IDIAP RESEARCH REPORT

2D Multi-Person Tracking: A Comparative Study in AMI Meetings

Kevin Smith [a] Sascha Schreiber [b]
Vítezslav Beran [c] Igor Potúcek [c]
Gerhard Rigoll [b] Daniel Gatica-Perez [a]

IDIAP–RR 06-37

April 2006

paru dans
Classification of Events, Activities, and Relationships (CLEAR) 2006

[a] IDIAP Research Institute, Switzerland
[b] Technische Universität München (TUM), Germany
[c] Brno University of Technology (BUT), Czech Republic

IDIAP Research Institute **www.idiap.ch**
Rue du Simplon 4 P.O. Box 592 1920 Martigny – Switzerland
Tel: +41 27 721 77 11 Fax: +41 27 721 77 12 Email: info@idiap.ch

Rapport de recherche de l'IDIAP 06-37

2D Multi-Person Tracking: A Comparative Study in AMI Meetings

Kevin Smith Sascha Schreiber Vítezslav Beran Igor Potúcek
Gerhard Rigoll Daniel Gatica-Perez

April 2006

paru dans
Classification of Events, Activities, and Relationships (CLEAR) 2006

Résumé. In this paper, we present the findings of the Augmented Multiparty Interaction (AMI) project investigation on the localization and tracking of 2D head positions in meetings. The focus of the study was to test and evaluate various multi-person tracking methods developed in the project using a standardized data set and evaluation methodology.

Fig. 1 – *Examples from seq14 of the AV16.7.avi data corpus.* Left : Typical meeting room data with four participants (free to stand, sit, walk). Center : Participant heads near the camera are not fully visible and often move in and out of the scene. Right : The data set also contained challenging situations such as this (four heads appear and are annotated in this image).

1 Introduction

One of the fundamental goals of the AMI project is to formally and consistently evaluate tracking methods developed by AMI members using a standardized data set and evaluation methodology. In a meeting room context, these tracking methods must be robust to real-world conditions such as variation in person appearance and pose, unrestricted motion, changing lighting conditions, and the presence of multiple self-occluding objects. In this paper, we present an evaluation methodology for gauging the effectiveness of various 2D multi-person head tracking methods and provide an evaluation of four tracking methods developed under the AMI framework in the context of a meeting room scenario.

The rest of this paper is organized as follows : Section 2 describes the method of evaluation, Section 3 briefly describes the tracking methods, Section 4 presents the results of the evaluation, and Section 5 provides some concluding remarks[1].

2 Evaluation Methodology

To objectively compare the tracking methods, a common data set was agreed upon (Sec. 2.1) and evaluation procedure [13] was adopted (Sec. 2.2).

2.1 Data Set

Testing was done using the AV16.7.ami corpus, which was specifically collected to evaluate localization and tracking algorithms[2]. The corpus consists of 16 sequences recorded from two camera angles in a meeting room using four actors. Seven sequences were designated as the training set, and nine sequences for testing. The sequences depict up to four people performing common meeting actions such as sitting down, discussing around a table, etc (see Figure 1). Participants acted according to different predefined agendas for each scene (they were told the order in which to enter the room, sit, or pass each other), but the behavior of the subjects was otherwise natural. The sequences contain many challenging phenomena for tracking methods including occlusion, cameras blocked by passing people, partial views of backs of heads, and large variations in head size (see Table 1).

The corpus was annotated using bounding boxes for head location for use in training and evaluation [3]. Annotators were instructed to fit the bounding boxes around the perimeters of the participants heads, which were ambiguous in some cases. To reduce annotation time, every 25th frame was annotated (evaluations were performed only on annotated frames).

2.2 Measures and Procedure

In [13], the task of evaluating tracker performance was broken into evaluating three tasks : fitting ground truth persons (or $\mathcal{GT}$s) with tight bounding boxes (referred to as *spatial fitting*), predicting

[1] This paper orginally appeared with minor changes in Proceedings Multimodal Interaction and Related Machine Learning Algorithms (MLMI) 2006

[2] We are thankful to Bastien Crettol for his support with the collection, annotation, and distribution of the AV16.7ami corpus, and to the participants for their time.

the correct number and placement of people in the scene (referred to as *configuration*), and checking the consistency with which each tracking result (or estimate, $\mathcal{E}$) assigns identities to a $\mathcal{GT}$ over its lifetime (referred to as *identification*). Several measures are defined to evaluate these tasks, each dependant on the fundamental *coverage test*. The tasks measured in [13] are similar in many ways to those in [7], but the methods for measuring differ in a fundamental way : the mapping of $\mathcal{E}$s and $\mathcal{GT}$s. Measures in [7] are computed using a one-to-one mapping, whereas [13] defines measures using many-to-one w.r.t. $\mathcal{E}$s and many-to-one w.r.t. $\mathcal{GT}$s. We believe the latter to be a superior method, since situations can arise where there is no clearly correct one-to-one mapping between the $\mathcal{E}$s and $\mathcal{GT}$s.

