ARTICLE IN PRESS

Available at
www.ElsevierComputerScience.com
POWERED BY SCIENCE DIRECT®

Signal Processing ▮ (▮▮▮▮) ▮▮▮–▮▮▮

SIGNAL PROCESSING

www.elsevier.com/locate/sigpro

Audio-visual perception of a lecturer in a smart seminar room

R. Stiefelhagen*, K. Bernardin, H.K. Ekenel, J. McDonough, K. Nickel, M. Voit, M. Wölfel

Interactive Systems Labs, Universität Karlsruhe (TH), Germany

Received 1 July 2005; received in revised form 5 December 2005; accepted 1 February 2006

Abstract

In this paper we present our work on audio-visual perception of a lecturer in a smart seminar room, which is equipped with various cameras and microphones. We present a novel approach to track the lecturer based on visual and acoustic observations in a particle filter framework. This approach does not require explicit triangulation of observations in order to estimate the 3D location of the lecturer, thus allowing for fast audio-visual tracking. We also show how automatic recognition of the lecturer's speech from far-field microphones can be improved using his or her tracked location in the room. Based on the tracked location of the lecturer, we can also detect his or her face in the various camera views for further analysis, such as his or her head orientation and identity. The paper describes the overall system and the various components (tracking, speech recognition, head orientation, identification) in detail and presents results on several multimodal recordings of seminars.

Keywords: Smart rooms; Multimodal–multisensor interfaces; Audio-visual tracking; Head pose estimation; Face recognition; Far-field speech recognition

1. Introduction

In recent years there has been much research effort spent on building smart perceptive environments, such as smart living rooms [1], smart lecture and meeting rooms [2–5] or smart houses [6,7]. Such smart spaces are usually equipped with a variety of sensors which allow for automatic acquisition of information about the users and their activities. The challenge then is to build smart spaces which support humans during their activities inside them without obliging them to concentrate on operating complicated technical devices.

In the framework of the project CHIL, Computers in the Human Interaction Loop (http://chil.server.de), we are developing services that aim at proactively assisting people during their daily activities and, in particular, during their interaction with others. Here, we focus on office and lecture scenarios, as they provide a wide range of useful applications for computerized support.

To provide intelligent services in a smart lecture environment it is necessary to acquire basic information about the room, the people in it and their interactions. This includes, for example, the number of people, their identities, locations, postures, body

*Corresponding author.
E-mail address: stiefel@ira.uka.de (R. Stiefelhagen).

0165-1684/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.sigpro.2006.02.043

and head orientations, and the words they utter, among others.

In this work, we describe our efforts at building technologies to automatically extract such information in a smart seminar room. In particular, we describe our latest perceptual components to locate and track the lecturer in the room, to transcribe his or her speech, both from close-talking microphones (CTMs) as well as from far-field microphone arrays (MAs), to estimate his or her head orientation and finally to visually recognize the lecturer's identity.

All four components—localization and tracking, speech recognition, estimating head orientation, face recognition—provide valuable information that can be used both for the annotation and indexing of multimodal seminar recordings, as well as to provide necessary context information to build real-time services that support the lecturer or students in the smart room.

Locating the lecturer in a seminar room is mandatory for many applications: First, knowing the lecturer's position can be used to improve far-field speech recognition using MAs. The experiments which we present in this paper clearly show that the accuracy in determining the speaker location has a direct influence on the quality of automatic speech recognition (ASR) measured in terms of word error rate (WER). Second, a person's location and trajectory of movement can provide important context for analyzing his or her activities. It can be useful, for example, to know whether a lecturer stands close to a whiteboard or not.

Once a lecturer has been located, we can detect his or her head around the estimated location for further analysis, such as analyzing his or her head orientation and identity.

A person's head orientation is a reliable cue to determine his or her focus of attention [8]. In seminars or lectures it could, for example, be used to analyze the interaction between the students and the lecturer as well as the level of attention of single students or the audience as a whole. Several user studies reported strong evidence that people naturally look at the objects or devices with which they interact [9,10]. The lecturer's head orientation could therefore also be useful to tell the "smart room" what the lecturer currently interacts with, for instance, a whiteboard, his or her laptop or the audience.

1.1. System overview

Fig. 1 shows a block-diagram of the perceptual components in our smart room.

First, the person tracking module uses the input coming from several T-shaped MAs and the videos coming from four calibrated cameras in order to produce an estimate of the lecturer's position in the room. This estimated position is then used for acoustic beamforming and further face alignment.

The face alignment module searches the lecturer's face in the four video streams using the predicted 3D position given by the AV person tracking module. The output of the face alignment module are the aligned faces of the lecturer in the different camera views, which are then used both by the face identification module and the head pose estimation module.

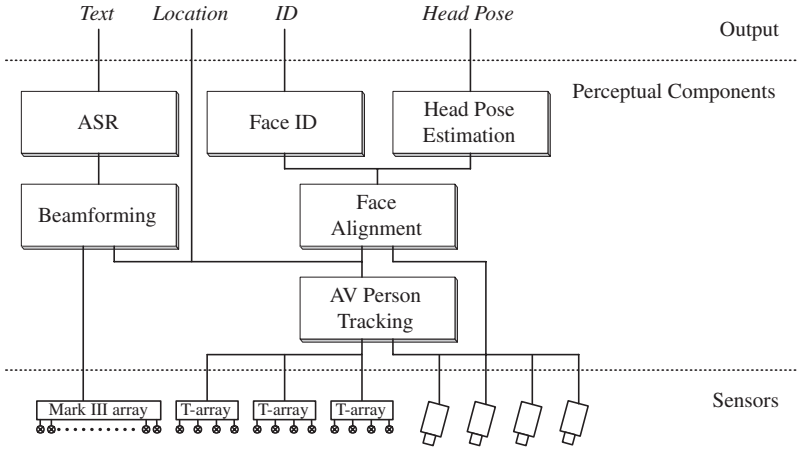

Fig. 1. System overview. Perceptual components are processing input from the sensors and use output from other perceptual components.

On the acoustic side, the beamforming module is analyzing the sensor stream coming from the Mark III MA. Together with the position of the lecturer, given by the person tracking module, it produces a beamformed speech signal which is then used by the ASR module.

In this work we briefly describe each of these perceptual components and present experimental results on several recordings of seminars that took place in our lab since 2003.

In order to allow concurrent parallel access to the various sensor streams as well as to the data streams provided, for example, by the AV-tracking module, all sensors and perceptual components are running on the NIST Smartflow System (NFSF) developed at the US *National Institute of Standards and Technologies* (NIST); see http://www.nist.gov/smartspace/nsfs.html. NSFS is a client–server-based system designed to optimize high-bandwidth data transfer in a network transparent way. Data produced by a client is automatically copied and distributed through shared memory to local recipients or over the network to distant recipients that have subscribed to it. Networks of clients running on the Smartflow system are dynamically reconfigurable, allowing for rapid prototyping and experimenting, run-on activation and deactivation of groups of clients, robust recovery from individual component failure, and easy fusion of multiple information streams.

