Activity Recognition and Room-Level Tracking in an Office Environment

Christian Wojek, Kai Nickel, Rainer Stiefelhagen
{cwojek, nickel, stiefel}@ira.uka.de
Interactive System Laboratories
Universität Karlsruhe (TH)
Germany

***Abstract*— We present an approach for multi-person activity recognition in an office environment with simultaneous tracking of users on the room-level. Audio as well as video features, gathered from a simple setup, are used to employ a multi-level Hidden Markov Model (HMM) framework. Evaluation on unconstrained real world data recorded on several days in five offices with one camera and one microphone per room is presented for activity recognition. We track the users by a distributed camera network which has to cope with blind gaps between different camera views. For location estimation, we apply a Bayesian filter on top of the activity recognition results. Results on a dedicated tracking sequence of one hour length show the algorithm's performance.**

I. Introduction

Activity Recognition has been a very active area of research in the past few years. The scope of work reaches from pure surveillance oriented tasks to applications in smart rooms. Those allow to analyze human interactions and to facilitate them with further services. For instance, it can be inferred how available users are and how convenient it is for them to be interrupted. Danninger *et al.* [1] for example show a scenario where phone calls and messages are routed according to the user's availability. Moreover, user intentions can be understood and interactions within groups can be interpreted.

A. Previous work

Previous work in the area of activity recognition is often embedded in a surveillance task. Oliver *et al.* [2] for example present an approach where coupled Hidden Markov Models are used to recognize human interactions in synthetic pedestrian scenes. Similar research was conducted by Stauffer and Grimson [3] who propose a real-time tracking algorithm and learn a hierarchy of classifiers to recognize usual trajectories; based on those unusual events are detected. Moreover, Brand and Kettnacker [4] propose to apply entropy minimization in order to obtain the parameters of Hidden Markov Models. They prove their algorithm to work well on a surveillance task where they monitor a junction for unusual events. Furthermore, a simple one person office activity detection based on video features under constant conditions is presented. In [5], Ivanov and Bobick augment context free-grammars with probabilities yielding a stochastic parsing framework to recognize visual activities where events on the lower level are recognized by probabilistic methods. This real-time system masters a gesture recognition task as well as a parking lot surveillance domain. Though, the most similar to our work is presented by Oliver *et al.* [6] who exploit multi-layer HMMs for a one person office activity recognition system. They not only use audio and video information but also computer interaction to detect what is happening in front of the user's computer. Their work differs from ours by the fact that their system works under stable lighting conditions where color based features are feasible. Moreover, only a small part of the room is captured and hence face recognizers can be applied to determine the number of people easily. Other important contributions are made by Zhang and McCowan *et al.* [7], [8] who address the problem of meeting understanding with a multi-layered HMM approach. In contrast to our setup, each user is equipped with a lapel microphone and additionally a microphone array is in use. The focus on the work is to explore several ways of coupling video and audio streams with different model setups where data is gathered from scripted meetings.

For the tracking subproblem most similar work has been done based on RFID sensors or similar tags. Wilson and Atkeson [9] for example equipped a home for elderly people with binary sensors and tracked them based on the readings from sensors such as pressure mats. Moreover, occupants were attached a motion sensor and activity recognition yielded whether they were moving or not. Finally, Schulz *et al.* [10] attach infrared badges to persons and use laser range-finders to simultaneously identify and track persons in an indoor environment.

B. System overview

The system presented here aims to understand situations that take place in a multi-person office environment. It employs features (section II) that are captured from audio and video sources and applies a multi-level HMM framework (section III) to recognize activities. Using video, events such as A PERSON AT DESK A or PERSON LEAVING OFFICE B are recognized. Using audio, speech and silence are classified. The higher level fuses both modalities to a more semantic description of what is happening: MEETINGS, DISCUSSIONS, PAPERWORK, PHONE CALLS or NOBODY PRESENT are the office situations that are to be distinguished by our system. Finally, a room-level tracking based on Bayesian filtering (section IV) yields the identities of the persons who are involved. For the evaluation (section V) it is essential to note that we recorded quite challenging, unconstrained

(a) Office A (b) Office B (c) Office C (d) Office D (e) Lab

Fig. 2. Camera views of the 5 monitored rooms; For the office rooms note that big windows cause varying lighting conditions.

