Probabilistic Integration of Sparse Audio-Visual Cues for Identity Tracking

Keni Bernardin
Universität Karlsruhe, ITI
Am Fasanengarten 5
76131, Karlsruhe, Germany
keni@ira.uka.de

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

Alex Waibel
Universität Karlsruhe, ITI
Am Fasanengarten 5
76131, Karlsruhe, Germany
waibel@ira.uka.de

ABSTRACT

In the context of smart environments, the ability to track and identify persons is a key factor, determining the scope and flexibility of analytical components or intelligent services that can be provided. While some amount of work has been done concerning the camera-based tracking of multiple users in a variety of scenarios, technologies for acoustic and visual identification, such as face or voice ID, are unfortunately still subjected to severe limitations when distantly placed sensors have to be used. Because of this, reliable cues for identification can be hard to obtain without user cooperation, especially when multiple users are involved.

In this paper, we present a novel technique for the tracking and identification of multiple persons in a smart environment using distantly placed audio-visual sensors. The technique builds on the opportunistic integration of tracking as well as face and voice identification cues, gained from several cameras and microphones, whenever these cues can be captured with a sufficient degree of confidence. A probabilistic model is used to keep track of identified persons and update the belief in their identities whenever new observations can be made. The technique has been systematically evaluated on the CLEAR Interactive Seminar database, a large audio-visual corpus of realistic meeting scenarios captured in a variety of smart rooms.

Categories and Subject Descriptors

I.4.8 [**Image Processing and Computer Vision**]: Scene Analysis

General Terms

Algorithms, Experimentation, Performance

Keywords

smart environments, sensor fusion, modality fusion, human perception

MM'08, October 26–31, 2008, Vancouver, British Columbia, Canada.

1. INTRODUCTION

Smart spaces and environments that perceive their occupants' actions and offer intelligent human-centered services have been a topic of research for quite some time. In this context, the tracking and identification of persons (referred to here as "identity tracking") plays an important role, as it provides fundamental contextual knowledge upon which further analysis of activities or interactions can be performed [1, 2]. The overall goal is to simultaneously keep track of multiple identities evolving in the space using unobtrusive sensors, such as distantly placed cameras and microphones. Further, this is to be accomplished in everyday scenarios imposing little or no constraint on the behavior of users, such as in meeting rooms, office areas, living rooms, etc.

One of the main problems facing identity tracking is that in such realistic scenarios, reliable cues for person-specific identification are hard to obtain with the sensors described above. Generally observable features based on a person's overall appearance, such as the color of clothing, body height, etc., can be ambiguous (e.g. when all persons wear black) and may well vary considerably with time or environmental conditions (e.g. taking off one's jacket, sitting down). On the other hand, more invariant and person-specific features such as those gained by face or voice identification may only seldom be observable (such as when a good view of the face is available or when the person takes his/her turn speaking in a conversation).

The main idea followed here to overcome this problem is to opportunistically integrate reliable identification cues for each person whenever they become available and to keep track of identified persons until further observations can be made.

The difficulties to be dealt with are twofold: Firstly, single observations gained through face or voice identification are inherently noisy, being influenced by lighting conditions, low resolution, imperfect facial alignment, environmental noise, crosstalk, etc. This implies that identification cues need to be accumulated in time and multiple modalities should be used to increase the accuracy of identification. Secondly, in realistic scenarios, the tasks of automatically detecting and tracking persons in the first place cannot be assumed solved with perfect accuracy. Persons may be missed, tracks may be confused or lost. This means that person identities need to be correctly recovered when observations again become available.

While some amount of work has been done on the fields of tracking and identification using sensor networks with overlapping or even non-overlapping views, none of the ap-

proaches so far tackle all the related problems efficiently. Most integrated approaches rely on general appearance features, such as color (or on RFID tags and other worn sensors), and build on the assumption that features for identification are jointly available with features for tracking with every observation made. While some approaches, such as [3], use person-specific features such as provided by face identification, they still rely on the continuous availability of high resolution face images in very restrictive setups. Approaches that use acoustic features for identification typically assume that the number of persons is known a priori, that speakers take frequent turns, and do not keep track of their locations except in very restrictive setups [4]. More importantly: Almost all approaches found in the literature that target multiple users are limited to applications where the detection (and spatio-temporally local tracking) of persons can be realized flawlessly and build on the results of this step for identification [4, 6, 5, 9, 10].

In this paper, a new methodology is presented for the multimodal tracking and identification of multiple persons by fusing reliable tracking and ID cues whenever they become available. The method:

- Opportunistically integrates person-specific identification cues that can only sparsely be observed for each person over time
- Keeps track of the locations of multiple identified persons while ID cues are not available
- Combines the acoustic and visual modalities to increase its robustness and flexibility
- Does not rely on accurate detection and tracking, but rather considers both person locations and identities as attributes to be estimated.