The remainder of this paper is organized as follows: Section 2 describes the sensor setup in our smart lecture room and the data set we used for experiments. Section 3 describes our approach for audio-visual tracking of a lecturer. Section 4 presents our work on far-field speech recognition in the lecture room. In Section 5, we present an approach to estimate the lecturer's head orientation from multiple cameras. Section 6 introduces our work on recognizing the face of the lecturer from multiple cameras. In Section 7 experimental results of all components on real seminar recordings are presented. In Section 8, we present our conclusions and plans for future work.

2. Sensor setup in the smart room and data set

The data used for the experiments described in this work were collected during a series of seminars held by students and visitors at the Universität Karlsruhe (TH), Germany, since Fall 2003. The subjects spoke English, but mainly with German or other European accents, and with varying degrees of fluency. This data collection was done in a very natural setting, as the students were far more concerned with the content of their seminars, their presentation in a foreign language and the questions from the audience than with the recordings themselves. Moreover, the seminar room is a common work space used by other students who are not seminar participants. Hence, there are many "real-world" events heard in the recordings, such as door slams, printers, ventilation fans, typing, background chatter, and the like.

The seminar speakers were recorded with a Countryman E6 CTM, a 64-channel Mark III MA developed at the NIST mounted on the wall, four T-shaped MAs with four elements apiece mounted on the four walls of the seminar room and three Shure Microflex table-top microphones located on the work table. The positions of the table-top microphones were not fixed. A diagram of the seminar room is given in Fig. 2. All audio files have been recorded at 44.1 kHz with 24 bits per sample. The high sample rate is preferable to permit more accurate position estimations, while the higher bit depth is necessary to accommodate the large dynamic range of the far field speech data. For the recognition process, the speech data was down-sampled to 16 kHz with 16 bits per sample. In addition to the audio data capture, the seminars were simultaneously recorded with four calibrated video cameras that are placed at a height of 2.7 m in the room corners. Their joint field of view covers almost the entire room. The images are captured at a resolution of 640×480 pixels and a framerate of 15 frames per second, and stored as jpg-files for offline processing.

The data from the CTM was manually segmented and transcribed. The data from the far distance microphones were labeled with speech and non-speech regions. The location of the centroid of the speaker's head in the images from the four calibrated video cameras was manually marked every 10th frame (i.e., approximately every 0.7 s). Based on these marks, the true 3D position of the speaker's head, which served as ground truth for our localization experiments, could be calculated with an accuracy of approximately 10 cm [11].

3. Audio-visual lecturer tracking

In our scenario, the task of lecturer tracking poses two basic problems: localizing the lecturer (in terms

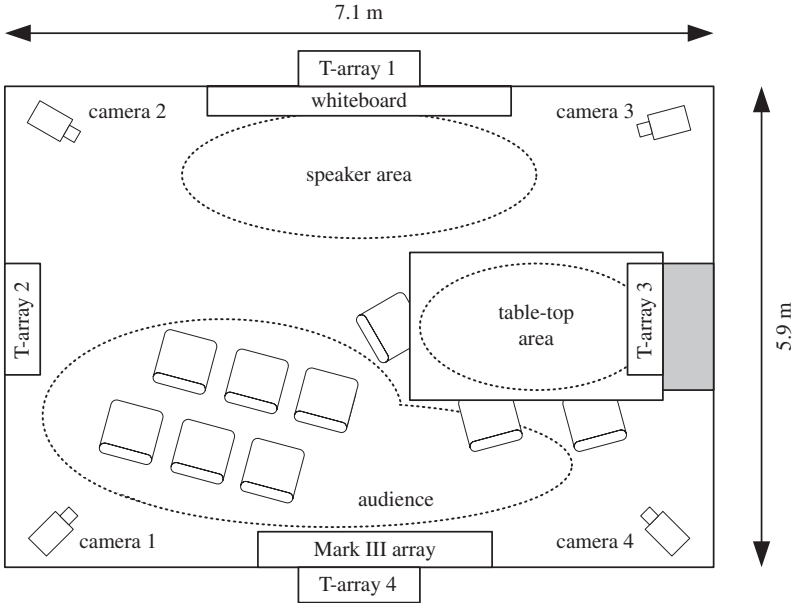

Fig. 2. The CHIL seminar room layout at the Universität Karlsruhe (TH).

of 3D head coordinates) and disambiguating the lecturer from other people in the room. In the proposed approach, we jointly process images from multiple cameras and the signal from multiple microphones in order to track the lecturer. The algorithm is based on the assumption, that the lecturer—among all other people in the room—is the one that is speaking and moving most of the time.

The central issue in audio-visual tracking is the question of how to combine different sensor streams in a beneficial way. In our approach, we integrate audio and video features such that the system does not rely on a single sensor or a certain combination of sensors to work properly. In fact, each single camera and each microphone pair alone can contribute to the track.

We use a particle filter framework [12] to guide the evaluation of the sensor streams and to generate the tracking hypothesis. Particle filters have been shown to be applicable successfully to audio-visual tracking for example by [13] for a video telephony application, by [14] for multiperson tracking or by [15] for multiparty conversation in a meeting situation.

3.1. Tracking using a particle filter

Particle filters represent a generally unknown probability density function by a set of m random samples $s_{1\ldots m}$. Each of these particles is a vector in state space and is associated with an individual weight $\pi_{1\ldots m}$. The evolution of the particle set is a two-stage process which is guided by the observation and the motion model:

(1) *The prediction step*: From the set of particles from the previous time instance, an equal number of new particles is generated. In order to generate a new particle, a particle of the old set is selected at random with a probability that is proportional to its weight, and then propagated by applying the motion model. In the simplest case, this can be additive Gaussian noise, but higher order motion models can also be used.
(2) *The measurement step*: In this step, the weights of the new particles are adjusted with respect to the current observation z_t: $\pi_{i,t} = p(z_t|s_{i,t})$. This means computing the probability of the observation given that the state of particle $s_{i,t}$ is the true state of the system.

As we want to track the lecturer's head centroid, each particle $s_i = (x, y, z)$ represents a coordinate in space. The ground plane is spanned by x and y, the height is represented by z. The particles are propagated by simple Gaussian diffusion, thus representing a coarse motion model. Using the features described in Sections 3.3 and 3.4, we can

calculate a weight for each particle at time t by combining the probability of the acoustical observation A_t and the visual observation V_t using a weighting factor α:

$$p(z_t|s_i) = \alpha \cdot p(A_t|s_i) + (1-\alpha) \cdot p(V_t|s_i). \quad (1)$$

The weighting factor α was adjusted dynamically according to the acoustic confidence measure described in Section 3.4. The average value of α was approximately 0.4, so that more weight was given to the video features.