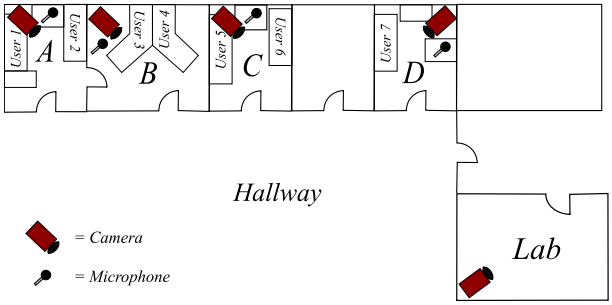


Fig. 1. Audio and video sensors in five monitored office rooms. Except for the lab, each room is equipped with one wide-angle cam and one omni-directional microphone.

Fig. 3. Left: Foreground segmentation of example input image 2(b) (the higher the distance to the background model the more intense is the green channel); Right: Optical flow extraction for a person entering an office (displacement is magnified for better demonstration)

real world data for several days, where illumination varied from artificial light to diffuse daylight. Figure 1 depicts the domain's layout and the distribution of sensors. Note that the hallway is not in the view of any camera.

II. Features

The selection of appropriate features is crucial to any kind of classification or recognition algorithm. For the audio signal we have to deal with various noise sources such as moving chairs or fan noise caused by computers. Additionally, the quality of the signal is dependent on the location of the speaker. There are also two offices which are connected through a door which is open for most of the time, such that a microphone possibly captures speech from the neighboring office in case the signal is loud enough.
Video processing is complicated by the varying lighting conditions. This means, that any kind of illumination dependent cue such as histogram backprojection to identify skin color is inappropriate for this kind of problem. Moreover, persons are perceived from multiple views. Depending on their position we get either frontal, side or even back views of their head, so that face detectors can hardly be employed to decide about the number of persons in the room.

A. Audio features

As we only want to separate spoken conversation from ambient noise, we use the well known audio features *signal power* and *zero crossing rate*. In addition, we employ an autocorrelation based method by Cheveignè and Kawahara [11] to extract *pitch*. As it was chosen to process the video stream with a frequency of 1Hz, the audio features had to be downsampled. This was accomplished by taking the mean and variance of each feature described above over one second. The resulting six dimensional vector o_t^a serves as input vector to the audio classification HMMs.

B. Video features

Since color dependent segmentation algorithms are prone to varying lighting conditions which result from the large windows, it was decided to use motion as main indicator for foreground regions. In order to do so, the color image is converted to grey scale and the segmentation algorithm described next is applied.

1) Segmentation: Due to the illumination changes a fixed background model can not be used. Furthermore, purely taking difference images of successive frames is not desireable, as this is not capable of detecting little motion well. Consequently, the background model $bg(i)$ needs to be adapted for each pixel location $i \in \mathbb{N}^2$ according to the learning rule:

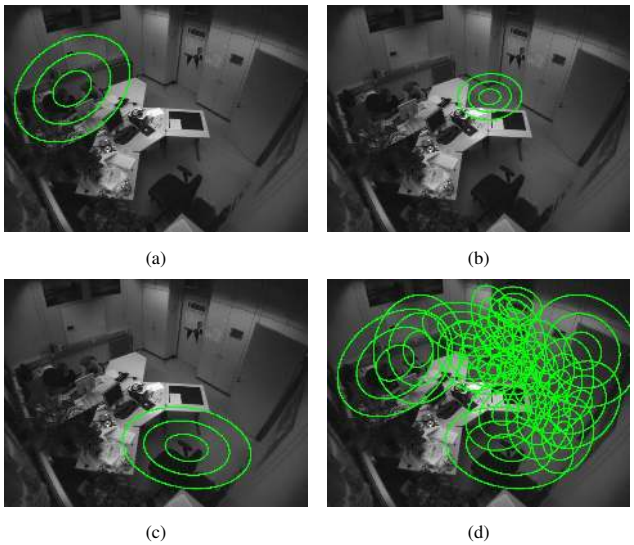
$$bg_t(i) = \alpha \cdot bg_{t-1}(i) + (1-\alpha) \cdot p_t(i)$$

Here bg_t denotes the background model at time t, p_t the current image and α the learning rate. To decide whether a pixel belongs to the foreground fg a threshold m on the difference image between background and input image is used:

$$fg_t(i) = \begin{cases} 0 & \text{if } |p_t(i) - bg_{t-1}(i)| \leq m \\ |p_t(i) - bg_{t-1}(i)| & \text{if } |p_t(i) - bg_{t-1}(i)| > m \end{cases}$$

All pixels i with $fg_t(i) > 0$ belong to foreground regions. Figure 3 shows an example segmentation.