2.2.1 Coverage Test. The coverage test determines if a $\mathcal{GT}$ is being tracked by an $\mathcal{E}$, if a $\mathcal{E}$ is tracking a $\mathcal{GT}$, and reports the quality of the tracking result. For a given tracking estimate $\mathcal{E}_i$ and ground truth $\mathcal{GT}_j$, the coverage test measures the overlap between the two areas using the *fitting F-Measure* $F_{i,j}$ [11]

$$F_{i,j} = \frac{2\alpha_{i,j}\beta_{i,j}}{\alpha_{i,j}+\beta_{i,j}} \qquad \alpha_{i,j} = \frac{|\mathcal{E}_i \cap \mathcal{GT}_j|}{|\mathcal{GT}_j|} \qquad \beta_{i,j} = \frac{|\mathcal{E}_i \cap \mathcal{GT}_j|}{|\mathcal{E}_i|} \tag{1}$$

where recall (α) and precision (β), are well-known information retrieval measures. If the overlap passes a fixed coverage threshold ($F_{i,j} \geq t_c$, $t_c = 0.33$), then it is determined that $\mathcal{E}_i$ *is tracking* $\mathcal{GT}_j$ and $\mathcal{GT}_j$ *is tracked* by $\mathcal{E}_i$.

2.2.2 Configuration. In this context, configuration means the number, the location, and the size of all people in a frame. A tracking result is considered to be *correctly configured* if and only if exactly one $\mathcal{E}_i$ is tracking each $\mathcal{GT}_j$. Four types of errors may occur, which correspond to the four configuration measures :

- **FN** - False negative. A $\mathcal{GT}$ is which not tracked by an $\mathcal{E}$.
- **FP** - False positive. An $\mathcal{E}$ exists which is not tracking a $\mathcal{GT}$.
- **MT** - Multiple trackers. More than one $\mathcal{E}$ is tracking a single $\mathcal{GT}$. An MT error is assigned for each excess $\mathcal{E}$.
- **MO** - Multiple objects. An $\mathcal{E}$ is tracking multiple $\mathcal{GT}$s. An MO error is assigned for each excess $\mathcal{GT}$.

An example of each error type is depicted in Fig. 2, where the $\mathcal{GT}$s are marked with green colored boxes, the $\mathcal{E}$s with red and blue. One can also measure the difference between the number of $\mathcal{GT}$s and the number of $\mathcal{E}$s :

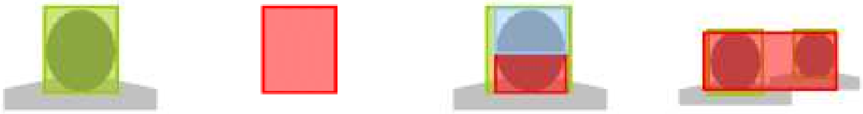
- **CD** - Counting distance. For a given frame, the difference between the number of $\mathcal{E}$s ($N^t_{\mathcal{E}}$) and $\mathcal{GT}$s ($N^t_{\mathcal{GT}}$) normalized by the number of $\mathcal{GT}$s ($N^t_{\mathcal{GT}}$).

$$\mathbf{CD} = \frac{N^t_{\mathcal{E}} - N^t_{\mathcal{GT}}}{max(N^t_{\mathcal{GT}}, 1)} \tag{2}$$

2.2.3 Identification. In the context of this evaluation, identification implies the persistent tracking of a $\mathcal{GT}$ by a particular $\mathcal{E}$ over time. Though several methods to associate identities exist, we adopt an approach based on a *majority rule* [13]. A $\mathcal{GT}_j$ is said to be identified by the $\mathcal{E}_i$ which passes the coverage test for the majority of $\mathcal{GT}_j$s lifetime, and similarly $\mathcal{E}_i$ is said to identify the $\mathcal{GT}_j$ which

TAB. 1 – Challenges in the AV16.7.ami data corpus test set (yes = y, no = n).

	seq01		seq02		seq03		seq08		seq09		seq12		seq13		seq14		seq16	
	L	R	L	R	L	R	L	R	L	R	L	R	L	R	L	R	L	R
duration (sec)	63		48		208		99		70		103		94		118		89	
total # heads	1	1	1	1	1	1	2	2	2	2	3	3	3	3	4	4	4	4
frontal heads	1	1	1	1	1	1	2	0	2	0	3	0	3	0	2	2	4	2
rear heads	1	1	1	1	1	1	0	2	0	2	0	3	0	3	2	2	4	4
event : occlusion	n	n	n	n	n	n	y	n	y	y	y	y	y	y	y	y	y	n
event : camera blocked	y	y	y	y	n	n	y	y	n	y	n	y	n	y	y	y	y	y
event : sit down	n	n	n	n	y	y	y	y	n	n	y	y	y	y	y	y	n	n



False negative (**FN**) False positive (**FP**) Multiple trackers (**MT**) Multiple objects (**MO**)

FIG. 2 – The four types of configuration errors. $\mathcal{GT}$s are represented by green boxes, $\mathcal{E}$s by red and blue boxes.

passes the coverage test for the majority of $\mathcal{E}_i$s lifetime (this implies that associations between $\mathcal{GT}$s and $\mathcal{E}$s will not necessarily match).
There can arise two types of identification failures, quantified by five measures.