A particle's weight is set to zero if the particle leaves the lecture room[1] or if its z-coordinate leaves the valid range for a standing person ($1.2m < z < 2.1m$). The final hypothesis about the lecturer's location over the whole particle set $s_{1\ldots m}$ (in our case $m = 300$) can be derived by a weighted summation over the individual particles at time t:

$$\Lambda_t = \frac{1}{\sum_{i=1}^{m} \pi_{i,t}} \sum_{i=1}^{m} \pi_{i,t} \cdot s_{i,t}. \quad (2)$$

3.2. Sampled projection instead of triangulation

A common way to infer the 3D position of an object from multiple views is to locate the object in each of the views and then to calculate the 3D position by using triangulation [16]. This approach, however, has several weak points: Firstly, the object has to be detected in at least two different views at the same time. Secondly, the quality of triangulation depends on the points of the object's images that are chosen as starting points for the lines of views: if they do not represent the same point of the physical object, there will be a high triangulation error. Furthermore, searching for the object in each of the views separately without incorporating geometry information results in an unnecessarily large search space.

In the proposed method, we followed an approach also taken by [17] and avoid the problems mentioned above by not using triangulation at all. Instead, we make use of the particle filter's property to predict the object's location as a well-distributed set of hypotheses: many particles cluster around likely object locations, and fewer particles populate the space between them. As the particle set represents a probability distribution of the predicted object's location, we can use it to narrow down the search space. So instead of searching a neighborhood exhaustively, we only look for the object at the particles' positions.

Note that it is also possible to avoid the problem of triangulation in a Kalman filter framework, as we have shown in other work [18].

3.3. Speaker localization: video features

For the task of person tracking in video sequences, there are a variety of features to choose from. In our lecture scenario, the problem comprises both locating the lecturer and disambiguating the lecturer from the people in the audience. A snapshot from a lecture showing all four camera views is shown in Fig. 3. As lecturer and audience cannot be separated reliably by means of fixed spatial constraints as, e.g., a dedicated speaker area, we have to look for features that are more specific for the lecturer than for the audience.

Intuitively, the lecturer is the person who is standing and moving (walking, gesticulating) most, while people from the audience are generally sitting and moving less. In order to exploit this specific behavior, we decided to use dynamic foreground segmentation based on adaptive background modeling as primary feature; a detailed explanation can be found in [19]. In order to support the track indicated by foreground segments, we use detectors for face and upper body. Both features (foreground F and detectors D) are linearly combined using a mixing weight β (for our experiments β was fixed to 0.7, this value was optimized on a development set), so that the particle weights for view j are given by

$$\overline{p}(V^j|s_i) = \beta \cdot p(D^j|s_i) + (1-\beta) \cdot p(F^j|s_i). \quad (3)$$

To combine the different views, we sum over the weights from the v different cameras in order to obtain the total weight of the visual observation for the particular particle

$$\overline{p}(V|s_i) = \frac{1}{v} \sum_{j=1}^{v} \overline{p}(V^j|s_i). \quad (4)$$

To obtain the desired (pseudo) probability value which tells us how likely this particle corresponds to the visual observation we have to normalize over all particles

$$p(V|s_i) = \frac{\overline{p}(V|s_i)}{\sum_i \overline{p}(V|s_i)}. \quad (5)$$

[1] We restrict the particles to be within the full width of the room's ground plane ($0 < y < 7.1m$) and half of the depth ($0 < x < 3m$).

Fig. 3. Snapshot from a lecture showing all four camera views (clockwise: cameras 1,2,3,4). The native resolution is 640 × 480 pixels.

3.4. Speaker localization: audio features

As the lecturer is usually the person speaking, audio features coming from multiple microphones can be used to detect his position.

Consider the jth pair of microphones, and let $\mathbf{m}_{j1}$ and $\mathbf{m}_{j2}$, respectively, be the positions of the first and second microphones in the pair. The ith particle $\mathbf{s}_i$ denotes the position of the speaker in a 3D space. Then the *time delay of arrival* (TDOA) between the two microphones of the pair can be expressed as

$$T_j(\mathbf{s}_i) = T(\mathbf{m}_{j1}, \mathbf{m}_{j2}, \mathbf{s}_i) = \frac{\|\mathbf{s}_i - \mathbf{m}_{j1}\| - \|\mathbf{s}_i - \mathbf{m}_{j2}\|}{c}, \tag{6}$$

where c is the speed of sound. To estimate the TDOAs, a variety of well-known techniques [20,21] exist. Perhaps the most popular method is the *phase transform* (PHAT), which can be expressed as

$$R_{12}(\tau) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \frac{X_1(e^{j\omega\tau})X_2^*(e^{j\omega\tau})}{|X_1(e^{j\omega\tau})X_2^*(e^{j\omega\tau})|} e^{j\omega\tau}\, d\omega, \tag{7}$$

where $X_1(e^{j\omega})$ and $X_2(e^{j\omega})$ are the Fourier transforms of the signals of a microphone pair in a MA. Normally, one would search for the highest peak in the resulting cross correlation to estimate the position. But since we are using a particle filter, as described in Section 3.1, we can simply set the PHAT value at the time delay position $T_j(\mathbf{s}_i)$ of the MA pair j of a particular particle s_i as

$$\overline{p}(A^j|\mathbf{s}_i) = \max(0, R_j(T_j(\mathbf{s}_i))). \tag{8}$$

As the values returned by the PHAT can be negative, but probability density functions must be strictly non-negative, we found that setting all negative values of the PHAT to zero yielded the best results.

To get a better estimate we repeat this over all m pairs of microphones (in our case 12), sum their values and normalize by m:

$$\overline{p}(A|\mathbf{s}_i) = \frac{1}{m}\sum_{j=1}^{m} \overline{p}(A^j|\mathbf{s}_i). \tag{9}$$

Just like for the visual features, we normalize over all particles in order to get the acoustic observation likelihood for each particle:

$$p(A|\mathbf{s}_i) = \frac{\overline{p}(A|\mathbf{s}_i)}{\sum_i \overline{p}(A|\mathbf{s}_i)}. \tag{10}$$

The weighting factor α used in Eq. (1) was set as

$$\alpha = \frac{m_0}{m} \cdot 0.6, \tag{11}$$

where m is the total number of microphone pairs and m_0 the number of PHAT values above 0. The

maximum weighting factor of 0.6 for the audio features has been determined experimentally.

4. Far field speech recognition

Interest within the ASR research community has recently focused on the recognition of speech where the microphone is located in the medium field, rather than being mounted on a headset and positioned next to the speaker's mouth. Using a MA can improve the performance over a single microphone on the signal-to-noise ratio as well as on WER if the speaker's location is known or can be estimated as described in previous sections.