The developed method is a non-parametric model-based approach based on Bayesian filtering of high level track and ID observations. It uses an EM learning algorithm to estimate the probability densities of a person's presence, location and identity based on the last k observations made. The proposed approach has been tested on a large annotated audio-visual corpus, the CLEAR Seminar Database [24, 21], comprising a total of 200 minutes from 20 different recorded small meetings. This database, captured in smart rooms using distantly placed sensors, features visual streams from several cameras on which tracking and face identification can be performed, as well as audio streams from several microphone arrays for speaker tracking and identification.

In the following section, a brief description is given of the tracking and identification components that feed their input to the probabilistic fusion module. Section 3 then explains the developed identity tracking method. In section 4, the integrated system is thoroughly evaluated, as well as compared to a baseline system, which builds on the results of the tracking step to infer identities in a sequential way. Finally, section 5 gives a summary and concludes.

2. TRACKING AND IDENTIFICATION CUES

In this section, the various components used for tracking, speaker localization, face recognition and speaker identification are presented. As they are not the focus of this paper, they are only briefly described, with references to previous work, giving more detailed explanations, made at the appropriate places.

Figure 1: The output of the person tracking component. The person tracks are indicated by colored circles at their estimated 3D positions

2.1 Multiple Person Tracking

The person tracking component is responsible for estimating the x,y,z scene coordinates of the persons in the environment (see Figure 1). It utilizes the images from five different synchronized and calibrated views of the room, and a variety of features to automatically detect and track multiple persons. The details of the tracker can be found in [18]. The performance of the tracker has been thoroughly evaluated on the CLEAR 2007 test set, and reached a tracking precision ($MOTP$) of 15cm and a tracking accuracy ($MOTA$) of 70%. The MOT measures are performance evaluation metrics for multiple object trackers. The former measures spatial accuracy whereas the latter measures the ability to find the correct number of objects and to keep consistent tracks in time (A detailed explanation of the MOT metrics can be found in [15, 16]). As these numbers and a further review of the overall accuracies reached in the CLEAR evaluations (see [21]) show, for such realistic and challenging scenarios, the flawless detection and tracking of multiple occupants is still not a realistic prerequisite.

2.2 Face Identification

Face identification is performed in each of the four available corner camera views whenever a frontal face can be found. Again, the difficulty lies in the fact that usually only low resolution snapshots of faces can be obtained from such views, resulting in poor identification confidence. Moreover, faces may be oriented downwards, tilted, occluded by heads or other objects, turned away from cameras, such that usable frontal faces can rarely be captured. Of course, these difficulties are partly compensated by the availability of multiple views of the scene. When a face area is found in an image, a local appearance-based identification technique is applied, using Discrete Cosine Transform (DCT) features and nearest neighbor classification. The details of the algorithm are given in [12, 13]. The technique accumulates identification results and confidences over several frames to increase its accuracy. It has been evaluated on the CLEAR 2007 database, in a closed set identification task featuring 28 individuals, using pre-segmented test segments of varying

lengths. It achieved 84.6% accuracy for the hardest condition in terms of data availability (15s training segments, 1s test segments) and 96.4% for the easiest condition (30s train, 20s test).

For the here presented evaluation, the classifiers are again trained on the same set of 28 known individuals, but the task is now that of open set identification, with 39 additional unknown faces appearing in the database. Moreover, the association of faces to persons for confidence accumulation is no longer known a-priori and has to be derived automatically. This makes the identification task considerably harder than in the aforementioned evaluation. An additional difficulty for face identification lies in the accurate alignment of cropped faces in low-resolution views, a problem which is still not satisfactorily solved (see [21]). In earlier work [17], we circumvented this problem by using active cameras to focus in on target persons and capture high resolution facial shots suitable for alignment and reliable identification. As no active camera views are available in the CLEAR 2007 database, manually annotated face bounding boxes are used here, as in the previous evaluation, instead of an automatic detection and alignment step. Nevertheless, most of the obtained faces are still badly recognizable because of the other mentioned difficulties, such that a confidence based filtering of face ID cues is applied, as explained later in this section.