- **FIT** - Falsely identified tracker. Occurs when a $\mathcal{E}_k$ which passed the coverage test for $\mathcal{GT}_j$ is not the identifying tracker, $\mathcal{E}_i$. FITs often result when $\mathcal{E}_i$ suddenly stops tracking $\mathcal{GT}_j$ and another $\mathcal{E}_k$ continues tracking $\mathcal{GT}_j$.
- **FIO** - Falsely identified object. Occurs when a $\mathcal{GT}_k$ which passed the coverage test for $\mathcal{E}_i$ is not the identifying person, $\mathcal{GT}_j$. FIOs often result from swapping $\mathcal{GT}$s, i.e. $\mathcal{E}_i$ initially tracks $\mathcal{GT}_j$ and subsequently tracks $\mathcal{GT}_k$.
- **OP** - Object purity. If $\mathcal{GT}_j$ is identified by $\mathcal{E}_i$, then OP is the ratio of frames in which $\mathcal{GT}_j$ and $\mathcal{E}_i$ passed the coverage test ($n_{i,j}$) to the overall number of frames $\mathcal{GT}_j$ exists (n_j).
- **TP** - Tracker purity. If $\mathcal{E}_i$ identifies $\mathcal{GT}_j$, then TP is the ratio of frames in which $\mathcal{GT}_j$ and $\mathcal{E}_i$ passed the coverage test ($n_{j,i}$) to the overall number of frames $\mathcal{E}_i$ exists (n_i).
- *identity F-Measure* - combines **OP** and **TP** using the F-measure such that if either component is low, identity F-Measure is low : $identity\ FMeasure = \frac{2\ OP\ TP}{OP+TP}$.

2.2.4 Procedure. To evaluate the ability of each tracking method for the tasks of spatial fitting, configuration and identification over diverse data sets, the following procedure is followed for each sequence :

Evaluation procedure for a data sequence.

1. `for` each frame in the sequence
   - `determine` *tracking maps* by applying the coverage test over all combinations of $\mathcal{E}$s and $\mathcal{GT}$s.
   - `record` configuration measures (FN,FP,MT,MO, CD) and fitting F-Measure from tracking maps.
2. `determine` *identity maps* for tracked $\mathcal{E}$s and $\mathcal{GT}$s using the *majority rule*.
3. `for` each frame in the sequence
   - `record` identification errors (FIT,FIO) from the identity maps.
4. `normalize` the configuration and identification errors and `compute` the purity measures for the entire sequence (the instantaneous number of ground truths and estimates are $N_{\mathcal{GT}}$ and $N_{\mathcal{E}}$ respectively, and the total number of frames is T).

$$\overline{FP} = \frac{1}{T}\sum_{t=1}^{T}\frac{FP_t}{max(N_{\mathcal{GT}}^t,1)} \quad , \quad \overline{FN} = \frac{1}{T}\sum_{t=1}^{T}\frac{FN_t}{max(N_{\mathcal{GT}}^t,1)},$$

$$\overline{MT} = \frac{1}{T}\sum_{t=1}^{T}\frac{MT_t}{max(N_{\mathcal{GT}}^t,1)} \quad , \quad \overline{MO} = \frac{1}{T}\sum_{t=1}^{T}\frac{MO_t}{max(N_{\mathcal{GT}}^t,1)},$$

$$\overline{FIT} = \frac{1}{T}\sum_{t=1}^{T}\frac{FIT_t}{max(N_{\mathcal{GT}}^t,1)} \quad , \quad \overline{FIO} = \frac{1}{T}\sum_{t=1}^{T}\frac{FIO_t}{max(N_{\mathcal{GT}}^t,1)},$$

TAB. 2 – Properties of the various head tracking approaches.

	Method A	Method B	Method C	Method D
Learned Models	binary, color, head shape	skin color, shape	skin color	face/nonface weak classifiers
Initialization	automatic	automatic	automatic	automatic
Features	background sub, silhouette, color	motion detection, skin color, head/shoulder shape	background sub, skin color, local charact.	skin color, gabor wavelets
Mild Occ.	robust	robust	robust	robust
Severe Occ.	semi-robust	semi-robust	sensitive	sensitive
Identity Recovery	swap, rebirth	swap, rebirth	rebirth	none
Comp. Exp.	∼1 frame/sec	∼3 frame/sec	∼20 frame/sec	∼0.2 frame/sec

$$\overline{OP} = \frac{1}{N_{\mathcal{GT}}} \sum_{j=1}^{N_{\mathcal{GT}}} \frac{n_{i,j}}{n_j}, \qquad \overline{TP} = \frac{1}{N_{\mathcal{E}}} \sum_{i=1}^{N_{\mathcal{E}}} \frac{n_{j,i}}{n_i}, \qquad \overline{CD} = \frac{1}{T} \sum_{t=1}^{T} |\mathbf{CD}|$$

$N_{\mathcal{GT}}$ and the number of frames (such as $\overline{FP}$). For these measures, the number reported could be thought of as a rate of error. For instance, $\overline{FP} = .25$ could be interpreted as : "for a given person, at time t, 0.25 FP errors will be generated on average."

