Robust Omni-directional Multi-cue Tracking for Multiple Person Meeting Scenarios

Sascha Schreiber and Gerhard Rigoll

Abstract

This paper addresses the problem of tracking an unknown number of humans in indoor environments with a monocular camera. Especially for cluttered or noisy video data tracking people has proven as quite challenging. However robust tracking results even for scenes with a very noisy background will be provided by our novel multi-cue approach. Based on a factored sampling technique, providing several hypotheses for possible locations of the tracked object, a modified active shape model approach is employed to obtain a weighting for each hypothesis. This framework is extended especially for challenging background scenarios by utilizing salient points to revaluate all hypotheses. Thus our algorithm provides a stable trajectory even in very cluttered environments with non-rigid object shapes. As an important advantageous aspect of this approach, only a few hypotheses are needed to track people consistently, resulting in a very time efficient algorithm, while comparable methods require at least between 100 and 1000 hypotheses. To enable multiple object tracking, an intelligent layer is introduced which evaluates all hypotheses in the frame based on additional low-level cues. In this way our method is able to detect fully automatically the number of persons visible in the frame and track all available persons throughout the video sequence.

1. Introduction

Within the last few years the automatic analysis of video data gained more and more in importance not at least due to a steadily growing amount of computing power. In this context a wide spectrum of applications has arisen comprising topics like video surveillance [3], medical tasks as well as intelligent indoor spaces [1]. To achieve these elusive goals, one basic requirement for higher level processing steps is the localization and tracking of single objects in image sequences. Basically the main tracking approaches can be split into two general categories, using different cues for tracking an object. Approaches belonging to the first group (e.g. blob detection [9]) consider the whole image to extract objects. In opposite to this, other approaches, classified to the second group, apply particle filtering to produce hypotheses for the object position and evaluate the image data only at these sample locations. Representatives for this class would be contour models [2, 4], articulated models [12] or color-based [13] techniques. But beside the evident advantage, that particle filters are not restricted on linear systems and do not assume Gaussian noise, there is one common problem of such approaches - a suitable measurement function has to be implemented. The novelty of our approach now lies in the originality of this measurement function, which does not only evaluate the image data but additionally tries to fit samples to plausible shapes. Together with a salient points tracker and a skin color detector our technique enables a reliable multi object tracking.

As we mentioned at the beginning, tracking will be used in numerous applications to enable further high level processing. Meeting projects [11, 17] represent such a possible application scenario. The goal of these projects is to automatically generate a protocol of the meetings by recognizing single person gestures, emotion [14], attention [16] and speech. For this reason the detection and tracking of participants is indispensable to extract features, that can be used for the classification of such gestures.

The structure of the paper is as follows. After a short description of the meeting data, the functionality of our tracking algorithm is introduced and will be explained in detail. Afterwards results for our tracking implementation are shown on the basis of different scenarios. Finally the paper concludes with a short summarization.

2. Data Acquisition

For our research on meeting scenarios we used a scripted meeting corpus, which has been recorded in a smart meeting room. In this room a typical meeting environment was emulated comprising a centrally located 4.8m × 1.2m rectangular table, a white-board as well as a projector screen. Additionally there has been installed a fully synchronized, multi-channel audio-visual recording equipment consisting of 24 microphones and 3 closed-circuit television cameras. A total dataset of 16 meetings with an average duration of 1-3 minutes has been recorded providing high quality video material with a PAL-resolution of 720 × 576 pixels and a frame rate of 25 Hz. In this dataset each subset of four sequences contains a certain amount of meeting participants, comprising scenarios with 1 to 4 persons.

VS 2006, Graz, Austria
IEEE, Jones, Graeme A. (Ed.)

Every sequence has been manually labeled in every 25th frame by a bounding box around the heads.

3. Tracking of Single Meeting Participants

3.1. Skin Color Detection

Although it might be impossible to realize robust tracking of human body parts only with low-level features like the skin color, it is nevertheless a key feature for the localization of hands and heads. In order to extract skin colored regions in the image, the RGB-color intensities are transformed into the normalized rg-chromatic color space to compensate varying lightning conditions in the images. Furthermore tests have shown, that in this 2D color space the skin-color distribution can be described by a Gaussian function, which provides a probability

$$p(skin) \propto \exp[-\frac{1}{2}(\begin{pmatrix} r \\ g \end{pmatrix} - \mu)^T C^{-1}(\begin{pmatrix} r \\ g \end{pmatrix} - \mu)] \qquad (1)$$

for each pixel to be skin-colored. For our experiments we have manually labeled roughly 400000 skin colored pixels originating from more than 200 different pictures. Observing the rg-chroma space for this training material, the mean vector μ and the covariance matrix C in the equation above have been computed as:

$$\mu = \begin{pmatrix} 0.4212 \\ 0.3151 \end{pmatrix}, C = \begin{bmatrix} 0.4440 & -0.2164 \\ -0.2164 & 0.1459 \end{bmatrix} \cdot 10^{-3}$$

After a threshold operation a binary mask as depicted in Figure 1b indicates areas with skin colored pixels. To avoid initialization of hypothesis on skin colored areas which obviously do not indicate a head, the aspect ratios of all skin colored blobs are analyzed. For this reason a bounding box is fitted around each blob. The ratio between height and width of this rectangle has to lie between 0.6 and 0.9 to be considered as a probable head and thus to serve as possible initialization location for new hypotheses. In Figure 1b all these remaining blobs are marked by rectangles.

