the detection of a hidden periodic signal, but simply the identification of recurring hand motion, we set the time-lag to a constant value of 50 frames, which was empirically found to be suitable. Sliding windows of 40 frames ($\approx$ 5 seconds) were used to calculate the autocorrelation.

3.1.2 Additional features

Since motion features capture only a specific aspect of activities, head pose detection is used to complement our set of features. Recognizing head poses robustly is highly dependent on the angle and distance of the person to the camera. If one camera is used to capture several persons, some angles might lead to wrong conclusions. Detecting head poses in order to recognize interactions between persons requires additional knowledge about the spatial relationship among them. Instead a reduced head pose estimation is used which recognizes whether a person's head is upraised or lowered. This recognition is performed using a hair color model, similar to the skin color model used for tracking the face position. The hair color model is initialized on the region above the detected face in the initialization step. The detection of possible hair and skin regions, combined with the knowledge of the head position, allows the calculation of the ratio between skin region size and hair region size. The lower row in Figure 3 shows that this feature is suitable for the recognition of the writing activity. Except for minor errors it is possible to deduct the writing activity from this feature exclusively.

Based on the idea by Zhang et al,[3] we also introduce a group feature. This feature is used in the second classification stage and is valid for all participants. In this work, the group feature is employed to identify whether

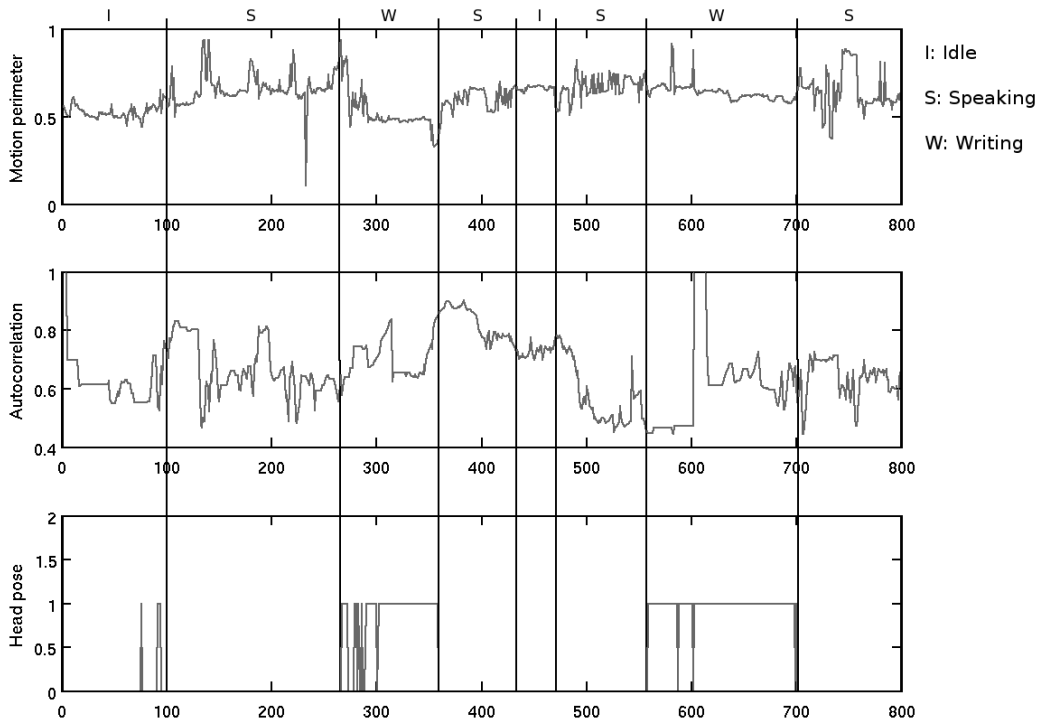

Figure 3. The plot shows the feature sequences for the motion perimeter, autocorrelation and head pose, corresponding to the position data given in Figure 2. Head pose of value 1 indicates a lowered head.

a person is presenting in front of the group. The segmentation is performed using an approximate median background subtraction.[19] The activity of the person standing in front of the group is not further investigated, since it is unlikely to reveal additional information about the overall situation.