2.3 Speaker Identification

Speaker localization is performed using the audio streams captured from at least four distributed microphone arrays on the room walls, while identification is made using just one audio channel. As compared to face identification, the difficulty in far-field acoustic identification lies in segmenting speech, separating multiple speakers, and dealing with low signal to noise ratios, reverberations, laughter, etc. Additionally, identification can usually only be made for one person at a time, the active speaker, while crosstalk is generally detrimental both for localization and identification. The algorithm applied here uses a set of Gaussian Mixture Models (GMMs) trained for each known speaker and Mel Frequency Cepstral Coefficients (MFCCs) as features. Additionally, a silence model was trained and used in parallel to the speaker models in testing, such that the tasks of segmentation and identification are performed jointly simply by analyzing the GMMs' resulting MAP probabilities. The details of the algorithm can be found in [14]. Just as for face ID, its performance has been tested in the CLEAR 2007 evaluations on the Interactive Seminar database [21], in a closed set identification task involving 28 individuals and using pre-segmented intervals of clean speech. It reached 86.7% accuracy for the hardest testing condition (15s train, 1s test) and 99.1% for the easiest condition (30s train, 20s test).

Again, for the here presented evaluation, the speaker identification task becomes an open set problem. A total of 27 speakers were trained in (of which 24 are also visually known, i.e. three known persons can only be identified using their voice). This means that 40 additional individuals occurring in the database are acoustically unidentifiable. An additional difficulty lies in automatically detecting clean segments of speech, usable for identification. In this approach, the audio stream is segmented into equal 1s segments and identification is made on each. Reliable speaker ID results are again recognized by analyzing the distribution of GMM MAP probabilities, and unreliable segments are not used in confidence accumulation. A definite disadvantage of the acoustic modality over the visual one is that speech cues can much more rarely be obtained for less active speakers, often resulting in a large delay before identification is possible.

2.4 Confidence Estimation and Spatial Localization

For both the frame-based results of face identification and the 1s interval speaker identification results, confidence measures in the found identities are derived in the form of discrete non-parametric probability distributions (pdfs) over the set of known ID labels. For face ID, the k-nearest neighbor classification result is a set of k distances to the training sample vectors closest to the test vector. Similarly, for Speaker ID, the result are the n highest MAP probabilities for the set of 28 speaker/silence GMMs. In both cases, the identity pdf is calculated by min-max normalization of the resulting values, followed by an additional normalization to unit sum. Only if a definite peak in the resulting pdf can be found (here, a threshold of 60% is used for peak detection) will it be used in the integration step to accumulate confidences and distinguish known from unknown persons.

Finally, identification cues are spatially localized whenever possible to allow the association to available person tracks. For visual identification, this is done by exploiting the expected width of a frontal face and camera calibration information. By using the detected face box center and width in a camera image, the distance to the camera can be estimated and a 3D scene location computed. As small variations in pixel width can cause great variations in the estimated distance, the derived 3D location comes with a certain amount of uncertainty, which is modeled by a 3-dimensional uncertainty covariance matrix. Acoustic localization is performed on the microphone array channels using Generalized Cross Correlation Features (GCC-Phat) and a Joint Probabilistic Data Association Filter framework (see [11] for details). The result of source localization is the 3D scene position of the current most active sound source, as well as an associated uncertainty covariance matrix. The acoustic source localization system was also evaluated on the CLEAR 2007 database and reached an accuracy ($MOTA$) of 55% and a precision ($MOTP$) of 14cm. One should note that speech localization inaccuracies can lead to cases where a speaker identification is possible, but localization is not. In this case, the found ID cue can later not be associated to a person track based on spatial mapping, but can well serve as a hint that the concerned person is actually present. In the next section, we will show how the integrated tracking approach makes use of this information.

3. AUDIO-VISUAL FUSION AND IDENTITY TRACKING

The main idea behind the design of our ID Tracking approach is to opportunistically integrate reliable but sparsely available cues for identification whenever they become available, and to keep tracking recognized persons in the absence of such. Audio-visual ID cues and person tracks of varying accuracy coming from the different system components are expected to arrive with varying regularity.

This raises the need for a fusion technique that handles incomplete and possibly very sparse information.

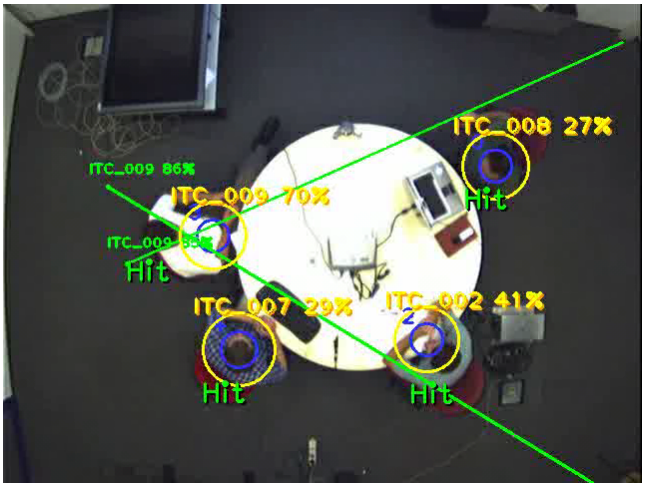

Figure 2: The output of the integrated identity tracking system. The blue and yellow circles represent the person tracker hypotheses and the person models, respectively. The identities for recognized persons are printed on top of the respective models. The green lines indicate face identification hits for the current frame, in this example made by two of the four corner cameras for one room occupant