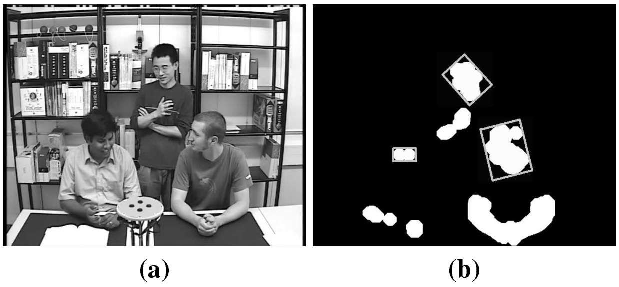
(a) **(b)**

Figure 1: Binary mask (b) representing only skin-colored areas in the original image (a)

3.2. Particle Filter Framework

Tracking persons in video data is challenging and elusive due to the complexity of the human body. Furthermore we have to deal with dense visual clutter in our meeting scenarios, and therefore Kalman Filtering has turned out as not very suitable for tracking under such contrarious conditions. Thus an algorithm is applied that uses factored sampling ([6],[8],[7]), which provides simultaneous alternative hypotheses $\mathbf{s}_t$ modeling the probability distribution $\mathbf{w}_t$ at each time step t. Based on the observations $\mathbf{z}_t$, representing the image features, the aim is to track the position of the persons throughout the posterior probability $p(\mathbf{w}_t|\mathbf{z}_{1:t})$. In most cases, there is no functional representation available for this conditional probability, but it can be derived iteratively by

$$p(\mathbf{w}_t|\mathbf{z}_{1:t}) \propto p(\mathbf{z}_t|\mathbf{w}_t) \int_{\mathbf{w_{t-1}}} p(\mathbf{w}_t|\mathbf{w}_{t-1}) p(\mathbf{w}_{t-1}|\mathbf{z}_{1:t-1}). \quad (2)$$

Updating the posterior distribution $p(\mathbf{w}_{t-1}|\mathbf{z}_{1:t-1})$ from the previous time step by prediction with dynamics $p(\mathbf{w}_t|\mathbf{w}_{t-1})$ leads to the effective prior $p(\mathbf{w}_t|\mathbf{z}_{1:t-1})$ for the actual time step. Finally multiplying the prior distribution with our measurement $p(\mathbf{z}_t|\mathbf{w}_t)$ results in the current state density $p(\mathbf{w}_t|\mathbf{z}_{1:t})$.

For the computational processing this filtering distribution is approximated now by a sample-set $S_t = \{\mathbf{s}_t^{(i)}, \pi_t^{(i)}, i = 1, \ldots, N\}$. In this sample-set each hypothesis $\mathbf{s}_t^{(i)}$, also called particle, consists of a q-dimensional vector, which will be further described in Section 3.3, and thus represents one possible shape in the image with a weight $\pi_t^{(i)}$. In Figure 2, hypotheses for both participants of the meeting are depicted. At the beginning the N particles are initialized uniformly distributed on skin colored regions in the image, which we have obtained by the skin color detection described in Section 3.1. Starting with this initial sample-set $\{\mathbf{s}_0^{(i)}, \pi_0^{(i)} = \frac{1}{N}, i = 1, \ldots, N\}$ our aim is to derive a sample-set of constant size N for each of the following time steps. Therefore we choose $N_1 < N$ particles from the old particle set $\{\mathbf{s}_{t-1}^{(i)}, \pi_{t-1}^{(i)}, i = 1, \ldots, N\}$

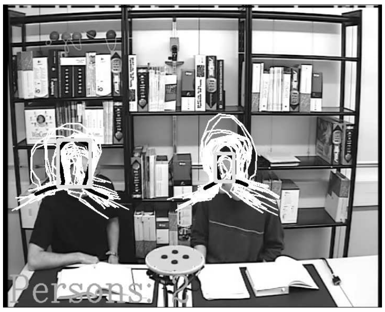

Figure 2: Image taken from a typical meeting scenario. Particles are placed on different locations in this image, but will finally concentrate on the heads over the temporal progress

at time step t, each with its probability $\pi_{t-1}^{(i)}$. After this procedure some of the old elements will be lost, while others may appear more than only one time in our new set $\{\tilde{\mathbf{s}}_t^{(i)}, \pi_{t-1}^{(i)}, i = 1, \ldots, N_1\}$. In the next step each element of the new set is predicted by a linear dynamical model with constant velocity and thus the new sample-set $\{\mathbf{s}_t^{(i)}, \pi_{t-1}^{(i)}, i = 1, \ldots, N_1\}$ is generated. To complete this set, $N_2 = N - N_1$ further particles are initialized on skin colored regions as it was already done for deriving the basic set $\{\mathbf{s}_0^{(i)}\}$.