2) Optical Flow: Good features to track are located using the criterion by Shi and Tomasi [12]. The feature's displacement then can be obtained with the feature tracker by Lucas and Kanade [13] of which a pyramidical implementation by Bouguet is used. As sometimes outliers are returned, the



(a) (b) (c) (d)

Fig. 4. Images (a) - (c) illustrate Gaussian mixture components (within 3 standard deviations) obtained for areas where users often sit, image (d) shows all Gaussians resulting from data driven clustering

median values of x and y displacement will be used for further processing. Figure 3 shows an example where the obtained optical flow is depicted with arrows.

C. Local feature model

In order to describe an action, local features are to be considered, whereas other parts of the image are irrelevant. Moreover, HMMs can not be applied reasonably with an entire image as input vector and so dimensionality has to be reduced. Hence, a local description needs to be found and meaningful image areas have to be detected. We consider those to be the k components of a mixture of Gaussians learned from the 2D location of foreground pixels. To learn the distribution's hidden parameters $\Theta = \{\Sigma_0, \mu_0, w_0, \ldots, \Sigma_{k-1}, \mu_{k-1}, w_{k-1}\}$, it is important that the training set for the EM algorithm contains about the same number of frames for each action that is supposed to be detected later on. Here Σ_i, μ_i and w_i denote the covariance matrix, the mean and the i^{th} mixture component's weight. Figure 4 illustrates the results that were obtained for one of the offices. For all further steps the video features are computed locally for each Gaussian i within all pixels $M_i = \{m \in \mathbb{N}^2 | (m-\mu_i)^T \Sigma_i^{-1} (m-\mu_i) \leq 3\}$ which have a Mahalanobis distance of less than three. The video feature vector finally consists of:

- The cumulated foreground mass:

$$cd_i = \sum_{j=1}^{|M_i|} fg_t(m_j)$$

- The joint probability from the local feature model on all foreground pixels $F_i = M_i \cap \{m \in \mathbb{N}^2 | fg_t(m) > 0\}$ restricted to mixture component i:

$$jp_i = \prod_{j=1}^{|F_i|} \frac{1}{2\pi |\Sigma_i|^{1/2}} e^{-\frac{1}{2}(x_j-\mu_i)^T \Sigma_i^{-1} (x_j-\mu_i)}$$

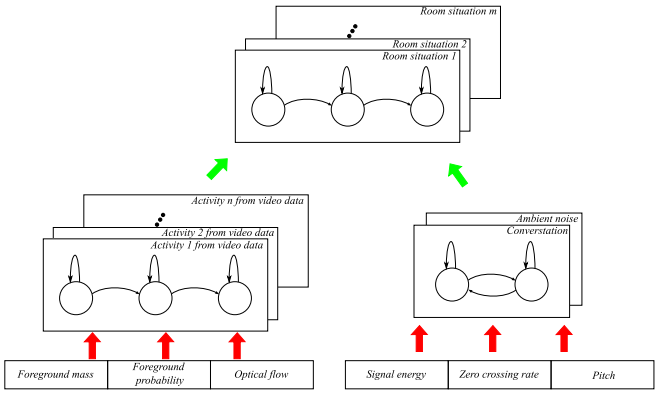


Fig. 5. Structure of the multi-layer HMM for a single office. The lower level recognizes events whereas the higher level represents *room situations*

- The optical flow's median in x and y direction (of_i^x and of_i^y)

Accordingly, the full video feature vector, that is fed into the lowest level HMM, consists of $k \cdot 4$ components.

$$o_t^v = \{cd_0, jp_0, of_0^x, of_0^y, \ldots, cd_{k-1}, jp_{k-1}, of_{k-1}^x, of_{k-1}^y\}^T$$

To denote an interval of observations from t_1 to t_2 the notations $o_{t_1:t_2}^v$ and $o_{t_1:t_2}^a$ will be used.

III. Activity Recognition

HMMs [14] have been proven to be a suitable framework for various activity recognition tasks. Nevertheless, for a high dimensional feature space, a lot of training data is needed to estimate the hidden parameters if all features are fed into the same so called early integration HMM.