3 Tracking Methods

Four head tracking methods built within AMI were applied to the data corpus and evaluated as described in Section 2. Each method approached the problem of head tracking differently, and it is noteworthy to list some of the qualitative differences (see Table 2). These methods are described briefly below.

3.1 Method A : Trans-Dimensional MCMC (developed at IDIAP).

Method A uses an approach based on a hybrid Dynamic Bayesian Network that simultaneously infers the number of people in the scene and their locations [12]. The state contains a varying number of interacting person models, each consisting of a head and body model. The person models evolve according to a dynamical model and a Markov Random Field (MRF) based interaction model (to prevent trackers from overlapping). The observation model consists of a set of global binary and color observations as well as individual head silhouette observations (to localize heads). The function of the global binary observation model is to predict the number of people in the scene. Inference is done by trans-dimensional Markov Chain Monte Carlo (MCMC) sampling (because of its ability to add/remove people from the scene and its efficiency).

3.2 Method B : Probabilistic Active Shape (developed at TUM).

Method B uses a double-layered particle filtering (PF) technique [5, 6] consisting of a control layer (responsible for the detection of new people and evaluating the person configuration) and a basic layer (responsible for building a local probability distribution for each head). Locations for new people are derived from skin colored regions, which are detected using a normalized rg skin color model. Heads are modeled using a deformable active shape model consisting of 20 landmark points [1, 2]. The basic layer PF samples and predicts a set of hypotheses for each person. Using the active shape model, a likelihood for the existence of a head in the image represented by the respective hypothesis can be computed. These sets of hypotheses are passed to the control layer PF, which evaluates and determines the configuration of heads by incorporating skin color validation and the local likelihood to verify the number of people being tracked.

3.3 Method C : KLT (developed at BUT).

Method C, proposed in [4] is based on the KLT feature tracker [8]. The method works by searching for potential people through performing background subtraction and skin color detection (using an RG

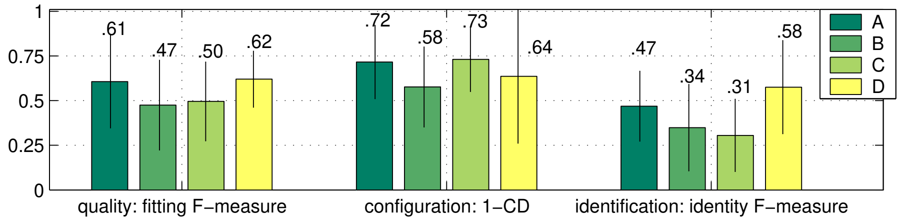


FIG. 3 – Results for the three tracking tasks (spatial fitting, configuration, and identification). The *fitting F-measure* shows the spatial fitting, or tightness of the bounding boxes. The quantity $1 - \overline{CD}$ is indicative of the ability of a method to estimate the configuration. The ability of a method to maintain consistent identities is measured by the *identity F-Measure*. The numbers above each bar represent the mean for the entire data set, and the lines represent the standard deviations.

skin color model) on the raw image. Connected component analysis is performed on the segmented image to find patches suitable for head detection. Ellipse-like shapes are then fitted to the patches and define a set of head centers. A KLT tracker, which extracts meaningful image features at multiple resolutions and tracks them by using a Newton-Raphson minimization method to find the most likely position of image features in the next frame, is initialized at each head center. Additionally, a color cue and rules for flocking behavior (alignment, separation, cohesion, and avoidance) are used to refine the tracking.

3.4 Method D : Face Detector (developed at BUT).
Method D, proposed in [10], is based on skin color segmentation and face detection. A learned skin color model is used to segment the image. Connected component analysis and morphological operations on the skin color segmented image are used to propose head locations. Face detection is then applied to the skin color blobs to determine the likelihood of the presence of a face. The face detection is based on the well-known AdaBoost [14] algorithm which uses weak classifiers to classify an image patch as a face or non-face. Method D replaces the simple rectangular image features with more complex Gabor wavelets [9]. The face detector was trained on normalized faces from the CBCL data set (1500 face and 14000 non-face images) and outputs a confidence, which is then thresholded to determine if a face exists. Faces are associated between frames using a proximity association defined on the positions of the detected faces.

4 Evaluation

The four methods were evaluated for their performance at the tasks outlined in Section 2 : spatial fitting, configuration, and identification. Methods A and B were tested on 360×288 non-interlaced images ; Methods C and D were tested on 720×576 interlaced images after applying an interpolating filter. This discrepancy may affect the relative performance of the methods, but we believe the effect to be minimal. In the following, we present a summary of the overall performance of the tracking methods, followed by a detailed discussion of each task[3].

4.1 Overall Performance
The fitting F-Measure is an indicator of the spatial fitting (see Figure 3). Spatial fitting refers to how tightly the $\mathcal{E}$ bounding boxes fit the $\mathcal{GT}$. The fitting F-Measure is only computed on correctly tracked people, and a value of one indicates perfectly fit bounding boxes. Lower numbers indicate looser, misaligned, or missized tracking estimates. Results for the fitting F-Measure indicate that methods A and D performed comparably well at about .60. Measures B and C performed at approximately .50. The spatial fitting depends on many aspects of the method including the features, motion model, and method of inference. Intuition suggests that the boosted Gabor wavelets of Method D and the head silhouette feature of Method A were most precise in this case, but these results cannot be solely attributed to these features without further experiments.