Due to this step, persons can be automatically tracked again even if they have previously left the camera view or the track has been lost some times ago. Finally the weights $\pi_t^{(i)}$ for the elements of the new particle set are computed by measuring the head likelihood $p(\mathbf{z}_t|\mathbf{w}_t)$, which is described in the following section. Thus the the weights can be updated by

$$\pi_t^{(i)} \propto \pi_{t-1}^{(i)} p(\mathbf{z}_t|\mathbf{w}_t) \tag{3}$$

3.3. Active Shape Model based Measurement

In the literature a lot of different cues for tracking humans in cluttered environments are introduced. Applied to our indoor scenario most of them such as the face or the body have turned out to be not suitable for tracking because these features are relatively often either completely invisible or partially occluded. For this reason we exploit the shape of the head as an alternative key feature for our tracking algorithm introduced in the previous section. Due to the variations of the shape, which are caused e.g. by turning the head, a flexible model based on the work of Cootes et al. [5] was chosen to represent the shilouette of the head.

Generating the model, k points are positioned manually along the shape, resulting in a $2k$-dimensional landmark vector

$$\mathbf{x} = (x_1, y_1, \ldots, x_k, y_k). \tag{4}$$

This procedure is repeated for l pictures to capture as much variations as possible for the model. In order to be able to compare equivalent points in the training set, the euclidean transformations have to be removed from all shape vectors $\{\mathbf{x}_i, i = 1, \ldots, l\}$ and they have to be aligned, so that the mean squared error over the sum of distances of all shapes is minimized. Thus we can exploit the statistics, which we want to use for modeling new shapes, by calculating the mean

$$\overline{\mathbf{x}} = \frac{1}{l}\sum_{i=1}^{l} \mathbf{x}_i \tag{5}$$

and the covariance matrix

$$L = \frac{1}{l-1}\sum_{i=1}^{l} (\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})^T \tag{6}$$

over all training samples $\mathbf{x}_i$. Finally the eigenvectors η_i and the corresponding eigenvalues λ_i of L are computed. Gathering all eigenvectors with the r highest eigenvalues into a matrix Φ, we can generate any shape from the training set using

$$\mathbf{x}' \approx \overline{\mathbf{x}} + \Phi\mathbf{c}. \tag{7}$$

where the vector $\mathbf{c}$ is used for weighting each eigenvector η_i in the matrix Φ to produce the variations of the shape. In Figure 3 the effects on the shape are visualized for variations of the three most important eigenvectors. Together with the euclidean parameters σ (scale), ψ (rotation) and τ (translation), the r-dimensional vector $\mathbf{c}$ creates one hypothesis $\mathbf{s}^{(i)} = [\sigma, \psi, \tau, \mathbf{c}]$ mentioned in the previous section. As already explained, these hypotheses are utilized to approximate the probability distribution $p(\mathbf{z}_t|\mathbf{w}_t)$, which can be derived by evaluating the quality of each particle $\mathbf{s}_t^{(i)}$ (indicated by its weight $\pi_t^{(i)}$) on the true image data.

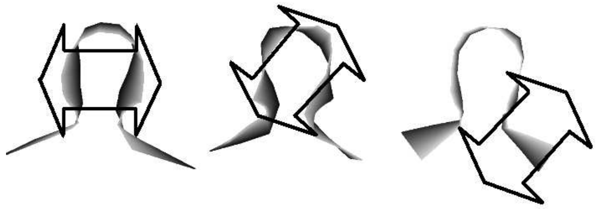

Figure 3: Variation of the weights for the three most important eigenvectors. Each eigenvector changes specific parts of the shape primarily in one main direction, indicated by the arrows.

In the first step of the measurement, the gradient image as depicted in Figure 4 is calculated. To discard as much of the background noise as possible, the gradient image is masked, so that only interesting regions remain. These interesting regions are defined by motion, which is obtained by observing the absolute value of the difference image between the actual and the previous frame, and additionally by the surroundings of the actual shapes. After that the model is iteratively adapted to the image data based on the gradient image as follows:

Iteration_start

The normal vector $\{\mathbf{n}_i, i = 1, \ldots, k\}$ through each landmark (available from the mean model shape $\overline{\mathbf{x}}$) is computed and along this straight line the dot product $\mu_{i,j}$ of the unit vector normal and the gradient $\mathbf{g}(x, y)$ at each pixel position inside a certain distance ϵ to the respective landmark $\{\mathbf{p}_i = (x_i', y_i'), i = 1, \ldots, k\}$ (cf. Figure 4) is calculated.

$$\mu_{i,j} = \mathbf{n}_i \circ \mathbf{g}(\mathbf{p}_i + j \cdot \mathbf{n}_i), \forall i \in \{1, \ldots, k\}, j \in [-\epsilon, \epsilon] \tag{8}$$

For each landmark the pixel $(\tilde{x}_i, \tilde{y}_i)$ with the highest score $\mu_{i,j}$ is chosen for a new contour $\tilde{\mathbf{x}}$, which represents best the image data.