A. Multi-layered HMMs

Decomposing the parameter space into several layers reduces the amount of training data required and gives a better intuition on the learning process. The basic idea of a layered HMM is that HMM layer $l+1$ is connected to layer l by using its output probabilities as observations. Each layer can be trained on labeled data on its own by employing the well known Baum-Welch parameter estimation algorithm. Moreover, the inference interval i_l can be chosen individually for each layer.
Our approach (cf. figure 5) uses two sets of HMMs on the lowest level. The first set called $V^1 = \{V_0^1, \ldots V_{N-1}^1\}$ is trained to recognize activities from the video stream based on the current observations $o_{t-i_1:t}^v$, whereas the second one denoted as $A^1 = \{A_0^1, A_1^1\}$ is used to distinguish between spoken conversation and ambient noise based on $o_{t-i_1:t}^a$. The inferential outputs $P(o_{t-i_1:t}^v | V_i^1)$ and $P(o_{t-i_1:t}^a | A_j^1)$ are then passed on to the set of second level HMMs $S^2 = \{S_0^2, \ldots, S_{M-1}^2\}$ which recognize the situation that is taking place in an office.

B. Model Training and Inference

Various HMMs are trained dependent on the action to be recognized:

- For each room a set of two ergodic HMMs was trained to separate ambient noise from speech. As inferential output $\frac{P(o^a_{t-i_1:t}|A^1_0=Speech)}{P(o^a_{t-i_1:t}|A^1_1=Ambient)}$ was passed on.
- Moreover, several ergodic HMMs V^1 are trained with a fixed inference length to detect persons in a room's meaningful areas. For some of them another HMM was trained to recognize a visitor sitting next to them, which typically resulted in more motion.
- To improve separability yet another HMM $\overline{V}_i$ was trained with data that showed counterexamples of the activity to be recognized. Here again the ratio $R^v_{i,t} = \frac{P(o^v_{t-i_1:t}|V^1_i)}{P(o^v_{t-i_1:t}|\overline{V}^1_i)}$ was passed on to the second level.
- Finally, HMMs with a left-right topology were trained in order to recognize persons leaving or entering an office. For those, the mean training example length was used as inference interval.

Feature selection on the video features was necessary to avoid a priori dependences among different activities. Relevant features for each activity were determined by taking example sequences where activities were observed individually. Only features from those foreground mixture components were selected, which were necessary to cover at least 80% of all foreground pixels.

On the second level situations like NOBODY IN THE OFFICE, PHONE CALL, MEETING, DISCUSSION and PAPERWORK were trained on the same feature vector consisting of the output probabilities of the first level. The situation with the maximum output probability on the second level constitutes the final activity recognition result.

IV. ROOM-LEVEL TRACKING

It is not only important to know what a person is doing, but also where he currently is. This motivates to track the users across several rooms by using the information gathered by the first level HMMs. In general, this problem can be described as a dynamic Bayes net where events are observed which depend on real state and data association. Our approach exploits a Bayes filter framework [15], where the state vector x_t contains the belief that a certain person is in a certain room or out of sight. A separate tracker for each person is run and data association is performed in two stages: A standard nearest neighbor filter is applied to restrict the observations z_t to states close to the highest belief state. Moreover, the observation model is designed in a way that persons can only be observed at certain places depending on the room. The graph in figure 6 depicts the state space and the possible transitions.

A. Bayes filter approach

In general, a Bayes filter can be expressed as:

$$p(X_t = x_t|z_t) = k_t \cdot p(z_t|X_t = x_t)\cdot \sum_{x' \in X} p(X_t = x_t|X_{t-1} = x')p(X_{t-1} = x'|z_{t-1})$$

where $p(X_t = x_t|X_{t-1} = x')$ denotes the *dynamic model* and $p(z_t|X_t = x_t)$ the *observation model*. As the state space consisting of the set of possible locations is quite small, the sum can in this case be evaluated analytically. The transition probabilities for the dynamic model can be learned from labeled data. The observation model, which partially solves the data association problem, shall be described in more detail.