The CHIL seminar data comprised of spontaneous speech lectures and oral presentations collected by both near and far-field microphones present significant challenges to both modeling components used in ASR, namely the language and acoustic models. With respect to the former, the currently available CHIL data primarily contains seminars on technical topics. Recognizing such speech is a very specialized task since it contains many acronyms and therefore is quite mismatched to typical language models (LM) currently used in the ASR literature. Furthermore, large portions of the data contain spontaneous, disfluent, and interrupted speech, due to the interactive nature of seminars and the varying degree of the speakers' comfort with their topics. On the acoustic modeling side, the seminar speakers exhibit moderate to heavy German or other European accents in their English speech. The above problems are compounded by the fact that, at this early stage of the CHIL project, not enough data are available for training new language and acoustic models matched to this seminar task, and thus one has to rely on adapting existing models that exhibit gross mismatch to the CHIL data. Clearly, these challenges present themselves in both CTM data, as well as the far-field data, where of course they are exacerbated by the much poorer quality of the acoustic signal.

In this section, we investigate the influence of the speaker position on the WER of an ASR system after beamforming, and compare it to speech recognition using a CTM.

4.1. Beamforming

In this work, we used a simple delay and sum beamformer implemented in the subband domain. Subband analysis and resynthesis were performed with a cosine modulated filter bank [22, Section 8]. In the complex subband domain, beamforming is equivalent to a simple inner product

$$y(\omega_k) = \mathbf{v}^H(\omega_k)\mathbf{X}(\omega_k),$$

where ω_k is the center frequency of the kth subband, $\mathbf{X}(\omega_k)$ is the vector of subband inputs from all channels of the array, and $y(\omega_k)$ is the beamformed subband output. The speaker position comes into play through the *array manifold vector* [23, Section 2]

$$\mathbf{v}^H(\omega_k) = [e^{j\omega_k\tau_0(\mathbf{x})}\ e^{j\omega_k\tau_1(\mathbf{x})}\ \cdots\ e^{j\omega_k\tau_{N-1}(\mathbf{x})}],$$

where $\tau_i(\mathbf{x}) = \|\mathbf{x} - \mathbf{m}_i\|/s$ is the propagation delay for the ith microphone located at $\mathbf{m}_i$. After beamforming, the subband samples are synthesized back into a time domain signal, and then downsampled to 16 kHz for ASR purposes.

4.2. Language model training

To train LM for interpolation, we used corpora consisting of broadcast news (160M words), *proceedings* (17M words) of conferences such as ICSLP, Eurospeech, ICASSP or ASRU and *talks* (60k words) by the Translanguage English Database. Our final LM was generated by interpolating a 3-g LM based on broadcast news and proceedings, a class-based 5-g LM based on broadcast news and proceedings and a 3-g LM based on the talks. The perplexity is 144 and the vocabulary contains 25,000 words plus multiwords and pronunciation variants.

4.3. Acoustic model training

As relatively little transcribed data are available for acoustic model training, we used the *Broadcast News* [24] corpus along with the close talking channel of meeting corpora [25], to provide a total of 300 h of training material.

The speech data was sampled at 16 kHz. Speech frames were calculated using a 10 ms Hamming window. For each frame, 13 *Mel-minimum variance distortionless response* (Mel-MVDR) cepstral coefficients were obtained through a discrete cosine transform (DCT) from the Mel-MVDR spectral envelope [26]. Thereafter, linear discriminant analysis was used to reduce the utterance-based cepstral mean normalized features plus seven adjacent to a final feature number of 42. Our baseline model consisted of 300,000 Gaussians with diagonal

covariances organized in 24,000 distributions over 6000 codebooks.

4.4. Acoustic adaptation

To compensate for the mismatch between the training and test condition we have adapted the far distance acoustic model in consecutive steps:

(1) Four iterations of Viterbi training on far distance data from NIST [27] and ICSI [4] over all channels on top of the acoustic trained models to better adjust the acoustic models to far distance.
(2) A supervised MLLR in combination with FSA and VTLN on the far distance (single distance or MA processed) CHIL development set: this step adapts to the speaking style of the lecturer and the channel (in particular to the room reverberation). In the case of non-native speakers the adaptation should also help to cover some non-native speech.
(3) A second, now unsupervised MLLR, FSA and VTLN adaptation based on the hypothesis of the first recognition run: this procedure aims at adapting to the particular speaking style of a speaker and to changes within the channel.

The acoustic model of the reference close talking channel was adapted in a similar manner; see [25] for further details.

5. Multiview head pose estimation

Tracking a lecturer's head orientation can give valuable cue to determine his or her focus of attention. This could be useful to index seminar recordings, to detect context switches such as interruptions, discussions, etc. and in particular to tell the "smart room" about the lecturer's focus of attention; e.g., the audience, a whiteboard, a laptop, etc.

In this section, we present our approach to estimating a lecturer's head orientation. By using multiple cameras we cover the entire room and are able to combine head pose estimates coming from various camera views to form a more robust hypothesis.

To estimate head pose in each view, we use an appearance-based approach as proposed in [28,29], as it has proven to provide useful results even from low-resolution facial images such as the ones captured with the smart room cameras.

The work presented here extends our previous work, which was based on single cameras only, in that it fuses the estimates from multiple cameras, thus improving robustness and allowing for coverage of the whole smart room. Since head pose is initially estimated with respect to each camera, our approach is flexible and allows for easy change of camera positions and use of additional cameras without the necessity of retraining the neural networks for pose estimation.

5.1. Head alignment

Since the estimated position of the lecturer given by the tracking module does not provide consistently aligned bounding boxes of the face, a further face alignment step becomes necessary before faces can be extracted for later processing.

In order to align and extract the lecturer's face in each camera view, we use a frontal and profile face detector which are based on Haar-feature cascades as proposed in [30]. The search space for these face detectors is limited to a window around the initially estimated position of the lecturer and projected into the respective camera view as shown in Fig. 4. The big boxes around the lecturer's head depict the search windows for the face detectors.

Since the face detectors sometimes fail to detect a face, we predict the face bounding boxes in those camera views, in which the lecturer's face could not be detected. This can be done if the face was detected in at least two other camera views. From those detected faces, we compute the lecturer's 3D position by triangulation and project a 3D cuboid around the 3D head location into those camera views where no face was detected.

If face detection errors cause more than one face to be detected in a camera view, those face bounding boxes that lead to the minimal triangulation residual are chosen as the correct.