3.1.3 Feature selection

In order to select suitable features for the first classification stage, we investigated the distribution of the extracted features in feature space. Figure 4 a) shows a series of two-dimensional scatterplots. Each scatterplot shows the distribution of the extracted features in a two-dimensional feature space. None of the feature combinations in Figure 4 a), corresponding to velocity and acceleration, exhibit even a moderate separating ability. The same applies for feature combinations of velocity and acceleration with all other features, which are not given in this figure. Figure 4 b) and c) show the distribution of features in two-dimensional feature space spanned by motion perimeter and autocorrelation. Although there is no perfect separation, both activities clearly form clusters in Figure 4 c). The *writing* activity is neglected in this plot since it can be recognized based on head pose. From the set of features, i.e. head and hand positions, velocity, acceleration, motion perimeter, the autocorrelation feature and the head pose, only the last three are relevant for the separation of the individual activities. The remaining nine features do not provide a description of the original head and hands motion which could be used to distinguish the individual activities. These features are therefore not further considered in this work.

3.2 Classification

The classification is performed as a two-stage classification process. The complete process from the feature extraction to the recognition of the situation is displayed in Figure 6. First the head and both hands are detected and features are extracted from the position data. Afterwards, the individual activity is recognized

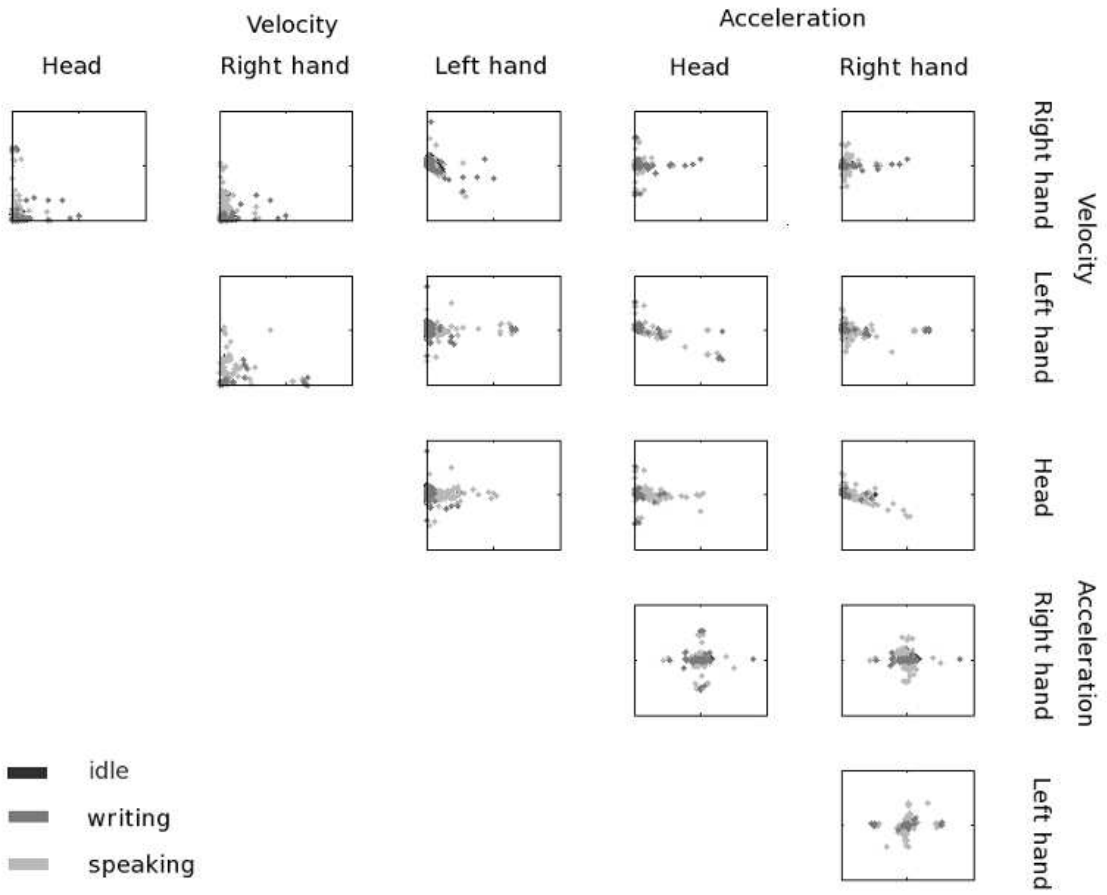

Figure 4. This plot shows distributions of features corresponding to the three individual activities in two-dimensional feature space spanned by velocity and acceleration. Labels given at the top of the figure identify the x-axes, labels given on the right identify the y-axes. No combination of velocity and acceleration features accomplish to separate the individual activities in feature space. Points identifying *idle* are vastly occluded by the other activities and therefore barely perceptible even at larger resolutions.