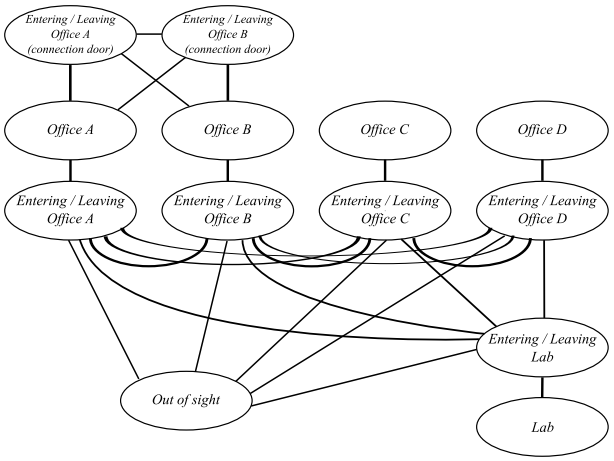


Fig. 6. State space for the room-level tracking problem; one dedicated network for each person was being employed; self transitions are not depicted for the sake of readability

B. Observation model

The first level outputs $R^v_{i,t}$ are regarded as binary sensors after taking a threshold m_i. Moreover, for each person a priori the sensor h is known where he normally works. Therefore, the remaining seats vs^k_i can be taken by a visitor. Hence, the observation model can be distinguished into the *home state* hs and *visiting states* $\{vs_0, \dots vs_n\}$ and their respective entering/leaving states. For the home state the observation probability $p(z_t|X_t = hs)$ becomes 1 if a seat event at sensor h is triggered and ϵ otherwise. The connected entering/leaving states hs^{el} have a observation probability $p(z_t|X_t = hs^{el})$ of 1 assigned if a door event is triggered and a state change occurs at the according seat at the same time and ϵ otherwise. The observation model for visiting states is defined accordingly. Due to the latency resulting from the segmentation's background adaption and the inference length of the HMMs it was necessary to smooth the raw first level HMM inference results prior to evaluation of the observation model. Finally, if a person was too long out of view, the tracker is reinitialized as soon as an observation at the own desk occurred.

V. RESULTS

A. Experimental setup

We equipped four of our offices with audio and video sensors. Video has been recorded with a 640x480 pixel resolution at a framerate of 7.5Hz by fixed cameras. The cameras have about 90° field of view and are mounted close to the ceiling in the rooms' corners. Hence, almost the entire room can be observed by a camera. Figure 2 shows sample images taken from each office.

Description	*Recognition rate*	*False positive rate*	*Percentage of data*
Somebody at User 3's desk	92.2 %	4.0 %	75.3 %
Somebody at User 4's desk	98.4 %	1.8 %	65.6 %
Visitor behind User 3's desk	63.2 %	17.2 %	3.8 %
Visitor behind User 4's desk	78.1 %	13.9 %	2.4 %
Somebody around visitor's chair	98.3 %	11.5 %	1.2 %
Somebody enters	100.0 %	3.8 %	0.2 %
Somebody leaves	98.2 %	4.4 %	0.2 %
Somebody entering through side door	94.7 %	3.2 %	0.2 %
Somebody leaving through side door	91.0 %	2.5 %	0.3 %

TABLE I

1ST LEVEL RECOGNITION RESULTS FOR OFFICE B

Audio was sampled at a framerate of 16kHz and features were gathered on windows of 20ms with an overlap of 10ms. Omni-directional microphones are used and the signal is pre-amplified in order to sense audio signals from the entire room. To ensure the users' privacy we did not record the raw audio stream, but extracted the features online and stored them tagged with timestamps for later processing.

Overall, for the activity recognition part we collected data of six days, of which we used four days for training and two days for evaluation. As this set of data only contained few events of people changing the offices, we recorded a second set with a scripted sequence of events with the length of about an hour which we used to evaluate room level tracking.

B. Activity Recognition

1) First Level: An event was being detected whenever the output ratio exceeded an experimentally determined threshold on the probability ratios as described for the room-level tracking. Moreover, an overlap of one inference interval shall be allowed to account for the HMM's latency. Table I exemplarily shows the results for the office B on smoothed output. Smoothing was done by assigning the activity that occurred most often in a sliding window to the center position. It can be observed that recognition for a distinct place, like SOMEBODY AT A DESK works quite well. Whereas recognition of two persons who are close to each other like for VISITOR BEHIND A DESK works less well yielding a relatively high false positive rate. These missclassifications mainly occur when a single user moves a lot or when two persons cause relatively little motion. Moreover, it shall be noted that leaving and entering events are often confused due to the similarity of a opening door in this feature space.