5.2. Frontal vs. back of the head classification

In our experiments, we observed that neural networks for head pose estimation performed worse if views of the back of a head (showing hair only) were included in the training data set. This is why we try to automatically detect back-views. To do this, we trained a neural network classifier which outputs the a posteriori probability that a given

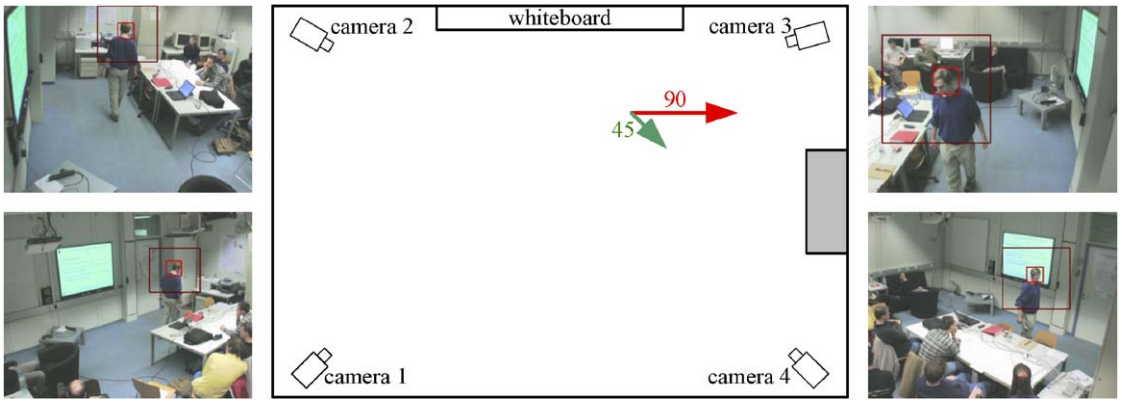

Fig. 4. Example output of our discrete head pose estimation system. The arrows indicate the final head pose estimation (long red arrow) and the groundtruth head pose (short green arrow). Further, the position of the arrows indicate the position of the user in the smart-room.

image depicts a frontal view in the range from left to right profile $[-90°, +90°]$. We use a three-layered, feed-forward network, trained with frontal views and views of the head's back only. For the latter the target output was defined to be 0, else 1. Finally, we use a likelihood threshold of 0.5 above which all captures are classified as (near-) frontal views of the head. As input to the neural net, a histogram-normalized grayscale image of the head as well as horizontal and vertical edge images, downsampled to 16×16 pixels each, are used.

The network was trained using standard error backpropagation, minimizing the output error on a cross evaluation set within 100 training cycles.

5.3. Single-view head pose estimation

In our system we first try to estimate the lecturer's head orientation relative to each camera in the range of $[-90°, +90°]$. Doing the estimation relative to each camera, instead of the world coordinate system, allows us to train and use only one single neural network to estimate head pose for all cameras. This has the advantage that all available facial images from all cameras can be used for training of the network. This also makes our system independent of the positioning of the cameras in the room and allows us to add further cameras without the necessity of retraining networks.

We follow our previous work [28,31] using a three-layered, feed-forward network with one single output unit, delivering continuous head pose estimation in the range of $[-90°, +90°]$. As input images, we used downsampled histograms of normalized grayscale images once more, as well as horizontal and vertical edge images of heads.

The network is trained with standard error back-propagation, using a data set that consists of frontal views of the head only, ranging from left to right profile. As noted above, we experienced a more robust performance of the single view system by limiting its training data and output to the $[-90°, +90°]$ range.

5.4. Building the joint hypothesis

We define $\Theta = \{\theta_i\}$, with $\theta_i \in \{0°, 45°, \ldots, 315°\}$ as the set of all possible head pose classes, relative to the room coordinate system. Moreover, at each timestamp we have $H = \{h_1, h_2, \ldots, h_n\}$, the set of all single estimations made, where n represents the number of cameras used.

In making a final decision about the true head pose, we score a pose hypothesis by summing up the a posteriori probabilities of all available estimations as follows:

$$\pi(\theta_i) = \sum_{j=1}^{n} P(\theta_i|h_j). \tag{12}$$

Finding the best-fitting hypothesis $\hat{\theta}$ then consists in maximizing the score:

$$\hat{\theta} = \arg\max_{\theta_i \in \Theta} \pi(\theta_i). \tag{13}$$

The a posteriori probabilities in (12) are derived from confusion matrices that were built for each camera when evaluating the classification performance of the trained neural network on the cross evaluation set. Here, the posterior probability of a class θ_i given the observation h_j can be computed as

$$P(\theta_i|h_j) = \frac{k_{ij}}{\sum_m k_{mj}}, \tag{14}$$

where k_{ij} denotes the matrix element in row i and column j. While the matrix columns define the different estimation classes and the rows describe the ground truth head pose classes.

As the a posteriori probabilities $P(\theta_i|h_j)$ are added in (12) instead of multiplied, the hypothesis scores are guaranteed to increase when more camera views are used. It is advantageous to use as many cameras as possible to stabilize the estimation.

6. Face recognition

Although face recognition under controlled laboratory conditions has been successfully addressed in the past, face recognition in unconstrained environments remains a challenging problem of significant research interest. A face recognition system should be robust against detection and alignment errors, insensitive to illumination and background variations and easily extendible to detect and recognize unknown persons. Furthermore, it should naturally weight the contributions of the frames from multiple cameras for face classification. With these requirements in mind, we developed a novel face recognition algorithm. To minimize sensitivity to illumination and background variations, the face appearance is modeled locally. That is, the detected and resized face is divided into 8×8 pixel resolution blocks and each block is represented with DCT coefficients. The DCT was chosen for its compact representation capability, fast computation and data-independent nature. Although the paradigm of local appearance-based face recognition can also benefit from other data-dependent or independent basis functions, data-independent bases are preferred, since there is no alignment step involved during training for extracting proper bases, as in the case of principal component analysis (PCA). Furthermore, in [32], it is shown that local appearance-based face recognition using DCT performs better than well-known holistic approaches, such as [33–35], and local appearance representation using PCA [36]. To achieve robustness against detection and alignment errors, artificial samples are generated from the original training face images by translating and scaling them; the artificial data generation process is not limited to only translation or scale, it can perform other variations like illumination, face rotation, view morphing, etc. To increase discrimination between candidate individuals and to provide robustness against false detections (detecting background as a face), in the classification step, a two-class linear discriminant analysis is performed. In this approach a single M-class linear discriminant classifier is divided into M two-class linear discriminant classifiers. The training data for the genuine class consists of samples from the true candidate, whereas the training data for the impostor class consists of the other people's samples plus random background samples. With this method, each class has its own $N \times 1$ projection vector, where N denotes the size of the feature vector. When a test image arrives, it is projected onto each individual's decision space, using the corresponding projection vector. The distribution of projected genuine and impostor data in 1D space is modeled with univariate Gaussians. The decision is taken by applying Bayes rule

$$P(C_{k,1}|x) = \frac{P(x|C_{k,1})P(C_{k,1})}{\sum_{i=1}^{2} P(x|C_{k,i})P(C_{k,i})},$$

where $C_{k,1}$ denotes the genuine class and $C_{k,2}$ denotes the impostor class of the kth individual. $P(C_{k,1})$ and $P(C_{k,2})$ are set to 0.5. From the equation above, three cases may be observed. In the first case, for every k, $P(C_{k,1}|x)$ may be smaller than 0.5. This can occur either from a background sample detected as a face or from an unknown face. In the second case, there may be more than one k such that $P(C_{k,1}|x)$ is bigger than 0.5. In this case the most probable candidate can be selected. In the third, ideal case, $P(C_{k,1}|x)$ is greater than 0.5 for only one individual.

The proposed classification approach has been tested on still images extensively, and compared with the traditional multiclass LDA-based classification scheme in [37]. It is found that the classification using two-class LDA provides significantly better results than classification using multiclass LDA. Besides the performance advantage, the proposed method has many others like simpler calculation of projection vectors, more discrimination between classes, easier update of the database with new individuals and inherent detection of unknown people [37].