In the next step the model calculates optimal parameters for the euclidean transformations to minimize the sum of

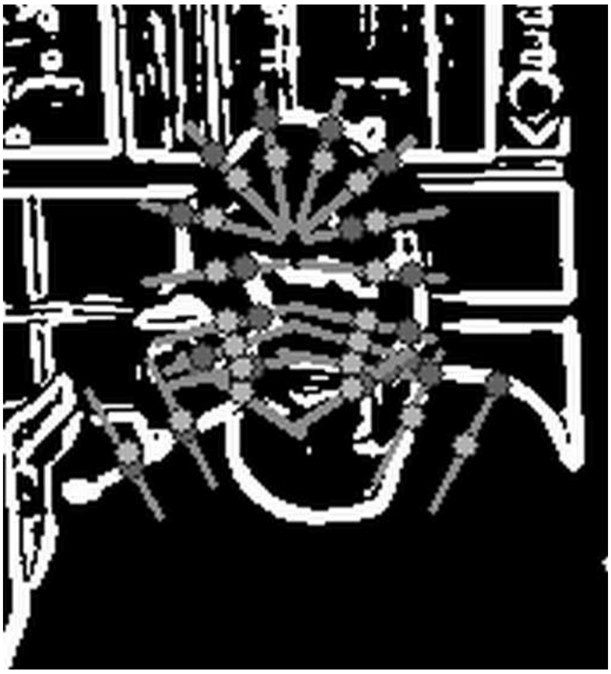

Figure 4: Gradient image, in which edges are represented by the absolute value of the gradient. Furthermore through each landmark (light gray points) the normal vectors have been plotted. Along these normals the landmarks are shifted to the pixel, where normal vector and gradient direction are most similar, resulting in the new landmark positions (dark gray points).

squared distances between the model landmark vector $\mathbf{x}'$ and the landmarks of the new contour $\tilde{\mathbf{x}}$. This is achieved by taking the partial derivatives of the squared distance E and solving to the variables σ, ψ and τ

$$E=\sum_{i=1}^{k}\left|\begin{pmatrix} a & b\\ -b & a\end{pmatrix}\begin{pmatrix}\tilde{x}_i\\ \tilde{y}_i\end{pmatrix}+\begin{pmatrix}\tau_x\\ \tau_y\end{pmatrix}-\begin{pmatrix}x_i\\ y_i\end{pmatrix}\right|^2$$

$$\begin{aligned} a &= \sigma\cos\psi,\\ b &= \sigma\sin\psi \end{aligned} \tag{9}$$

At this stage, we achieve a transformed contour $\hat{\mathbf{x}}$ by applying these parameters to the model $\tilde{\mathbf{x}}$. Finally the shape of the contour $\mathbf{x}$ is adapted to that of the new contour $\hat{\mathbf{x}}$ by computing an optimal model parameter $\mathbf{c}$ according to Equation 7

$$\mathbf{c} = \mathbf{\Phi}^T(\hat{\mathbf{x}} - \overline{\mathbf{x}}) \tag{10}$$

and thus we are given a transformed and adapted contour $\mathring{\mathbf{x}}$.
Iteration_end
This iteration block is repeated until the difference of the parameter $\mathbf{c}$ between two successive iterations falls below a given threshold. The complete weighting procedure is finished by the computation of the sum over the dot product between the unit vector normal and the gradient at the landmark position $\{\mathbf{p}_i = (\mathring{x}_i, \mathring{y}_i), i = 1,\ldots,k\}$ of the final contour $\mathring{\mathbf{x}}$:

$$\Omega = \sum_{i=1}^{k} |\mathbf{g}(\mathbf{p}_i)\circ\mathbf{n}_i|^2 \tag{11}$$

With the dot product in Equation 11 a score for the quality can be measured, to what extent the adapted model fits gradients in the image data. As this equation already implies, the score should not depend on the direction of the unit vector normal and thus we have to summarize over the absolute value of the dot product. This score now represents the probability for a plausible head contour described by the respective hypothesis $\mathbf{s}^{(i)}$ at time step t and is used to update the weight π_t according to Equation 3.

4. Multi Person Tracking

Due to the basic principle of the particle filter all hypotheses would finally concentrate only on one location in the image - to wit the one where the hypotheses have their highest weight. To prevent all hypotheses from converging to one and the same shape, a hyper layer is introduced to control the allocation of M different hypotheses sets $S_t^{(j)} = \{\mathbf{s}_t^{(i)}, \pi_t^{(i)}, i = 1,\ldots,N\}, j = 1,\ldots,M$, where each of these sets consists of N hypotheses to represent exactly one object. The hyper layer is organized similar to the basic particle filter described in Section 3.2, but here hypotheses comprise complete hypotheses sets. These hypotheses sets are sampled and predicted as it was done for the single hypotheses above. Then for each hypothesis of the sets, the active shape model is run to obtain a weight. Due to this measurement a weight $\Pi_t^{(j)}$ for the hyper layer hypotheses $S_t^{(j)}$ can be computed by

$$\Pi_t^{(j)} = \Pi_{t-1}^{(j)}\frac{1}{N}\sum_{s^{(i)}\in S^{(j)}}\Omega_i \tag{12}$$

This measurement still would not prevent hypotheses sets to converge in one image location. Thus some additional low-level cues, which are described in the following sections, are utilized to enhance the performance of ongoing tracking multiple objects.