2) Second Level: For the second level missclassifications at boundaries were not counted as well. Moreover, the final results have been smoothed. Confusion matrices of the most interesting office rooms B and D as well as recognition and false positive rates are given in table II and table III. It can be observed that the situation of NOBODY IN THE OFFICE and PAPERWORK can be recognized quite reliably in both offices. Problems arise for the distinction of meeting and paperwork if there is too little conversation, and for discussion and meeting if there is too much or too little motion. Here it shall be noted, that a discussion was defined as a conversation of two persons who work in the same room. Meetings were defined as conversations with an external visitor where both participants face each other or look onto the same display. Another reason for the missclassification of meetings as discussions is that the recognition of the visitor failed on the first level due to the fact that they were close to each other and too little motion occurred so that they were only perceived as one person.

Description		[1]	[2]	[3]	[4]
Nobody in the office	[1]	**3462**	10	144	11
Paperwork	[2]	695	**20341**	723	663
Discussion	[3]	76	123	**4890**	1524
Meeting	[4]	0	793	203	**2278**

Description	*Recognition rate*	*False positive rate*	*Percentage of data*
Nobody in the office	95.5%	0.5%	10.1%
Paperwork	90.7%	5.8%	62.4%
Discussion	73.9%	4.8%	18.4%
Meeting	69.6%	2.8%	9.1%

TABLE II

CONFUSION MATRIX (IN SECONDS) AND RECOGNITION RATES FOR THE SECOND LEVEL FOR OFFICE B

Description		[1]	[2]	[3]	[4]
Nobody in the office	[1]	**6995**	17	117	32
Paperwork	[2]	659	**12877**	352	1042
Phone call	[3]	76	977	**4294**	776
Meeting	[4]	26	1031	685	**2611**

Description	*Recognition rate*	*False positive rate*	*Percentage of data*
Nobody in the office	97.7%	0.5%	22.0%
Paperwork	86.2%	6.3%	45.8%
Phone call	70.1%	5.6%	18.8%
Meeting	60.0%	5.3%	13.4%

TABLE III

CONFUSION MATRIX (IN SECONDS) AND RECOGNITION RATES FOR THE SECOND LEVEL FOR OFFICE D

C. Room-level Tracking

The data set to evaluate room-level tracking contained 44 transitions. To show our algorithm's performance we evaluated both the overall percentage of correctly tracked frames and the percentage of correctly identified transitions. In the experimental setup we tracked seven persons despite of clutter which was caused by two additional persons who were working in the lab. As the hallway was not monitored due to privacy reasons blind gaps occurred between the cameras. On average 91.5% of the frames were tracked correctly and

User	*Frame accuracy*	*Overall transitions*	*Recognized transitions*
User 1	94.5 %	4	4
User 2	93.8 %	6	5
User 3	97.6 %	6	6
User 4	91.5 %	10	8
User 5	82.3 %	6	3
User 6	89.5 %	5	3
User 7	91.0 %	7	7
Overall	91.5 %	44	82.0 % ≙ 36

TABLE IV
RESULTS FOR ROOM-LEVEL TRACKING

Ground truth			**Tracking result**		
Begin	*End*	*Place*	*Place*	*Begin*	*End*
0.0	31.7	Office B	Office B	0.0	41.0
45.8	67.8	Lab	Lab	49.0	75.0
76.6	799.7	Office B	Office B	81.0	809.0
809.1	1109.6	Office D	Office D	810.0	1102.0
1117.1	2194.3	Office B	Office B	1103.0	2197.0
2194.3	2484.3	Office A	Office A	2198.0	2496.0
2495.5	2660.4	Office D	Office D	2497.0	2671.0
2667.8	3248.2	Office B	Office B	2672.0	3263.0
3248.2	3388.8	Out of view	Out of view	3264.0	3382.0
3388.8	3685.4	Office B	Office B	3383.0	3687.0
3685.4	3698.5	Lab	Lab	3688.0	3705.0
3708.7	3719.4	Office A	Office D	3709.0	3925.0
3719.4	3925.9	Office B			

TABLE V
EXEMPLARY TRAJECTORY FOR USER 4 (TIMES ARE GIVEN IN SECONDS)

36 transitions were correctly recognized. Table IV shows the detailed results. Table V shows ground truth and tracking results for one of the tracked persons.

Our experiments showed that tracking works quite well as long as persons are not out of sight for too long. Due to the sparsity of people crossing the hallway data association works quite well. Though, this situation becomes worse in case the track of a person out of sight is crossed by a second person who passes the hallway.