[3]Example videos and details can be found at http ://www.idiap.ch/∼smith/

The counting distance $\overline{CD}$ measures the difference between the number of $\mathcal{GT}$s and $\mathcal{E}$s for a given frame, and gives an imperfect estimation of the configuration performance, i.e. the ability of the method to place the correct number of $\mathcal{E}$s in the correct locations. $\overline{CD}$ is an imperfect summary because some types of errors such as FPs and FNs may cancel in the calculation of $\overline{CD}$, but it is still a good indicator. The quantity $1 - \overline{CD}$ is reported so that higher numbers indicate better configuration performance ($\overline{CD} \in [0, \infty)$ but in our experiments ranged from 0 to 1). Methods A and C performed best, at about .73, while method D performed at .64, and B at .58. An alternative way to measure the overall configuration performance is to sort the methods by rankings of the individual configuration measures (see Section 4.3 and Figure 5). Doing so, we find that Method C performs the best, followed by Method A, Method D, and finally Method B. Though not necessarily so, in this case this result is consistent with the findings of the counting distance.

The *identity F-Measure* measure indicates how consistently a method was able to identify the $\mathcal{GT}$s over time ; it is a combination of the $\overline{TP}$ and $\overline{OP}$ measures. In this case, method D clearly outperformed the others. This is somehow surprising, as it uses the simplest procedure for maintaining identity (spatial proximity between frames). More sophisticated methods such as models for swapping identities in Methods A and B, are perhaps not suited for this data. One the other hand, because Method D relies on specialized face detection, it's superior performance may not generalize to situations in which faces are not the target objects.

4.2 Spatial Fitting

As mentioned in Section 4.1, the fitting F-measure indicates the tightness of the fit of the bounding boxes to the $\mathcal{GT}$s. From Figure 4, it is apparent that certain sequences presented much more of a challenge than others. Figure 4 illustrates the variation of performance on specific pieces of data, something hidden by all-inclusive measures. Typically, fitting F-Measure values were similar for all the trackers at approximately 0.80, but for more challenging sequences such as 08R, 09R, 12R, 13R, and 16R, differences were more pronounced and fitting F-Measure values dipped as low as 0 in one case. Method D was the most spatially robust for the challenging sequences.

4.3 Configuration

Results for the four configuration error types and $\overline{CD}$ can be found in Figure 5.

The measure $\overline{FN}$ gives an estimation of the number of False Negatives (or undetected person ground truths) per ground truth, per frame. Method C performed the best in this respect, with .26 FN's per person, per frame. This low rate of missed $\mathcal{GT}$s may be attributed to KLT trackers selection of meaningful image features. Method B performed significantly worse, averaging approximately .49 FN, which may be due to difficulties in fitting the contour to the appearance of some heads. FNs were the most prominent type of configuration error among all four tracking methods, usually as a result of an unexpected change in the appearance of a head, partial views, lighting changes, entrances/exits, and size variations and occlusions (sometimes as extreme as in Figure 1).

The measure $\overline{FP}$ estimates the number of False Positive errors (or extraneous $\mathcal{E}$s) per ground truth, per frame. This was the second most common type of configuration error. Typical causes for FP errors include face-like or skin colored objects in the background (texture or color), shadows, and

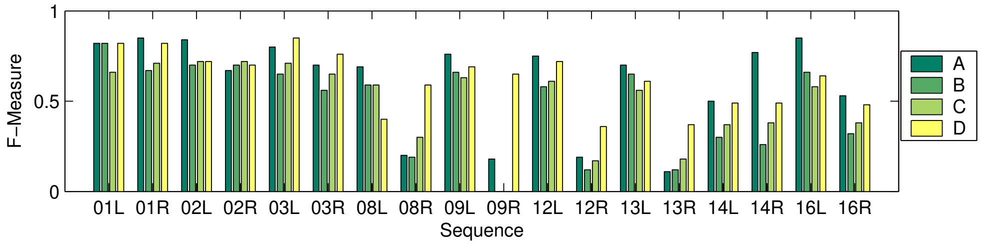


FIG. 4 – The fitting F-Measure shows how tightly the estimated bounding boxes fit the ground truth (when passing the coverage test).

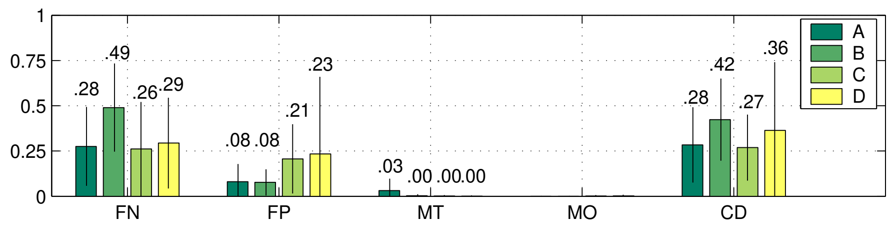


FIG. 5 – The configuration measures, $\overline{FN}$, $\overline{FP}$, $\overline{MT}$, $\overline{MO}$, and $\overline{CD}$, normalized over the test set.

background motion. Methods A and B were least prone to FP errors, with a rate of 0.08 FPs per person, per frame. Method A's low rate of FP errors can be attributed to the use of a body model, which only adds people when a body is detected (bodies are easier to detect than heads). This was followed by Method C with 0.21, and Method D with 0.23. Method D was particularly sensitive to FP generating conditions, as the standard deviation was roughly twice the mean, 0.42. Method D's FPs were generated by face-like or skin colored objects in the background and exposed skin on the arms of the participants.