3.1 Integrated Tracking and Identification

The probabilistic integrated identity tracking method considers person identities and locations as hidden variables to be jointly inferred in a Bayesian estimation process. Its main goal is to recognize the set of known persons in the smart environment, while considering their locations as additional information which may or may not be available from tracker outputs. As track information as well as captured identification hints are considered inherently flawed in this framework, no reliable cue is available for estimating the number of persons in the scene. In this work, we therefore limit the problem to the tracking and identification of a number of known persons among a greater set of unknown individuals. The known persons are those for which voice or face models have been trained in a priori. They will subsequently be referred to as "focus persons". In other words, we attempt to recognize the identities of a known number F of focus persons (from a larger set of N trained identities) evolving among a variable number M of unknown persons, and estimate their positions.

The developed algorithm works as follows: A set of person models $\{m_1...m_F\}$ is kept, one for each of the F focus persons, with $m_i = (id_i, st_i)$. The hidden variables id_i and st_i represent the person's identity and location respectively and are modeled by discrete non-parametric probability density functions. The person location, in this case, is not represented by his or her spatial x,y,z-coordinates in the scene. Rather, abstraction is made of the concrete locations, using available tracking information, where each track Tj represents a discrete state s_j with $j \geq 1$. Additionally, state s_0 represents the case where a person location does not coincide with any of the available tracks (in our case, when the person is present in the room but is not tracked). Person localization is therefore performed on a topological level, with the overall topology consisting of the room itself and all currently available tracks. Then, the location variable st_i becomes a discrete variable, just as the identity variable id_i. In addition to the models for known persons, a garbage model m_g is also kept, to which erroneous, noisy observations, or observations coming from unknown persons should be associated.

The observation sequence $\{o_1...o_t\}$ for our probabilistic model consists of the localized speaker ID and face ID cues obtained in time, with $o_i = (L_i, s_i)$ where the identity L_i is provided as discrete probability density function over all known identities, and the location s_i as a discrete state index derived in a track association step: The 3D location x_μ and covariance matrix Σ for each ID cue are used to evaluate track proximity and the association is made to the track T_j with location x_j maximizing $p(x_j|x_\mu, \Sigma) \approx N(x_\mu, \Sigma)$, resulting in the discrete state index j. An overlap threshold is however applied and observations which cannot be mapped to any specific track are assigned the state index 0.

For every new observation $o_i = (L_i, s_i)$, the person model pdfs are updated using an iterative EM-algorithm. For every model m_j, the similarity $d(L_i, id_j)$ between the observed identity pdf L_i and the modeled identity pdf id_j is measured and the association is made to m_k with

$$k = argmax_{1..F}(d(L_i, id_k)).$$

Here, the Bhattacharyya distance [23], with

$$d(p, q) = \sum_x \sqrt{p(x)q(x)}$$

is used as similarity measure for discrete pdfs. The identity and location pfds, id_k and st_k, are then updated using the last n observations for model m_k. This is done by storing the observations associated to each person model and using the last n stored observations to derive its current pdfs. Since acoustic and visual identification cues may come at a sensibly different rate, due to the availability of several camera views captured at high framerates, it makes sense to store location estimates, acoustic ID and visual ID pdfs in separate queues, q_s, q_a, and q_v. In the update step, separate probability density functions for visual and acoustic ID are first computed, by averaging the stored information for each modality, and the combined audio-visual pdf id_k, as used in the expectation step, is then obtained as a weighted sum of the two. Here, equal weights are assigned to the audio and visual modalities.

Finally, the highest MAP identity label for each person model m is derived from id_m, by assuming uniform prior distribution of identities, and the output of the identity tracking module are the F person models with the highest MAP probabilities. Figure 2 shows an example output of the integrated identity tracking system on a CLEAR seminar with four known participants.