4.1. Skin Color Validation

For the validation of the sets the ratio between the area covered by the mean shape and the corresponding skin blob, i.e. the mean blob with the smallest distance to the mean shape, is computed:

$$p_{skin} = \frac{\{A_{meanshape}\}\cap\{A_{skinblob}\}}{\{A_{skinblob}\}}. \tag{13}$$

To allow some tolerance, especially for situations like sitting down or occlusion, where no skin color is available, the hyper layer hypothesis weight is only updated, if p_{skin} is less than a given threshold:

$$\Pi_t^{(j)} = \Pi_t^{(j)} \cdot p_{skin} \tag{14}$$

Since the skin color map has been already computed (cf. Section 3.1), this measure means no additional computational expense.

4.2. Salient Points Validation

Salient points or interest points are landmarks in an image which are often intuitively obvious to a human like corners or edges of objects. In our context interest points are defined as prominent points within the human head, e.g. the eyes, mouth, nose or ears. To detect these features strong corners within the hypothetical shape of a head, represented by the mean of the hypotheses set, the Harris operator is applied. Thus features are determined by computing the eigenvalues $\lambda_{1,2}(x, y)$ of the tensor matrix

$$B = \begin{pmatrix} \sum_R (dI/dx)^2 & \sum_R (dI/dx \cdot dI/dy) \\ \sum_R (dI/dx \cdot dI/dy) & \sum_R (dI/dy)^2 \end{pmatrix}, \tag{15}$$

where R is a quadratic neighborhood of n pixels. The minor of the two eigenvalues is then assigned to its corresponding pixel location (x, y), resulting in an eigenvalue map T. On this map, a non-maxima suppression within a 3×3 neighborhood is performed to obtain local maxima. After that all corners with an eigenvalue smaller than an adaptive threshold, depending on the maximum of T, are discarded. Finally, for all remaining eigenvalues resp. corners it has to be ensured that all corners are distanced far enough to the next corner. Therefor the distances between all corners are computed and all corners distanced less than a minimum distance are rejected. These salient points are used for the improvement of the hypotheses' weight as follows:

If there have not been assigned any salient points to the hypotheses set before, the K corners with the highest eigenvalues are assigned. These salient points are predicted in the next frame by calculating the optical flow for every pixel of the actual image using the algorithm of Lucas and Kanade [10]. The new salient points are validated again and as Figure 5 demonstrates, some of them are positioned outside the shape ($\oplus$). The ratio

$$p_{salient} = \frac{\text{No. of salient points inside}}{\text{No. of overall salient points}} \tag{16}$$

between remaining salient points inside the shape and overall salient points. With this probability again the hyper layer hypothesis weight can be updated as follows:

$$\Pi_t^{(j)} = \Pi_t^{(j)} \cdot p_{salient} \tag{17}$$

If the number of salient points is smaller than K the missing ones are refilled by the those detected in the way described above, thus each hypotheses set $S_t^{(j)}$ always is assigned N salient points. The salient points validation can be interpreted as a measure for the excursiveness of the mean shape represented by the hypotheses set $S_t^{(j)}$. which should be of course very small.

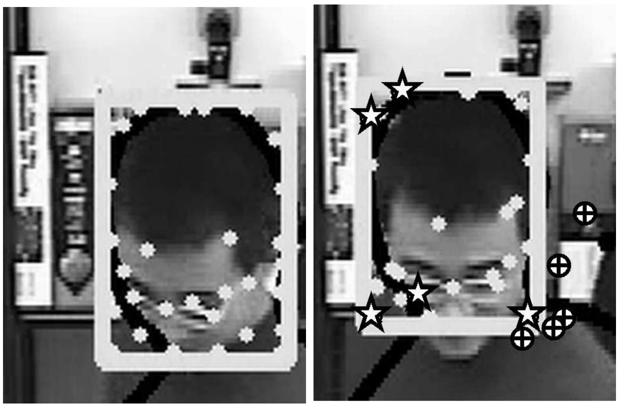

Figure 5: Frames with bounding box, containing salient points (light gray points). Some of the salient points ($\oplus$) left the bounding box after optical flow prediction. These points are replaced by new detected salient points (☆).