The measure $\overline{MT}$ estimates the number of Multiple Tracker errors (which occur when several estimates are tracking the same ground truth person). The only method significantly prone to this type of error was Method A. This susceptibility is due to the fact that Method A uses strong priors on the size of the body and head to help the foreground segmented image features localize the head. The priors of Method A are trained using participants in the far field of view, and are not robust to dramatic changes in size. When a participant appears close to the camera, Method A often fits multiple trackers to the larger head area. Methods B,C, and D do no suffer from this effect because they do not enforce constraints on the size of the head so strongly.

The measure $\overline{MO}$ estimates the number of Multiple object errors (which occur when one estimate tracks several ground truths) per person, per frame. This type of error generally occurs when a tracker estimate is oversized and expands to cover large areas of the image, or occasionally when people are near one another. All four methods tested were robust to this type of error. This robustness can be attributed to the modeling of head objects, interaction models, and motion models built into each of the methods.

The counting distance measure $\overline{CD}$ is described in Section 4.1.

4.4 Identification

Results for the identification measures can be found in Figure 6.

The $\overline{FIO}$ measure estimates the rate of Falsely Identified Object errors (when an $\mathcal{E}$ tracks a $\mathcal{GT}_k$ which is not the $\mathcal{GT}_j$ that the $\mathcal{E}$ *identifies*). Of the two types of identification errors (FIO and FIT), FIO errors occurred less frequently for all four methods. FIO errors are often generated when an $\mathcal{E}$ outlives the $\mathcal{GT}$ it is supposed to identify, and the $\mathcal{E}$ begins to track another $\mathcal{GT}$, though this was rare in our experiments. The other common mode of failure occurred when $\mathcal{E}$s confused $\mathcal{GT}$s, often as a

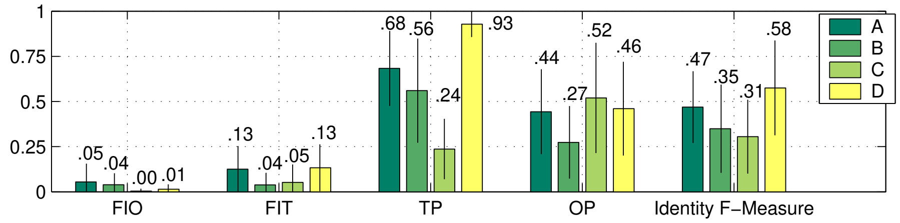


FIG. 6 – The identification measures, $\overline{FIO}$, $\overline{FIT}$, $\overline{TP}$, $\overline{OP}$, and *identity F-Measure* computed over the test set.

result of occlusion. This method of failure was seen most in Methods A and B with $\overline{FIO}$ rates of 0.05 and 0.04, respectively. Interestingly, both these methods modeled *identity swapping*, where $\mathcal{E}$s switch labels in an attempt to maintain identity. Spurious identity swaps could account for higher FIO rates. Method C was very robust to FIO errors, with a negligible FIO rate. Method D was nearly as robust, with a $\overline{FIO}$ of 0.01.

The $\overline{FIT}$ measure reports the rate of Falsely Identified Tracker errors (which occur when a $\mathcal{GT}$ person is being tracked by a non-identifying $\mathcal{E}$). There are two typical sources of FIT errors. The first occurs, as with the FIO error, when $\mathcal{E}$s swap or confuse $\mathcal{GT}$s. The second error source occurs when several short-lived $\mathcal{E}$s track the same $\mathcal{GT}$s. Both of these sources caused FIT errors in our test set, though it can be expected that FIT contributions from the first error source should roughly match the FIO error rate (and thus, any increase in the FIT over the FIO is caused by short-lived $\mathcal{E}$s). Methods A and D saw the most FIT errors, with $\overline{FIT}$ rates at 0.13 (0.13 FIT errors are generated per frame, per person). Method D's FIT errors can be almost exclusively attributed to multiple, short-lived $\mathcal{E}$s tracking the same $\mathcal{GT}$. Method B was the most robust to FIT errors with a rate of 0.04.