3.2 Baseline Sequential System

As a baseline to evaluate the advantages of the integrated ID tracking approach, a sequential algorithm was implemented which relies on an accurate detection and tracking step to estimate person identities. The baseline fusion system initializes a person model for each track delivered by the multiple camera tracker and uses these person models as the basis for spatial association of ID cues. Here again, a person model comprises acoustic, visual and audio-visual pdfs for the iterative learning of identities. In contrast to

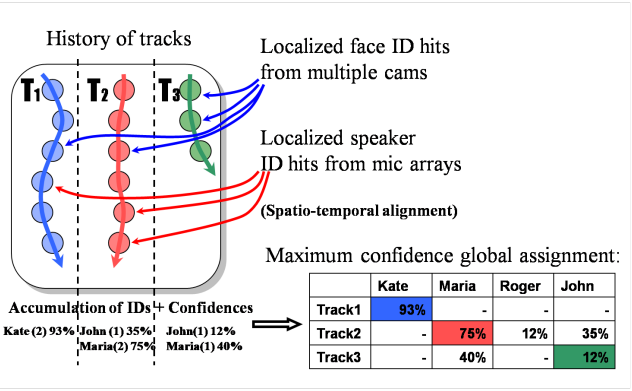

Figure 3: The process of mapping localized ID cues to person tracks. T_1 to T_3 represent the person tracker output. The blue and red arrows represent sporadically captured face and speaker ID cues, respectively. A spatio-temporal mapping is made for both types of cues, confidences are accumulated in time and a global assignment of IDs to tracks is made that optimizes the overall confidence level

the integrated system, though, the person location, here, is not a variable to be estimated, but is directly given by the 3D coordinates of the corresponding track. The association of visual and acoustic ID cues to person models is then made in the following way: The observed cues, derived from single frames for face ID and from 1s intervals for speaker ID, and tagged with their 3D location estimates, are compared to model locations and mapped to the closest overlapping model (with the overlap threshold set to 50cm). Cues which can not be associated to any of the available models (such as non-localized speech) are ignored.

After accumulation, the final identification hypothesis is not determined for each track independently, e.g. based on the highest MAP label, but rather by globally optimizing the hypothesis outputs jointly for all tracks. For each model m, the identification confidence $P(l|m)$ for each label l in consideration is derived from the model's audio-visual ID pdf. Finding the assignment of distinct identities to person models that maximizes the overall confidence is a combinatorial problem (a maximum weight assignment problem), which is solved here using Munkres' algorithm [22]. The optimal assignment is recomputed every time a new identification hit is received. The advantage of joint assignment is that mapping of the same ID to several persons is excluded, as the system will change the hypothesized ID for one track based on new information for another track. Figure 3 shows the process of spatio-temporal association and ID assignment. The output of the baseline system, just as for the probabilistic integration system, are the F tracks with the highest MAP identity confidence.

One obvious drawback of the baseline method is that only tracked persons can be identified and a learned identity is lost when the corresponding person track is lost. It then has to be relearned from subsequent observations as soon as tracking information is again available.

4. EXPERIMENTAL EVALUATION

This section describes the data and metrics used to evaluate combined tracking and identification performance, and presents comparative results for the baseline system and the integrated probabilistic fusion approach.

4.1 Evaluation Database

The developed method for integrated identity tracking has been extensively evaluated on the Interactive Seminar database used in the CLEAR 2007 evaluation [19, 20]. This database features recordings of multiple users in realistic small meeting scenarios, captured in a variety of smart rooms equipped with a multitude of audio-visual sensors (see Figure 4). It offers five calibrated and synchronized visual streams from corner and ceiling cameras, as well as synchronized audio streams from a minimum of four microphone arrays on the room walls. The dataset comprises 20 seminars from five recording rooms with varying audio-visual characteristics, with two annotated five minute segments per seminar, for a total of 200 minutes of recordings. In this dataset, a total of 67 individuals take part in small meetings, with typical meeting sizes of three to five persons. Of these 67 identities, 24 are trained in audio-visually, three are trained in using only the acoustic modality and four using only the visual one. The ratio of known to unknown persons varies with each meeting, with a slightly greater number, on average, of unknown persons.

4.2 Evaluation Goals and Metrics

The goal of evaluation is to measure the performance of the presented identity tracking technique at recognizing and tracking a subset of known focus persons interacting with several unknown ones in a smart environment. The evaluation procedure is defined in accordance: We define a cumulative identification score (ia), measuring the accuracy of the tracker at estimating focus person identities, and a localization score (la), measuring the tracker's ability to find the correct person positions within a certain tolerance level.

Let $L = \{l_1 \dots l_n\}$ be the set of labeled focus persons and $H = \{h_1 \dots h_m\}$ the hypothesis output by the identity tracker for one time frame.