for every participant. The HMM, which is used to recognize the individual activity, is initially trained on the data of several persons. This is done to provide a person independent recognition system and to reduce the training effort. The HMM used in this stage is therefore exactly the same for every participant. The individual recognition results are passed on, along with the group feature, to the second HMM which recognizes the situation as a whole. The second recognition stage was trained to distinguish the situations *discussion*, *presentation* and *silent work*. A *discussion* is specified as at least one person speaking and the others listening, *presentation* as one person standing in front of the group and the others listening, and *silent work* as the majority of the participants taking notes.

4. EVALUATION

For the purpose of evaluating our system, we defined several group meeting scenarios with predefined schedules and four different persons participating in every meeting. One camera per participant was used. All videos were captured at 7,5 frames per second and a total of 30 minutes of group meetings was captured.
Seven minutes of video data (12800 frames) consisting of four video streams (one per participant), were used to train the first stage classifier, i.e. the recognition of individual activities. The second stage classifier was trained on the resulting individual activities referring to the seven minutes video data.

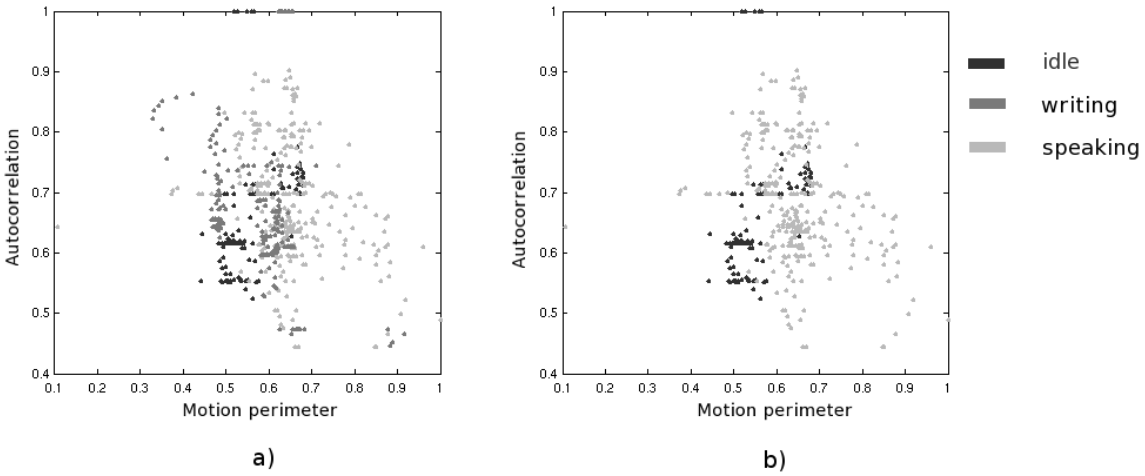

Figure 5. a) shows a scatterplot of all three individual activities in the two-dimensional feature space spanned by motion perimeter and autocorrelation, b) shows the same plot without the *writing* activity, leaving *idle* and *speaking*, since writing can already be recognized based on head pose. Although there is no perfect separation of activities in this feature space, it is clearly visible that features corresponding to *idle* form a cluster.