VI. CONCLUSION AND FUTURE WORK

We have presented a system that can recognize a set of characteristic activities in an office domain. Furthermore, it is able to track the users well across several rooms even under real world conditions. For future work we would like to investigate further features for the activity recognition part and to solve the data association for the multi person tracking problem in a more unconstrained way.

To improve first level activity recognition results, for example Haar cascades [16] trained on various head poses might be the right approach to detect the number of people in a scene in an illumination invariant manner.

To resolve the constraint that tracks are lost on the hallway while other users are crossing, data association might better be addressed with a Rao-Blackwellised particle filter [17] as proposed by Wilson *et al.* [9]. This approach regards the data association as well as the sensor space as common state space and solves tracking with a Bayesian filter as proposed here, but data association with a particle filter.

Moreover, color similarity based on clothing might be used to match observations correctly. This requires to calibrate the camera colors to ensure color constancy. Renno *et al.* give a survey on existing algorithms in [18].

ACKNOWLEDGMENTS

This work has been funded by the European Commission under Project CHIL (http://chil.server.de, contract #506909).

REFERENCES

[1] M. Danninger, G. Flaherty, K. Bernardin, H. K. Ekenel, T. Kluge, R. Malkin, R. Stiefelhagen, and A. Waibel, "The connector: facilitating context-aware communication." in *International Conference on Multimodal Interfaces*, 2005, pp. 69–75.

[2] N. Oliver, B. Rosario, and A. Pentland, "A bayesian computer vision system for modeling human interactions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 831–843, Aug. 2000.

[3] C. Stauffer and E. Grimson, "Learning patterns of activity using real-time tracking," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 747–757, 2000.

[4] M. Brand and V. Kettnaker, "Discovery and segmentation of activities in video," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 844–851, 2000.

[5] Y. Ivanov and A. Bobick, "Recognition of visual activities and interactions by stochastic parsing," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 852–872, Aug. 2000.

[6] N. Oliver, E. Horvitz, and A. Garg, "Layered representations for human activity recognition," in *International Conference on Multimodal Interfaces*, 2002, pp. 3–8.

[7] D. Zhang, D. Gatica-Perez, S. Bengio, and I. McCowan, "Modeling individual and group actions in meetings with layered hmms," *IEEE Transactions on Multimedia*, vol. 8, no. 3, pp. 509–520, June 2006.

[8] I. McCowan, D. Gatica-Perez, S. Bengio, G. Lathoud, M. Barnard, and D. Zhang, "Automatic analysis of multimodal group actions in meetings," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 3, pp. 305–317, Mar. 2005.

[9] D. Wilson and C. Atkeson, "Simultaneous tracking and activity recognition (STAR) using many anonymous, binary sensors," in *International Conference on Pervasive Computing*, 2005, pp. 62–79.

[10] D. Schulz, D. Fox, and J. Hightower, "People tracking with anonymous and ID-sensors using rao-blackwellised particle filters," in *International Joint Conferences on Artificial Intelligence*, 2003, pp. 921–928.

[11] A. de Cheveigné and H. Kawahara, "Yin, a fundamental frequency estimator for speech and music," *Journal of the Acoustical Society of America*, vol. 111, pp. 1917–1930, 2002.

[12] J. Shi and C. Tomasi, "Good features to track," in *IEEE Conference on Computer Vision and Pattern Recognition*, June 1994, pp. 593–600.

[13] B. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *DARPA Image Understanding Workshop*, 1981, pp. 121–130.

[14] L. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[15] D. Fox, J. Hightower, L. Liao, D. Schulz, and G. Borriello, "Bayesian filtering for location estimation," *IEEE Pervasive Computing*, vol. 2, no. 3, pp. 24–33, July–September 2003.

[16] P. Viola and M. Jones, "Robust Real-Time face detection," in *International Conference On Computer Vision*, July 9–12 2001, pp. 747–747.

[17] A. Doucet, N. de Freitas, K. Murphy, and S. Russell, "Rao-blackwellised particle filtering for dynamic bayesian networks," in *Conference on Uncertainty in Artificial Intelligence*, 2000, pp. 176–183.

[18] J.-P. Renno, D. Makris, T. Ellis, and G. Jones, "Application and evaluation of colour constancy in visual surveillance," in *International Workshop on Performance Evaluation of Tracking and Surveillance*, 2005, pp. 301–308.