The $\overline{TP}$ measure evaluates the consistency with which an $\mathcal{E}$ identifies a particular $\mathcal{GT}$. Misidentified $\mathcal{GT}$s cause FIO errors, but the TP measure gives equal weight to all tracking estimates. $\mathcal{E}$s with a short lifetime will not significantly influence the $\overline{FIO}$, and $\mathcal{E}$s with long lifetimes will dominate. Typically, in our experiments, the methods reported a higher $\overline{TP}$ than $\overline{OP}$. This indicates more $\mathcal{E}$s were generated than the number of $\mathcal{GT}$s in the sequence (in a temporal sense), and that they lasted for shorter lifetimes. Method D reported a $\overline{TP}$ of 0.93, which indicates that nearly all its $\mathcal{E}$s perfectly identified their $\mathcal{GT}$s. However, this does not indicate near-perfect identification. Method D's $\overline{OP}$, 0.46, while on par with the other methods, indicates that the $\mathcal{GT}$s were often tracked by multiple short-lived $\mathcal{E}$s. Method A reported the next highest $\overline{TP}$, with a value of 0.68, followed by Method B (0.56) and Method C (0.24). Method C was the only method to report a lower $\overline{TP}$ than $\overline{OP}$.

The $\overline{OP}$ measure evaluates the consistency with which a $\mathcal{GT}$ is identified by the same $\mathcal{E}$. Misidentifying $\mathcal{E}$s can cause FIT errors, but OP gives equal weight to all $\mathcal{GT}$s in the sequence. Short-lived $\mathcal{GT}$s will not significantly affect the $\overline{FIT}$, and $\mathcal{GT}$s with a long lifetime will dominate. Method C reported the best $\overline{OP}$.

4.5 Summary and Qualitative Comments

Giving equal weight to the three tracking tasks described in this document (configuration, identification, and spatial fitting) and using a simple ranking system, the best performing tracking method is D, followed by A, C, and B. Method D is the most reliable at identification and exhibits the highest spatial fitting. However, it does have several drawbacks. It is the slowest of the four methods and the most sensitive to occlusion. The face detector is based on skin color detection and is more sensitive to lighting conditions than the other methods. Skin-colored segments of the background pose a problem for the face detector (Method D exhibits the highest $\overline{FP}$), and the $\overline{FN}$ suffers as the detector struggles with non-frontal faces.

Ranked second among the four methods is Method A. Method A was the only method which did not model skin color, and was the only method which modeled the body to help localize the head. The use of a body model had several effects. First, Method A had the lowest $\overline{FP}$ rate, which can be attributed to the body model preventing spurious head $\mathcal{E}$s. The body model assisted in detecting heads, which kept the $\overline{FN}$ rate low. However, because of strong size priors on the head and body models, Method A performed poorly when tracking heads near the camera (resulting in MT errors). Method A was ranked second in spatial fitting and was also ranked second in maintaining identity, though incorrect swapping of $\mathcal{E}$ labels may have lowered this performance.

Method C was third overall among the four methods. It was the fastest computationally ; the only one approaching real-time frame rates. Method C had the highest configuration performance, boasting the lowest FN rate and negligible MT and MO errors. This can be attributed to the KLTs selection of meaningful image features. However, Method C performed worst in terms of spatial fitting and identification. The poor spatial fitting might be due to a lack of shape features or features specialized to the face (as in the face detector). Problems with identification were due to the lack of an explicit way to manage identity among the trackers.

Finally, Method B fell last overall, but ranked third for each of the three tracking tasks. In terms of spatial fitting, Method B was the highest performing method for several of the sequences, but suffered from poor performance on some of the more difficult multi-person sequences (12R, 14R, and 16R). Among the four trackers, the Method B was the most robust to partial occlusions. For Method B, identity was maintained by binning gray values of the face shape. A lack of color information, poor shape adjustment, and a swapping mechanism like that of Method A, may have caused identification problems for this method.

From this evaluation, we might draw some of the following conclusions :

1. Shape-based methods, such as B and C, perform as well or better at spatial fitting when stable, but are more prone to configuration failures, and less able to recover from such failures.
2. Methods employing background subtraction (such as A and C) seem to have an advantage estimating the configuration of the scene.
3. Attempts to model identity changes to handle difficult tracking scenarios such as dramatic changes in size and appearance or frequent occlusions may do more harm than good (as for Methods A and B).

5 Conclusion and Future Work

real-life scenarios which remain challenging for state-of-the-art tracking methods. These results represent the first evaluation of methods for multi-person tracking in meetings using a common data set in the context of the AMI project. Future work might incorporate multi-model information or concentrate on tracking other objects in different scenarios.

Acknowledgements This work was supported by the Swiss National Center of Competence in Research on Interactive Multi-modal Information Management (IM2), and the EC project Augmented Multi-party Interaction (AMI, publication AMI-175).

Références

[1] T. Cootes and C. Taylor, *Statistical models of appearance for computer vision*, 2004.

[2] T. Cootes, G. Edwards and C. Taylor, "A comparative evaluation of active appearance model algorithms", *British Machine Vision Conference*, Southampton, UK, Sept. 1998.

[3] D. Gatica-Perez "Annotation Procedure for WP4-locate", *AMI Internal Document*, Martigny, Switzerland, October 2004.

[4] M. Hradis, R. Juranek, "Real-time Tracking of Participants in Meeting Video", *Proceedings of CESCG*, Wien, 2006.

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

[6] M. Isard and A. Blake, "A Mixed-State CONDENSATION Tracker with Automatic Model-Switching", *International Conference on Computer Vision (ICCV)*, 1998.

[7] R. Kasture et. al., "Performance Evaluation Protocol for Face, Person, and Vehicle Detection & Tracking Analysis and Content Extraction (VACE-II)", ARDA Techinical Report, Tampa, FL, 2006.