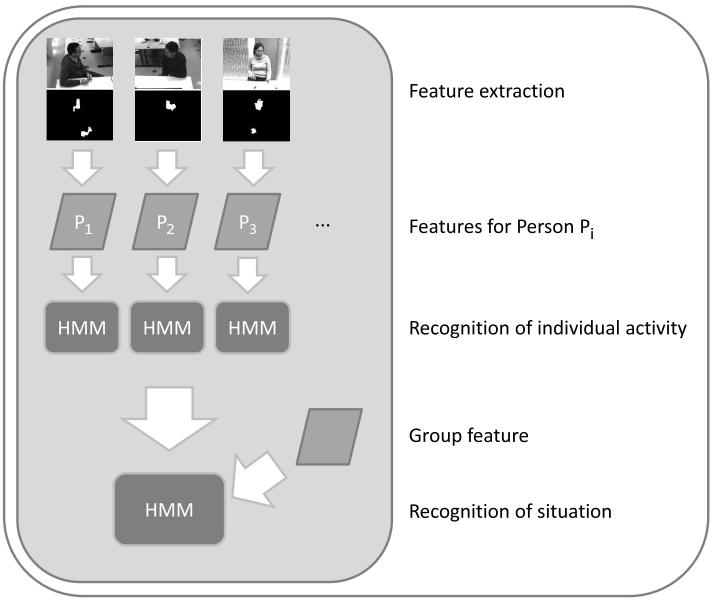

Figure 6. The process implemented by our recognition system.

The system was tested on the remaining 23 minutes of group meeting video data, i.e. 92 minutes (10400 frames) of video data for all participants. Since the recognition results are evaluated for situations only, four frames are processed in one time step for all participants. In the following, we will refer to time step to indicate the same frame in all four video streams.

The results of the test are summarized in Table 1. The situation with the best recognition rate is *silent work* with a perfect recognition rate and no false negatives. *Presentation* is recognized correctly in 60,8% of all time steps. Taking a closer look at the false negatives for this situation reveals that they were mostly recognized as discussions. This fact is very interesting, since the video shows one person standing in front of the group while other participants are talking to each other. Whether this situation should correctly be labeled as a presentation or discussion is controversial. Ambiguous situations of this kind make up 3% of the test data set. The misclassifications concerning *discussions* are caused by head poses which were falsely identified as lowered

for one or more participants. *Discussion* was therefore recognized as *silent work*.

Table 1. Evaluation results on a test set consisting of 41600 frames (10400 per participant). *Presentations* exhibit the highest error rate, due to contradictory behavior of participants. The overall correct classification rate is approximately 94%.

Situation	#Total time steps per situation	#Misclassified time steps	error rate
Presentation	1151	451	39,2%
Discussion	8667	266	3%
Silent work	582	0	0%
Overall	10400	647	6,2%

Considering the features used for recognizing the situations, the classification results are very promising. The overall correct classification rate is approximately 94%. Although the skin color based segmentation and tracking of both hands failed in several cases, this does not severly decrease the recognition results.

5. CONCLUSION

We presented a system for recognizing situations within group meeting. A set of features were extracted from position information of both hands and the head of each participant. We selected a suitable subset of features for the recognition of individual activities, i.e. the activity of one participant. Three features, motion perimeter, autocorrelation of the distance between both hands and the head pose are able to separate the individual activities *speaking*, *writing* and *idle/listening* in feature space. The system is implemented as a two-stage classification process using Hidden-Markov Models. The first stage recognizes individual activities, whereas the second stage recognizes the overall situation based on the individual activities and an additional group feature which identifies whether a person is presenting in front of the group. The system is capable of distinguishing the situations *discussion*, *presentation* and *silent work* with an overall correct classification rate of about 94%.
Although the extracted features are quite simple and are based on position information and head pose only, the classification of the situation is robust, due to the two-stage classification process. The system can easily be extended to the recognition of additional situations by increasing the number of features for motion description or by adding additional information.
Future research will include the recognition of group meetings in contrast to encounters of several group meetings. Distinguishing a group meeting with 10 to 15 participants from several smaller independent groups at the same table presents a challenge. This recognition will be performed by an analysis of postures and gaze directions, i.e. the identification of several speakers and group centers. Using social signals as investigated by Vinciarelli et al[20] might also help to significantly improve this aspect. Another research topic will be the recognition of situations inside group meetings under varying lighting conditions, as is often the case in real world scenarios.
Since our system is based on visual information only, we think that it can be extended to more general behavioral analysis tasks outside of group meetings, for example, in surveillance scenarios.