Let also ia and la be initially set to 0. For every evaluated time frame t,

1. Let g_t be the number of labeled identities in L_t.
2. For every identity l_i in L_t, verify if a corresponding identity is included in the set H_t. If yes, increase the identification score ia by one. If additionally the labeled and hypothesized person positions overlap (with the overlap threshold set to 50cm), increase also the localization score la.
3. For all identities in L_t for which no match has been found, verify if at least a not yet mapped identity in H_t spatially overlaps with it. In this case, increase the la score by one. All remaining identities are considered missed.

Now let $G = \sum_t g_t$, be the total number of ground truth identity labels in the sequence.

We then define the following metrics for tracking and identification performance: The Identification Accuracy $IA = ia/G$, representing the ability to correctly recognize identities independent of tracking performance, averaged over

(a) *AIT* (b) *UKA* (c) *FBK-IRST* (d) *IBM* (e) *UPC*

Figure 4: Scenes from the CLEAR 2007 Interactive Seminar database

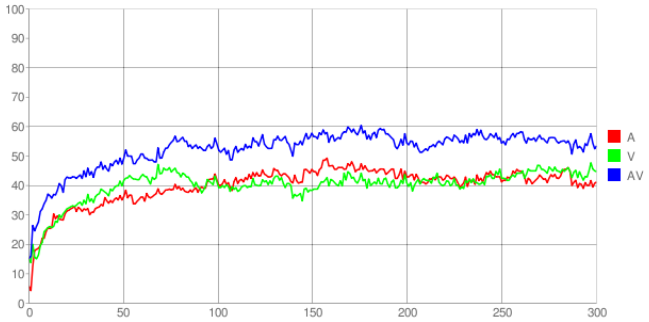

Figure 5: Evolution of the frame-based Identity Tracking Accuracy (ITA) in time, averaged over all seminars and segments, for audio, visual, and audio-visual identification

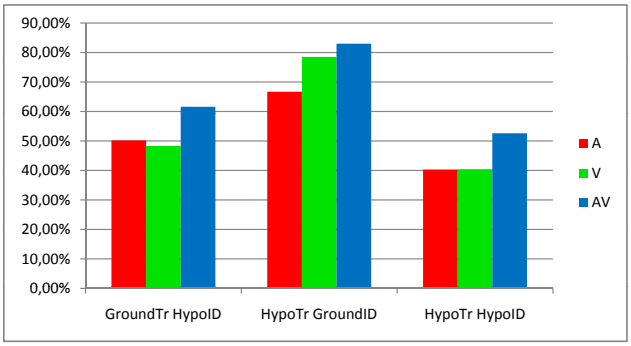

Figure 6: Comparative ITA results using ground truth or system hypothesis tracking and identification cues

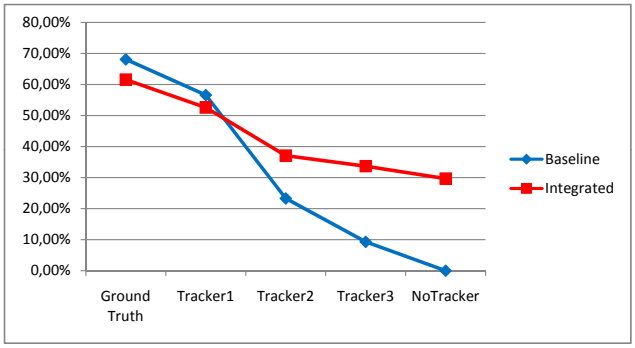

Figure 7: Results of the integrated identity tracking approach (ITA), compared to the baseline algorithm, with decreasing tracking accuracy

all labeled identities and time frames, and the Localization Accuracy $LA = la/G$, measuring the quality of position estimation. The overall Identity Tracking Accuracy over the entire sequence (ITA) is then defined as: $ITA = \frac{LA+IA}{2}$.

4.3 Evaluation Results

Figure 5 shows the evolution of frame-level ITA scores for the integrated probabilistic approach when using voice ID, face ID, or both cues for person identification. As input tracks to the system, the hypotheses of the multiple person tracker described in section 2.1 were used. As can be clearly seen, a noticeable advantage is to be gained from the fusion of modalities, both in the speed with which identification confidences rise, as well as in the overall accuracy reached. Figure 6 shows the average accuracies reached over all sequences and segments (*HypoTrHypoID* group in Figure 6).

To further investigate the effects of tracking or identification quality on overall accuracies, separate evaluations were also conducted using manually labeled person tracks together with automatically captured face and voice ID cues (*GrounTrHypoID* group in Figure 6), and using manually labeled face and voice recognition cues (derived from manual annotations of frontal faces in the images and of speaker activity in the far-field audio channels) in combination with automatic track hypotheses (*HypoTrGroundID* group in Figure 6).