The extension of this system to multiple cameras and to video sequences is straightforward; we need only accumulate the $P(C_{k,1}|x)$'s that are greater than 0.5 over multiple camera views and frames. By doing this, the temporal and multiview information is incorporated naturally, and good frames—i.e., the frames which contain properly detected, high-

resolution images—are inherently assigned a higher weight.

Overall, the proposed face recognition algorithm improves both the representation and the classification steps. In the following, we provide a summary of the algorithm.

Training:

(1) Artificially generate data to account for improper detection and alignment of faces, illumination variations, etc.
(2) Perform local appearance modeling. Perform DCT on 8×8 pixels blocks and choose only the DCT coefficients that contain more information by zig-zag scanning the DCT image.
(3) Perform class-specific projection for each class.
(4) Model the distribution of projected genuine and impostor data with univariate Gaussians.

Testing:

(1) Perform local appearance modeling (as in training step 2).
(2) Project the extracted feature vector onto each class.
(3) Use Bayes' rule to obtain the matching score for each class.
(4) Accumulate scores over multiple cameras and frames whenever the score is higher than 0.5.
(5) Choose the candidate with the highest score.

7. Experiments

In order to evaluate the performance of the various components described in this work, we ran experiments on several seminar recordings that were recorded in our smart room as described in Section 2.

7.1. Audio-visual tracking

The experiments for audio-visual tracking and speech recognition were performed on five seminars/speakers providing recordings of 130 min length, including 16,395 words.

The error measure used for tracking is the average Euclidean distance between the hypothesized head coordinates and the labeled ones.

It can be seen in Table 1, that even though the video only tracker performs considerably better than the audio only tracker, the performance can still be significantly increased by combining both modalities. This effect is especially pronounced during one recording in which the lecturer is standing most of the time in one dark corner of the room. For this seminar, tracking the speaker only by means of video features proved very difficult and resulted in a mean error of 116 cm. While the video only tracker exhibits the same performance in all frames, the precision of the audio only and the combined tracker is higher for the frames where speech is present.

Table 1
Averaged error in 3D head position estimation of a lecturer over all frames (approximately 130 min) and frames where speech was present (approximately 105 min)

Tracking mode	Average error (cm)	
	All frames	Speech frames
Audio only	46.1	41.7
Video only	36.3	36.5
Video and Audio	30.5	29.1

7.2. Beamforming and speech recognition

The speech recognition experiments described below were conducted with the *Janus recognition toolkit* (JRTk), which was developed and is maintained jointly by the Interactive Systems Laboratories at the Universität Karlsruhe (TH), Germany and at the Carnegie Mellon University in Pittsburgh, USA. All test documented here used the language and acoustic models described in Section 4. The experiments were performed on the same seminar data used for the testing the tracking components, as described in Section 2.

As mentioned before, the main advantage offered by a MA is the relatively large reduction in WER over a single channel, as can be seen in comparing the figures in Table 2. To make this relation more apparent, we have plotted the average position error of the source localization to the WER in Fig. 5. If the error of the labeled position to the ground truth is around 15 cm (our calculation of the accuracy is approximately 10 cm), then a linear relationship can be seen. Indeed, by using a simple beamforming algorithm to combine the channels of the MA using an estimated speaker position, we gained back 26.9% of the degradation in going from the CTM to a single channel of the MA.

Table 2
Word error rates (WERs) for a close talking microphone and a single microphone of the array and the microphone array with different position estimates

Tracking mode	WER (%)
Close talking microphone	34.0
Microphone array	
Single microphone	66.5
Estimated position (Audio only)	59.8
Estimated position (Video only)	59.1
Estimated position (Audio and video)	58.4
Labeled position	55.8

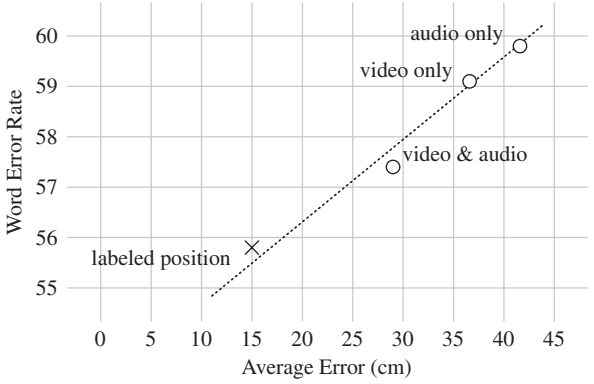

Fig. 5. Plot comparing the average position error to the word error rate.

7.3. Head pose estimation

Considering the educational smart-room scenario we already described earlier, we evaluated our implementation on real videos that were recorded during a seminar lecture in 2003. Overall we recorded seven persons, further splitting each recording into four segments of 5 min each, on which training and evaluation was realized separately.

For providing ground truth information regarding the true head pose, we manually annotated the videos with the observed head pose of the lecturer, classifying the head's pose manually into one of eight equidistant classes such as $0°, 45°, 90°, \ldots, 315°$.

In the multiuser scenario, we trained the underlying neural networks on segments 1 and 2, using segment 3 as cross evaluation set. Segment 4 was used for evaluation purposes, thus evaluating the networks with video data that has not been seen before in the training stage, though resulting from the same persons. In the unknown user scenario, we implemented a round robin evaluation, thus excluding a person's recording from training and cross evaluation when evaluation is being done on this person's video data.

Table 3
Systems performance on the head pose estimation task

	Correct class (%)	Correct or neighbor class (%)
Multiuser manual view selection	74.6	96.4
Multiuser automatic view selection	58.9	91.7
Unknown user automatic view selection	48.4	82.9

Table 3 summarizes the results. In the multiuser scenario, head orientation performed correctly with approximately 59% in our fully automatic scenario. This means, the networks were evaluated using unsupervised head extractions, thus including outliers and variance resulting from imperfect alignment of the corresponding bounding box. In this case, with the use of our earlier described head position tracking module, classifying frontal views of the head performed with an accuracy of 83.5%.

In case of manually annotated 3D positions of the head's centroid and manual removal of extreme outliers, the performance increased to approximately 74% correct detection of the pose class thus showing the impact of imperfect face detections and outliers in the complete system.

In the unknown user scenario, correct pose class detection was achieved 48.4% of the time. Here, the initial facial view recognition step achieved a correct recognition rate of 79.6%. In 82.9% of the trials, the estimated pose class fell in either the correct class or a neighboring class; i.e., the error was less than 45°. Although the performance obtained was worse than in the multiuser case, we can see that—as in the multiuser case—the performance increases as more face views are available for pose classification; see Table 4. Moreover, the results indicate that it may be advantageous to train the system with much more data in order to increase the networks' generalization capability on unseen people.