5. Results

Although a lot of research is concentrated on multiple object tracking there is little agreement amongst the community on how to evaluate or compare these methods. For the tracking results shown below in Table 1, a comprehensive list of error measures, introduced in [15], is used to enable nevertheless a qualitative and objective performance rating of our tracks. The basic procedure for the evaluation is as follows:

At first for each combination between a tracker output, referred to as estimate $\mathcal{E}_i$, and a labeled tracking target, denoted as ground truth object $\mathcal{GT}_j$, two measures, the precision and recall are computed as follows:

$$\text{Recall} \quad \alpha_{i,j} = \frac{|\mathcal{E}_i \cap \mathcal{GT}_j|}{|\mathcal{GT}_j|}$$

$$\text{Precision} \quad \beta_{i,j} = \frac{|\mathcal{E}_i \cap \mathcal{GT}_j|}{|\mathcal{E}_i|}$$

As it can be shown very easily, both α and β must be high to obtain good tracking results. For this reason, a coverage test using the F-measure

$$F_{i,j} = \frac{2\alpha_{i,j}\beta_{i,j}}{\alpha_{i,j} + \beta_{i,j}} \tag{18}$$

has to be passed, returning only a high value if $\alpha_{i,j}$ and $\beta_{i,j}$ are high. This test is considered to be passed, if $F_{i,j}$ exceeds a fixed threshold t_c and thus determines, that $\mathcal{GT}_j$ is being tracked by $\mathcal{E}_i$. The error measures, which can now occur in multiple object tracking and thus are computed for all tracking results in this paper, can be defined as follows:

a) **Measure** FP - False positive. There is an $\mathcal{E}$ indicating an object, where no $\mathcal{GT}$ is.

b) **Measure** FN - False negative. A $\mathcal{GT}$ is not tracked by an $\mathcal{E}$.

c) **Measure** MT - Multiple trackers. More than one $\mathcal{E}$ is associated with only one $\mathcal{GT}$. In order to obtain the subjective impression of a human spectator each excess $\mathcal{GT}$ is counted as a MT error.

d) **Measure** MO - Multiple objects. More than one $\mathcal{GT}$ is associated with only one $\mathcal{E}$. Again a MO error is assigned for each excess $\mathcal{GT}$.

e) **Measure** CD - Configuration Distance. The difference between the number of $\mathcal{E}$ and $\mathcal{GT}$, divided by the number of $\mathcal{GT}$ present in a given frame.

e) **Measure** FIT - Falsely identified tracker. An $\mathcal{E}_i$ which passed the coverage test for $\mathcal{GT}_j$ is different to that identifying this ground truth object before.

f) **Measure** FIO - Falsely identified object. A $\mathcal{GT}_j$ which passed the coverage test for $\mathcal{E}_i$ has not been the identified object in the frame before.

g) **Measure** OP - Object purity. If $\mathcal{GT}_j$ is the ground truth object which has been identified by $\mathcal{E}_i$ for most of the time, then OP is the ratio of frames $(n_{i,j})$ that $\mathcal{GT}_j$ is correctly identified by $\mathcal{E}_i$ to the overall number of frames (n_j) $\mathcal{GT}_j$ exists.

h) **Measure** TP - Tracker purity. If $\mathcal{E}_i$ is the estimate which has been identified by $\mathcal{GT}_j$ for most of the time, then TP is the ratio of frames $(n_{i,j})$ that $\mathcal{E}_i$ is correctly identified by $\mathcal{GT}_j$ to the overall number of frames (n_i) $\mathcal{E}_i$ exists.

To obtain an overall impression of the performance of the tracking results, all these error measures are divided by the number of ground truth objects visible in each frame, summed up and normalized by the overall amount of frames, e.g.:

$$\overline{FP} = \frac{1}{f}\sum_{t=0}^{f}\frac{FP_t}{max(N_t^{\mathcal{GT}},1)} \qquad (19)$$

In Table 1 the error measures are listed for 13 sequences, where every 25th frame of the tracking output was evaluated on the ground truth data. Frames from the other sequences have been used for training the active shape model and have therefor not been included into the tracking evaluation. For the assignment of ground truth and estimates a common threshold of 0.33 for the F-measure from the coverage test was used. As this table shows for meetings with only a few participants, our algorithm provides robust results. With the growing number of participants persons often leave and reenter the room. Thus hypotheses have to be initialized very often and basically there is not enough time to adapt the shapes to the image data.
In Figure 6 every third frame of a typical video sequence, containing quite a lot of challenges, is depicted. In the first frame one person is already in the room, being tracked very precisely. A second person enters the room and, as shown by the rectangle, is automatically detected and tracked. The second person walks towards the left side of the room, occluding the first person partially. At the beginning of this occlusion, the bounding box of the first person is slightly disturbed, but recovers even during the occlusion. For all runs of our tracking algorithm on the video sequences one common active shape model consisting of 20 landmarks points was used, which has been trained on 40 head-shoulder contours based on 5 different persons.

6. Summary and Conclusions

In this paper a system for the automatic tracking of multiple people has been presented. A novel tracking approach based on the combination of two powerful techniques, Active Shape Models and Particle Filters, has been introduced. This basic framework was further optimized so that only 20 hypotheses per object are necessary for tracking, only a fraction compared to most of the state-of-the-art particle filter approaches. This implementation has been tested on approximately two hours of video material containing special challenges like dense visual clutter in the background, partial/total occlusion and different skin color. Extracts of our results have been depicted, which show a robust tracking behavior even for critical situations like sitting down or partial occluded scenes. Although video sequences can be already processed in acceptable time, future work will deal with optimizing and accelerating our algorithm by incorporating additional low-level features, which will lead to a further improvement of available hypotheses in the framework. First attempts will also be run on integrating automatic recognition of person gestures into this framework, which will be necessary for generating a protocol of meetings by the computer.