One should notice here that although the measured accuracy of the person tracking component lies around 70%, this concerns the tracking of *all* the persons in a sequence, including unknown ones. On average, the accuracy concering only focus persons is much higher though (more than 90%), as these are usually the main speakers or presenters in the meeting, and can therefore more easily be tracked. This explains why accuracies do not drop significantly when going from manual tracks to hypothesis tracks.

Finally, the performance of the integrated approach is compared to the baseline system. This is shown in Figures 7 and 8. As the drop in accuracy when passing from labeled to hypothesis tracks (Tracker1) was not significant enough to illustrate the effects of tracker failure, additional tracking system hypotheses were simulated by manually removing the tracks for one (Tracker2) or two (Tracker3) focus persons from the tracker hypothesis, and finally by using no tracker output at all. As can be expected, the LA and IA scores drop considerably for the baseline system while for the integrated approach, at least the IA score stays relatively

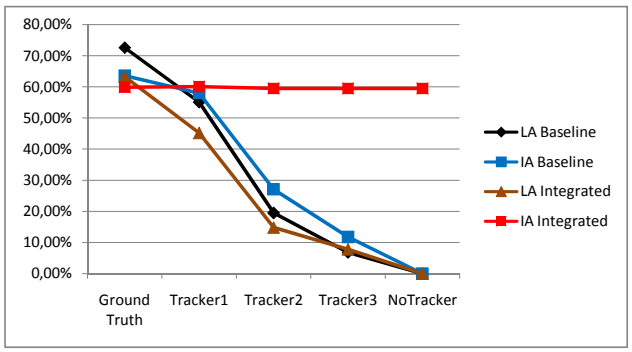

Figure 8: Localization accuracy (LA) and Identification accuracy (IA) for baseline and integrated systems, with decreasing tracking accuracy

constant through degrading tracker performance. This illustrates the advantage of considering tracking information only as an additional, possibly incorrect, source of information. The probabilistic integrated approach is capable of providing basic information about the room occupants, as long as some of the underlying tasks of person tracking, source localization, face recognition, or speaker identification, can be accomplished with a sufficient degree of accuracy.

5. CONCLUSION

In this paper, we have presented a novel technique for the tracking and identification of multiple persons in a smart environment using distantly placed audio-visual sensors. The technique builds on the opportunistic integration of tracking as well as face and voice identification cues, gained from several cameras and microphones, whenever these cues can be captured with a sufficient degree of confidence. A probabilistic model was introduced that jointly estimates person locations and identities based on audio-visual observations. The technique has been systematically evaluated on the CLEAR Interactive Seminar database, and compared to a baseline technique, performing tracking and identification in a sequential way. As results show, the integrated approach is much more robust to tracking failures and degrades gracefully with decreasing tracking accuracy. The results also show that the fusion of audio and visual modalities can help achieve higher identification accuracies, even in relatively uncontrolled situations with multiple persons, occlusions, cross-talk, etc. For an open set identification task, even under the challenging conditions posed by the CLEAR Seminar database, noticeable identity tracking accuracies could be reached using available state-of the art tracking, face and voice identification components.

6. ACKNOWLEDGMENTS

The authors wish to thank Hazim Ekenel, Tobias Gehrig and Qin Jin for their invaluable contributions to this work.

7. REFERENCES

[1] I. McCowan, D. Gatica-Perez, S. Bengio, G. Lathoud, M. Barnard, 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, March, 2005.

[2] R. Stiefelhagen, *"Tracking Focus of Attention in Meetings"*. IEEE Int. Conf. on Multimodal Interfaces - ICMI 2002, Pittsburgh, 2002.

[3] T. Choudhury, B. Clarkson, T. Jebara and A. Pentland, *"Multimodal Person Recognition using Unconstrained Audio and Video"*. Second Conference on Audio- and Video-based Biometric Person Authentication '99 (AVBPA '99), pp. 176–181, Washington DC

[4] J. Yang, X. Zhu, R. Gross, J. Kominek, Y. Pan, A. Waibel, *"Multimodal people ID for a multimedia meeting browser"*. Proceedings of the 7th ACM International Conference on Multimedia '99, Orlando, FL

[5] A. Hampapur, S. Pankanti, A. W. Senior, Y.-L. Tian, L. Brown, R. M. Bolle, *"Face Cataloger: Multi-Scale Imaging for Relating Identity to Location"*. IEEE Conference on Advanced Video and Signal Based Surveillance (AVSS 2003), July 2003, Miami, FL.

[6] S. Stillman, R. Tanawongsuwan, and I. Essa, *"A system for tracking and recognizing multiple people with multiple cameras"*. Technical Report GIT-GVU-98-25, Georgia Inst. of Tech., Graphics, Visualization, and Usability Center, 1998.