7.4. Face recognition

The objective of the face recognition task was to determine which lecturer from a known set spoke

during a given seminar. To evaluate performance on this task, five frames of frontal face images were selected from a development set for speaker enrollment. Some enrollment images are shown in Fig. 6. For testing, a total of 20 uniformly sampled, non-overlapping sequences of 100 frames apiece were selected from each camera. Fig. 7 shows some test images from each camera. A sequence of images from a single camera is shown in Fig. 8. The classification was performed over the 100 frame sequences. In the proposed system, the detection of faces was done automatically using only a frontal face detector. In the case of multiple face detections caused by people in the audience, the face rectangle that was closest to the face tracking estimate was chosen. The detected faces were scaled to a resolution of 40×32 pixels.

Table 4
Correct classification in percent in case of both a multiuser and an unknown user scenario

	1 frontal view (%)	2 frontal views (%)	3 frontal views (%)	Avg. (%)
Multiuser	37.5	58.3	72.5	58.9
Unknown users	27.3	55.2	55.3	48.4

In both cases, using more frontal views at the head enhances the system's performance.

During system training, the face images were roughly cropped by hand from the images; no alignment was performed. The training images were translated in x-, y-directions and scaled using five different coefficients. This way, 125 training images were derived from a single training image. In total, for each individual 625 training images were generated and used. For each individual, these 625 face images were used as genuine samples and the $6 \times 625 = 3750$ face images of the other known lecturers as well as 1400 images randomly selected

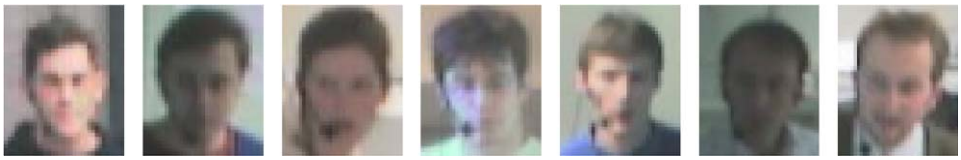

Fig. 6. Sample training face images.

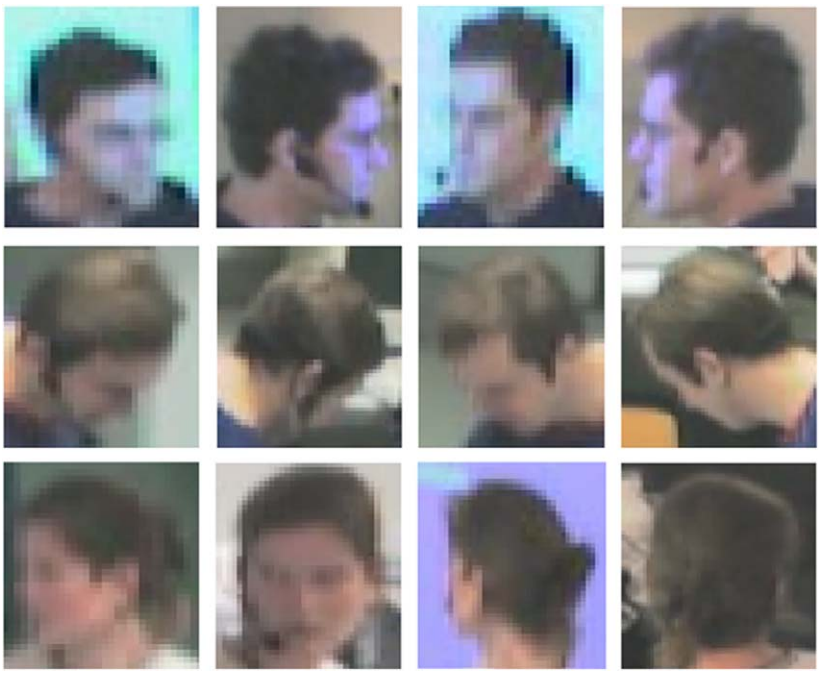

Fig. 7. Test samples at the same instant from different cameras (from left to right: cameras 1–4).

Fig. 8. Sub-sampled testing sequence from a single camera.

from the background were used as impostor samples. For each class, a two-class linear discriminant analysis was performed and a projection vector was computed.

Our system achieved a correct recognition rate of 65.4%. As classification is performed only with seven classes, this result may seem low at first blush. It should be kept in mind, however, that the classification task at hand is very difficult. One difficulty arose from single-view-to-multiview face recognition; only frontal face images were used during training, but all views were used for testing, and some 100 frame sequences contained no frontal face images; see Figs. 7 and 8. The high variation in illumination conditions also posed a challenge. The variations caused by the projector's beam and illumination sources in the room resulted in a decrease in performance. A final source of difficulty was the poor resolution of the faces of the subjects; the average face resolution was approximately 30×30 pixels.

8. Conclusions

Intelligent perceptive environments, such as smart meeting rooms or lecture rooms have been a very active research area for many years and are still in the focus of many large research projects worldwide.

In this paper we have presented several key components for audio-visual analysis of activities in a smart lecture room. More specifically, we have presented a system that tracks the location of a lecturer using acoustic and visual cues, that tracks his head pose, identifies his face and recognizes his speech. We have described in detail the overall system architecture, as well as each of the perceptual components, and presented experimental results on a number of multimodal recordings of seminars in our smart room.

In our system, audio-visual localization of the lecturer plays a key role: it is used by the components for face recognition and head pose estimation in order to restrict the search area for the lecturer's face to his location in the room. Localization of the lecturer is also used to improve ASR using distant microphones, by focusing on only the acoustic signal coming from the lecturer. Compared to using only a single remote microphone, the WER of our speech recognizer could be reduced from 66.5% to 58.4% when using a MA. The results of the experiments presented here have even shown that there is a direct relation between the accuracy of localization and the WER obtained from speech recognition.

The seminar recordings used for the experiments presented in this work provide a very interesting and challenging scenario. The challenges include coping with many people in the room, occlusions, poor image resolution, varying head pose, changing illumination conditions, moving speakers and far-field speech recognition.

All information that can be automatically gained about the lecturer using the proposed system provides valuable cues for annotating and indexing multimedia recordings of seminars, as well as building online proactive services supporting the lecturer and students in a classroom.

In our future research we will also investigate how head pose estimation can be used to improve the acoustic localization of the lecturer. In addition, the recognized identity of the lecturer could be used to select person-dependent acoustic models, which should lead to a further improvement in speech recognition. We will also aim at analyzing the identities, head pose, attention and activities of the audience in seminars. Finally, we plan to apply

these technologies to meeting analysis in the smart room.

Acknowledgments

This work has been partially funded by the European Commission under contract no. 506909 within the project CHIL (http://chil.server.de).

References

[1] B. Brumitt, B. Meyers, J. Krumm, A. Kern, S. Shafer, Easyliving: technologies for intelligent environments, in: Handheld and Ubiquitous Computing, September 2000.

[2] G.D. Abowd, C. Atkeson, A. Feinstein, C. Hmelo, R. Kooper, S. Long, N. Sawhney, M. Tani, Teaching and learning as multimedia authoring: the classroom 2000 project, in: Proceedings of the ACM Multimedia'96 Conference, November 1996, pp. 187–198.