7. Acknowledgement

This work was partly supported by the European Union 6th FWP IST Integrated Project AMI (Augmented Multi-party Interaction, FP6-506811, publication).

References

[1] T. Akiyama, J. Lee and H. Hashimoto, “Evaluation of Human Localization Using Color Model in Intelligent Space,” *International Conference on Control, Automation and Systems*, pp.198-201, October 2001.

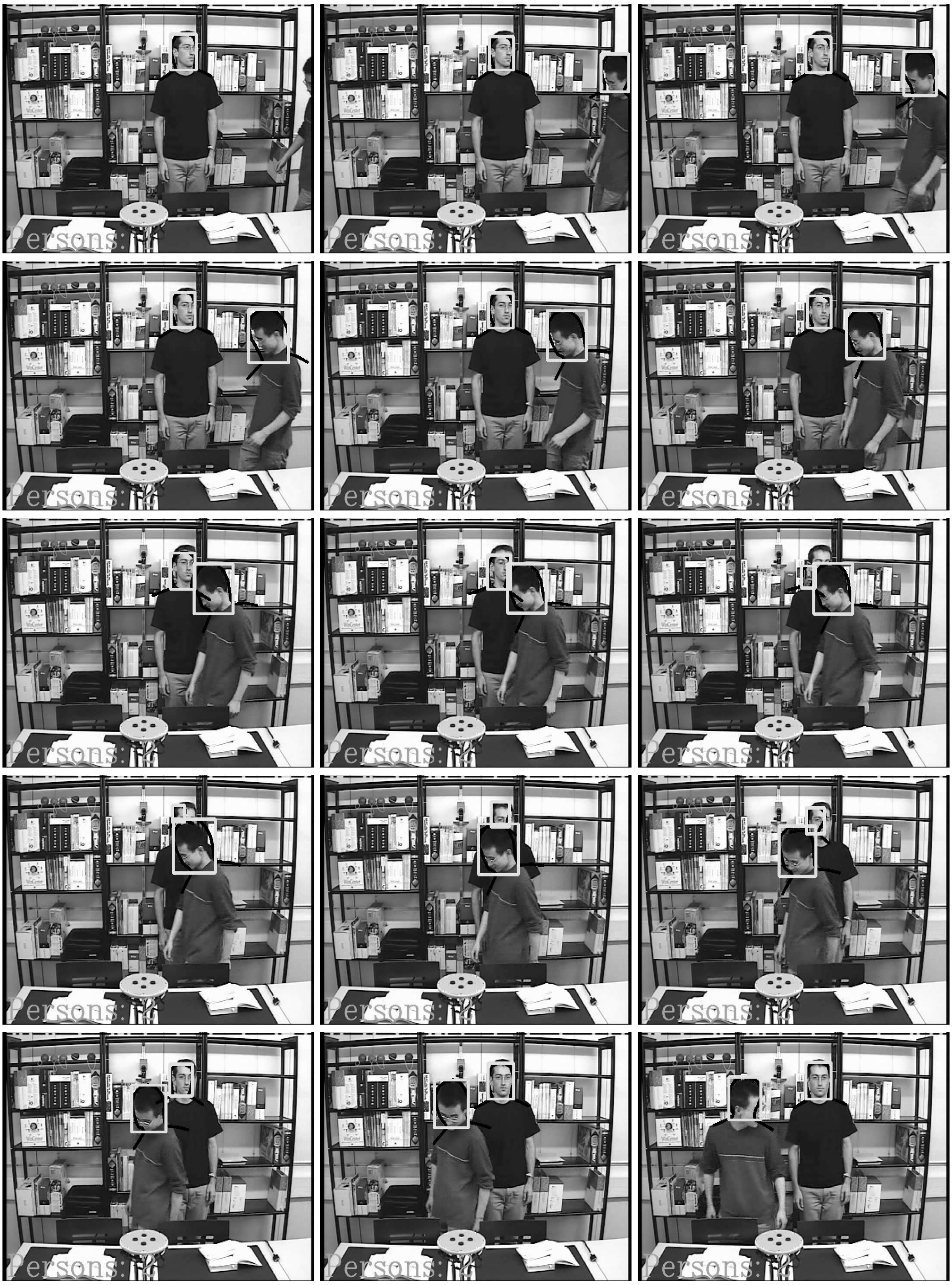

Figure 6: Meeting scenario with two persons. These exemplary frames already contain a lot of challenges like entering the room, partial occlusion and dense clutter in the background. Nevertheless the tracking output (light gray rectangles) and the number of identified tracks are very precisely.