ACKNOWLEDGMENTS

This work was developed within the Nexus project (collaborative research center/SFB 627), which is supported by the German Research Foundation (DFG).

REFERENCES

[1] Sacks, H., Schegloff, E. A., and Jefferson, G., "A simplest systematics for the organization of turn-taking for conversation," *Language* **50**(4), 696–735 (1974).

[2] McGrath, J. E., [*Groups, interaction and performance*], Prentice-Hall (1984).
[3] Zhang, D., Gatica-Perez, D., Bengio, S., and McCowan, I., "Modeling individual and group actions in meetings with layered HMMs," *IEEE Transactions on Multimedia* **8**(3), 509–520 (2006).
[4] Murphy-Chutorian, E. and Trivedi, M. M., "Head pose estimation in computer vision: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence* **31**(4), 607–626 (2009).
[5] Charif, N. and McKenna, J., "Tracking the activity of participants in a meeting," *Machine Vision Applications* **17**(2), 83–93 (2006).
[6] Zobl, M., Wallhoff, F., and Rigoll, G., "Action recognition in meeting scenarios using global motion features," in [*Action Recognition in Meeting Scenarios Using Global Motion Features*], **1**, 299–308 (2003).
[7] Mitra, S. and Acharya, T., "Gesture recognition: A survey," *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews* **37**(3), 311–324 (2007).
[8] Yu, H., Clark, C., Malkin, R., and Waibel, A., "Experiments in automatic meeting transcription using JRTK," in [*IEEE International Conference on Acoustics, Speech and Signal Processing*], **2**, 921–924 (1998).
[9] Morgan, N., Baron, D., Edwards, J., Ellis, D., Gelbart, D., Janin, A., Pfau, T., Shriberg, E., and Stolcke, A., "The meeting project at ICSI," in [*Human language technology research (HLT '01)*], 1–7 (2001).
[10] Dielmann, A. and Renals, S., "Dynamic bayesian networks for meeting structuring," in [*IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '04)*], **5**, 629–632 (2004).
[11] Waibel, A., Bett, M., Finke, M., and Stiefelhagen, R., "Meeting browser: Tracking and summarizing meetings," (1998).
[12] Waibel, A., Yu, H., Westphal, M., Soltau, H., Schultz, T., Schaaf, T., Pan, Y., Metze, F., and Bett, M., "Advances in meeting recognition," in [*Human language technology (HLT '01)*], 1–3 (2001).
[13] Bett, M., Gross, R., Yu, H., Zhu, X., Pan, Y., Yang, J., and Waibel, A., "Multimodal meeting tracker," in [*Conference on Content-based Multimedia Information Access (RIAO2000)*], (2000).
[14] "EU Project M4. Multi-Modal Meeting Manager." http://www.dcs.shef.ac.uk/spandh/projects/m4/.
[15] Choudhury, T., Rehg, J., Pavlovic, V., and Pentland, A., "Boosting and structure learning in dynamic bayesian networks for audio-visual speaker detection," in [*International Conference on Pattern Recognition*], **3**, 789–794 (2002).
[16] Shivappa, S., Trivedi, M., and Rao, B., "Hierarchical audio-visual cue integration framework for activity analysis in intelligent meeting rooms," in [*Computer Vision and Pattern recognition Workshop*], 107–114 (2009).
[17] Ferryman (ed.), J., "Fourth IEEE international workshop on performance of tracking and surveillance (PETS-ICVS). Graz, Austria," (2003).
[18] Viola, P. and Jones, M. J., "Robust real-time face detection," *International Journal of Computer Vision* **57**(2), 137–154 (2004).
[19] McFarlane, N. J. B. and Schofield, C. P., "Segmentation and tracking of piglets in images," *Machine Vision and Applications* **8**(3), 187–193 (1995).
[20] Vinciarelli, A., Pantic, M., and Bourlard, H., "Social signal processing: Survey of an emerging domain," *Image and Vision Computing* **27**(12), 1743 – 1759 (2009).