[8] M. Kölsch and M. Turk, "Fast 2D Hand Tracking With Flocks and Multi Cue Integration", Department of Computer Science, University of California, 2005.

[9] V. Kruger, "Wavelet Networks for Object Representation," thesis dissertation, Technischen Fakultat, Christian-Albrechts-Universitat zu Kiel, 2000.

[10] I. Potucek, S. Sumec, M. Spanel, "Participant activity detection by hands and face movement tracking in the meeting room", *Computer Graphics International (CGI)*, Los Alamitos, 2004.

[11] C.J. Van Rijsbergen, *Information Retrieval*, Butterworth-Heinemann, Newton, MA, USA, 1979.

[12] K. Smith, S. Ba, J.M. Odobez, D. Gatica-Perez, "Multi-Person Wander-Visual-Focus-of-Attention Tracking", *IDIAP-RR-05-80*, Nov 2005.

[13] K. Smith, S. Ba, J.M. Odobez, D. Gatica-Perez, "Evaluating Multi-Object Tracking", *CVPR Workshop on Empirical Evaluation Methods in Computer Vision (EEMCV)*, San Diego, CA, June 2005.

[14] J. Viola and M. Jones, "Robust Real-time Object Detection", Technical Report 2001/01, Com-paq CRL, February 2001.

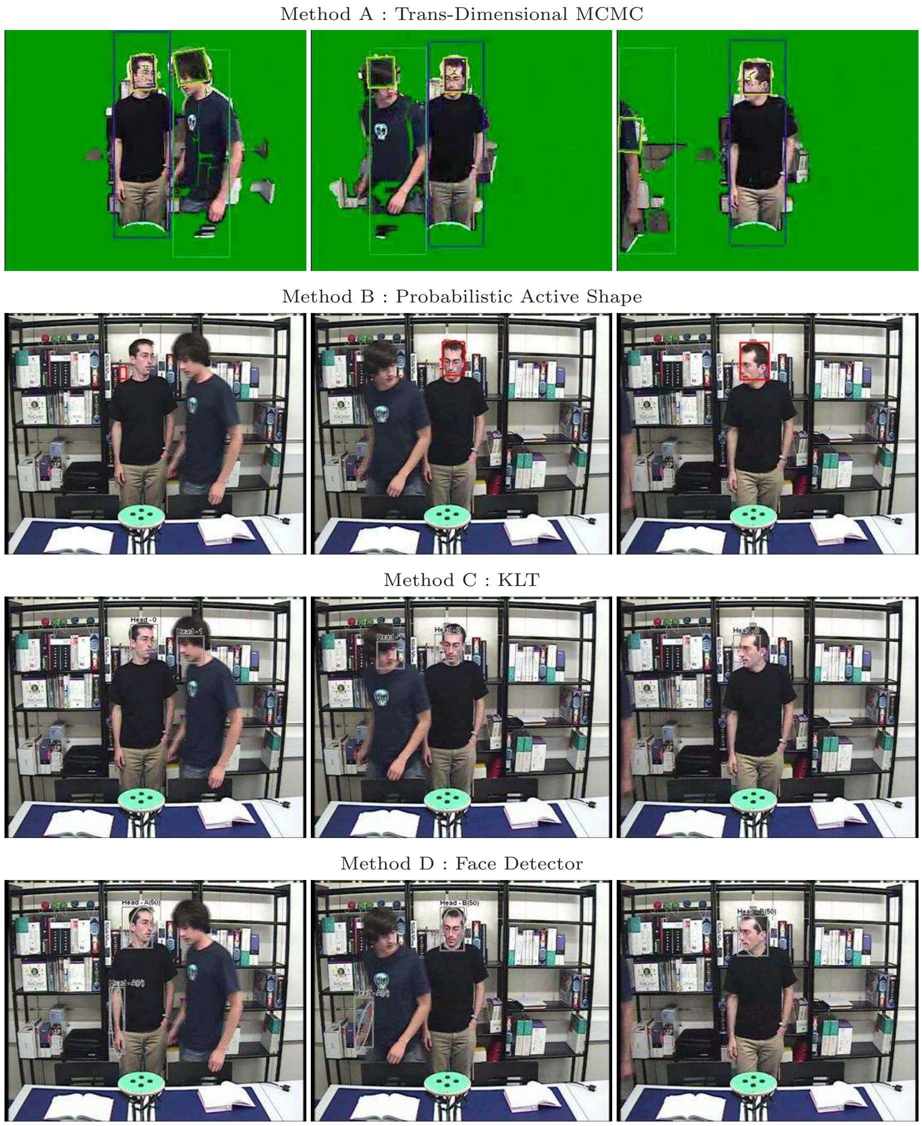


FIG. 7 – Results for frames 307, 333, and 357 of sequence 09L from the AV16.7.avi data corpus. Method A : body and head results shown. A FP error appears in frame 357. Method B : heads results appear as red bounding boxes. Two FN errors and an FP error occur in 307, and one FN error occurs in 333. Method C : head results appear as grey bounding boxes. Method D : results appear as grey bounding boxes, participant arms are mistaken for heads in 307 and 333.