Type of scenario	F-meas	**FN**	**FP**	**MT**	**MO**	**CD**	$\overline{\mathbf{FN}}$	$\overline{\mathbf{FP}}$	$\overline{\mathbf{MT}}$	$\overline{\mathbf{MO}}$	$\overline{\mathbf{CD}}$	**FIT**	**FIO**	$\overline{\mathbf{FIT}}$	$\overline{\mathbf{FIO}}$	$\overline{\mathbf{TP}}$	$\overline{\mathbf{OP}}$
1-Person	0.84	1	0	0	0	-1	0.02	0.00	0.00	0.00	0.02	0	0	0.00	0.00	1.00	0.88
1-Person	0.73	10	8	0	0	-2	0.20	0.16	0.00	0.00	0.16	0	0	0.00	0.00	0.58	0.52
1-Person	0.72	41	23	0	0	-18	0.20	0.11	0.00	0.00	0.14	0	0	0.00	0.00	0.70	0.56
1-Person	0.65	52	28	0	0	-24	0.25	0.13	0.00	0.00	0.12	0	0	0.00	0.00	0.66	0.51
1-Person	0.45	18	15	0	0	-3	0.28	0.23	0.00	0.00	0.08	0	0	0.00	0.00	0.06	0.05
1-Person	0.49	25	17	0	0	-8	0.39	0.27	0.00	0.00	0.13	0	0	0.00	0.00	0.19	0.14
2-Person	0.64	53	60	3	0	7	0.27	0.32	0.01	0.00	0.14	35	65	0.20	0.35	0.40	0.34
3-Person	0.66	21	25	3	0	8	0.18	0.25	0.03	0.00	0.21	14	34	0.13	0.35	0.51	0.38
3-Person	0.64	35	24	0	0	-10	0.48	0.33	0.00	0.00	0.54	0	0	0.00	0.00	0.17	0.02
3-Person	0.65	125	66	4	0	-14	0.51	0.35	0.02	0.00	0.40	34	53	0.16	0.23	0.39	0.24
4-Person	0.82	181	2	1	0	-72	0.79	0.02	0.01	0.00	0.81	9	15	0.07	0.10	0.70	0.05
4-Person	0.35	198	80	3	0	-43	0.67	0.30	0.01	0.00	0.43	18	35	0.07	0.12	0.35	0.13
4-Person	0.26	252	103	16	0	-43	0.71	0.31	0.04	0.00	0.41	27	60	0.06	0.13	0.41	0.13

Table 1: Tracking results for differnet constellations of people participating in the meeting.

[2] A. Baumberg and D. Hogg, “An Efficient Method for Contour Tracking Using Active Shape Models,” *Proceedings of IEEE Workshop on Motion of Non-Rigid and Articulated Objects*, pp. 194-199, 1994.

[3] D. Beymer, P. McLauchlan, B. Coifman and J. Malik, “A real-time computer vision system for measuring traffic parameters,” *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1997.

[4] S. Birchfield, “Elliptical Head Tracking Using Intensity Gradients and Color Histograms,” *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Santa Barbara, California, pp. 232-237, June 1998

[5] T.F. Cootes, D. Cooper, C.J. Taylor and J. Graham, “Active Shape Models - Their Training and Application.” *Computer Vision and Image Understanding*, Vol. 61, No. 1, pp. 38-59, January 1995.

[6] M. Isard and A.Blake, “Condensation - conditional density propagation for visual tracking,” *International Journal of Computer Vision 29(1)*, pp. 5-28, 1998.

[7] M. Isard and A. Blake, “A Mixed-State CONDENSATION Tracker with Automatic Model-Switching,” *Proceedings International Conference on Computer Vision*, pp. 107-112, 1998

[8] M. Isard and A. Blake, “ICONDENSATION: Unifying Low-level and High-level Tracking in a Stochastic Framework,” *Proc. of the Fifth European Conference on Computer Vision (ECCV ‘98)*, Vol I. pp. 893-908, Freiburg, Germany, June 1998.

[9] D. Koller, J. Weber and J. Malik “Robust Multiple Car Tracking with Occlusion Reasoning,” *European Conference on Computer Vision*, pp. 189-196, 1994.

[10] B. Lucas and T. Kanade “An Iterative Image Registration Technique with an Application to Stereo Vision,” *Proc. of 7th International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 674-679, 1981.

[11] N. Morgan, D. Baron, J. Edwards, D. Ellis, D. Gelbart, A. Janin, T. Pfau, E. Shriberg and A. Stolcke “The Meeting Project at ICSI,” *Proceedings of the Human Language Technology Conference*, San Diego, CA, March 2001.

[12] H. Ning, L. Wang, W. Hu and T. Tan, “Articulated Model-Based People Tracking Using Motion Models,” *Proceedings International Conference on Multimodal Interfaces*, pp. 383-388, 2002

[13] K. Nummiaro, E. Koller-Meier and L. Van Gool, “An Adaptive Color-Based Particle Filter,” *Image and Vision Computing*, Vol. 21, Issue 1, pp. 99-110, January 2003.

[14] T. S. Polzin and A. Waibel “Detecting Emotions In Speech,” *Proceedings of the CMC*, 1998.

[15] K. Smith, S. Ba, J. Odobez, and D. Gatica-Perez, “Evaluating multi-object tracking,” *Workshop on Empirical Evaluation Methods in Computer Vision (EEMCV)*, San Diego, CA, USA, June 2005,

[16] R. Stiefelhagen “Tracking Focus of Attention in Meetings,” *IEEE International Conference on Multimodal Interfaces*, Pittsburgh, PA, USA, October 2002.

[17] http://www.m4project.org/overview.html