[3] I. McCowan, S. Bengio, D. Gatica-Perez, G. Lathoud, F. Monay, D. Moore, P. Wellner, H. Bourlard, Modeling human interaction in meetings, in: ICASSP, 2003.

[4] A. Janin, J. Ang, S. Bhagat, R. Dhillon, J. Edwards, J. Macias-Guarasa, N. Morgan, B. Peskin, E. Shriberg, A. Stolcke, C. Wooters, B. Wrede, The icsi meeting project: resources and research, in: NIST 2004 Meeting Recognition Workshop, Montreal, May 2004 (AMI-06).

[5] The NIST smart space project. ⟨http://www.nist.gov/smartspace/⟩.

[6] C.D. Kidd, R.J. Orr, G.D. Abowd, C.G. Atkeson, I.A. Essa, B. MacIntyre, E. Mynatt, T.E. Starner, W. Newstetter, The aware home: a living laboratory for ubiquitous computing research, in: International Workshop on Cooperative Buildings—CoBuild'99.

[7] E.M. Tapia, S.S. Intille, K. Larson, Activity recognition in the home setting using simple and ubiquitous sensors, in: Proceedings of PERVASIVE 2004, Lecture Notes in Computer Science, vol. 3001, Springer, Berlin, 2004, pp. 158–175.

[8] M. Argyle, Social Interaction, Methuen, London, 1969.

[9] P.P. Maglio, T. Matlock, C.S. Campbell, S. Zhai, B.A. Smith, Gaze and speech in attentive user interfaces, in: Proceedings of the International Conference on Multimodal Interfaces, Lecture Notes in Computer Science, vol. 1948, Springer, Berlin, 2000.

[10] B. Brumitt, J.J. Cadiz. Let there be light: examining interfaces for homes of the future, in: Proceedings of Interact'01, pp. 375–382. Also available as Microsoft Technical Report TR-2000-92.

[11] D. Focken, R. Stiefelhagen, Towards vision-based 3-d people tracking in a smart room, in: International Conference on Multimodal Interfaces, Pittsburgh, PA, IEEE, October 2002, pp. 400–405.

[12] M. Isard, A. Blake, Condensation-conditional density propagation for visual tracking, Internat. J. Comput. Vision 29 (1) (1998) 5–28.

[13] J. Vermaak, M. Gangnet, A. Blake, P. Pérez, Sequential Monte Carlo fusion of sound and vision for speaker tracking, in: Proceedings of the IEEE International Conference on Computer Vision, vol. 1, 2001, pp. 741–746.

[14] N. Checka, K. Wilson, V. Rangarajan, T. Darrell, A probabilistic framework for multi-modal multi-person tracking, in: IEEE Workshop on Multi-Object Tracking (in Conjunction with CVPR), 2003.

[15] D. Gatica-Perez, G. Lathoud, I. McCowan, J.-M. Odobez, A mixed-state i-particle filter for multi-camera speaker tracking, in: Proceedings of the IEEE ICCV Workshop on Multimedia Technologies in E-Learning and Collaboration (ICCV-WOMTEC), 2003.

[16] I. Mikic, S. Santini, R. Jain, Tracking objects in 3d using multiple camera views, in: ACCV, 2000.

[17] D. Zotkin, R. Duraiswami, L. Davis, Joint audio-visual tracking using particle filters, EURASIP J. Appl. Signal Process. 2002 (11) (2002).

[18] T. Gehrig, K. Nickel, H.K. Ekenel, U. Klee, J. McDonough, Kalman filters for audio–video source localization, in: IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, New Paltz, New York, 2005.

[19] K. Nickel, T. Gehrig, R. Stiefelhagen, J. McDonough, A joint particle filter for audiovisual speaker tracking, in: International Conference on Multimodal Interfaces, ICMI, Trento, Italy, ACM, New York, 2005.

[20] M. Omologo, P. Svaizer, Acoustic event localization using a crosspower-spectrum phase based technique, in: Proceedings of the ICASSP, vol. II, 1994, pp. 273–276.

[21] J. Chen, J. Benesty, Y.A. Huang, Robust time delay estimation exploiting redundancy among multiple microphones, IEEE Trans. Speech Audio Proc. 11 (6) (November 2003) 549–557.

[22] P.P. Vaidyanathan, Multirate Systems and Filter Banks, Prentice-Hall, Englewood Cliffs, 1993.

[23] H.L. Van Trees, Optimum Array Processing, Wiley-Interscience, New York, 2002.

[24] Linguistic Data Consortium (LDC), English broadcast news speech (Hub-4) ⟨www.ldc.upenn.edu/Catalog/LDC97S44.html⟩.

[25] M. Wölfel, K. Nickel, J. McDonough, Microphone array driven speech recognition: influence of localization on the word error rate, Workshop on Multimodal Interaction and Machine Learning, 2005.

[26] M. Wölfel, J. McDonough, Minimum variance distortionless response spectral estimation, review and refinements, IEEE Signal Process. Magazine 22 (5) (September 2005) 117–126.

[27] V. Stanford, C. Rochet, M. Michel, J. Garofolo, Beyond close-talk—issues in distant speech acquisition, conditioning classification and recognition, in: ICASSP 2004 Meeting Recognition Workshop, 2004.

[28] R. Stiefelhagen, J. Yang, A. Waibel, Simultaneous tracking of head poses in a panoramic view, in: International Conference on Pattern Recognition, vol. 3, September 2000, pp. 726–729.

[29] R. Stiefelhagen, Tracking focus of attention in meetings, in: International Conference on Multimodal Interfaces, Pittsburgh, PA, IEEE, October 2002, pp. 273–280.

[30] P. Viola, M. Jones, Robust real-time object detection, in: ICCV Workshop on Statistical and Computation Theories of Vision, 2001.

[31] M. Voit, K. Nickel, R. Stiefelhagen, Multi-view head pose estimation using neural networks, in: Second Workshop on Face Processing in Video (FPiV'05), Victoria, BC, Canada, May 2005.

[32] H. Ekenel, R. Stiefelhagen, Local appearance based face recognition using discrete cosine transform, in: EUSIPCO'05, 2005.
[33] M.S. Bartlett, et al., Face recognition by independent component analysis, IEEE Trans. Neural Networks 13 (6) (2002) 1450–1454.
[34] P.N. Belhumeur, et al., Eigenfaces vs. fisherfaces: recognition using class specific linear projection, IEEE Trans. Pattern Anal. Machine Intell. 19 (7) (1997) 711–720.
[35] M. Turk, A. Pentland, Eigenfaces for recognition, J. Cognitive Sci. (1991) 71–86.
[36] R. Gottumukkal, V.K. Asari, An improved face recognition technique based on modular PCA approach, Pattern Recognition Lett. 25 (4) (2004).
[37] H. Ekenel, R. Stiefelhagen, Two-class linear discriminant analysis for face recognition, Technical Report, University of Karlsruhe, 2005.