[7] S. Stillman and I. Essa, *"Towards reliable multimodal sensing in aware environments"* Perceptual User Interfaces (PUI) Workshop, 2001.

[8] M. Trivedi, I. Mikic and S. Bhonsle, *"Active Camera Networks and Semantic Event Databases for Intelligent Environments"*. IEEE Workshop on Human Modeling, Analysis and Synthesis, June 2000.

[9] Dimitrios Makris, Tim Ellis, James Black, *"Bridging the Gaps between Cameras"*. IEEE Conference on Computer Vision and Pattern Recognition (CVPR'04) - Vol. 2, 2004

[10] W. Zajdel, B. J. A. Kröse, *"A sequential Bayesian algorithm for surveillance with nonoverlapping cameras"*. IJPRAI, Vol. 19, No. 8, Dec 2005

[11] T. Gehrig, K. Nickel, H. K. Ekenel, U. Klee, and J. McDonough, *"Kalman Filters for Audio-Video Source Localization"*. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, October. 2005.

[12] H. K. Ekenel, R. Stiefelhagen, *"Local Appearance based Face Recognition Using Discrete Cosine Transform"*. 13th European Signal Processing Conference (EUSIPCO), Antalya Turkey, September 2005.

[13] H. K. Ekenel, R. Stiefelhagen, *"A Generic Face Representation Approach for Local Appearance based Face Verification"*. CVPR IEEE Workshop on Face Recognition Grand Challenge Experiments, San Diego, CA, USA, June 2005.

[14] H. K. Ekenel, Q. Jin, *"ISL Person Identification Systems in the CLEAR Evaluations"*. Proceedings of the first International CLEAR evaluation workshop, Southampton, UK, April 2006.

[15] K. Bernardin, A. Elbs, R. Stiefelhagen, *"Multiple Object Tracking Performance Metrics and Evaluation in a Smart Room Environment"*. 6th IEEE Int. Workshop on Visual Surveillance, VS 2006, Graz, Austria, May 2006

[16] K. Bernardin and R. Stiefelhagen, *"Evaluating Multiple Object Tracking Performance: The CLEAR MOT Metrics"*. EURASIP Journal on Image and Video Processing, Special Issue on Video Tracking in Complex Scenes for Surveillance Applications, Vol. 2008, Article ID 246309, May 2008

[17] K. Bernardin and R. Stiefelhagen, *"Audio-Visual Multi-Person Tracking and Identification for Smart Environments"*. ACM Multimedia 2007, Augsburg, Germany, September 2007

[18] K. Bernardin, T. Gehrig, R. Stiefelhagen, *"Multi-Level Particle Filter Fusion of Features and Cues for Audio-Visual Person Tracking"*. Multimodal Technologies for Perception of Humans, Joint Proceedings of the CLEAR 2007 and RT 2007 Evaluation Workshops, May 2007, Baltimore, MD, USA, Springer LNCS 4625, 2008

[19] R. Stiefelhagen, K. Bernardin, R. Bowers, T. Rose, M. Michel and J. Garofolo, *"The CLEAR 2007 Evaluation"*. Multimodal Technologies for Perception of Humans, Joint Proceedings of the CLEAR 2007 and RT 2007 Evaluation Workshops, May 2007, Baltimore, MD, USA, Springer LNCS 4625, 2008

[20] D. Mostefa, N. Moreau, K. Choukri, G. Potamianos, S. M. Chu, A. Tyagi, J. R. Casas, J. Turmo, L. Christoforetti, F. Tobia, A. Pnevmatikakis, V. Mylonakis, F. Talantzis, S. Burger, R. Stiefelhagen, K. Bernardin, and C. Rochet, *"The CHIL Audiovisual Corpus for Lecture and Meeting Analysis Inside Smart Rooms"*. In Language Resources and Evaluation, No. 41, Springer, 2007.

[21] Rainer Stiefelhagen, Jonathan Fiscus and Rachel Bowers, *"Multimodal Technologies for Perception of Humans, Joint Proceedings of the CLEAR 2007 and RT 2007 Evaluation Workshops"*. Springer Lecture Notes in Computer Science, No. 4625, 2008.

[22] J. Munkres, *"Algorithms for the Assignment and Transportation Problems"*. Journal of the Society of Industrial and Applied Mathematics, Vol. 5(1), pp. 32-38, March 1957.

[23] T. Kailath, *"The Divergence and Bhattacharyya Distance Measures in Signal Selection"*. IEEE Trans. on Comm. Technology, Vol. 15, pp. 52–60, Feb. 1967

[24] CLEAR - Classification of Events, Activities and Relationships, http://www.clear-evaluation.org/