A Comparison of Multicamera Person-Tracking Algorithms

A.W.Senior, G.Potamianos, S.Chu, Z.Zhang, A. Hampapur
aws @ us.ibm.com
IBM T. J. Watson Research Center,
PO Box 704,
Yorktown Heights, NY 10598, USA.

Abstract

In this paper, we present a comparison of four novel algorithms that have been applied to the tracking of people in an indoor scenario. Tracking is carried out in 3D or 2D (ground plane) to provide position information for a variety of surveillance, HCI or meeting-support services. The algorithms, based on background subtraction, face detection, particle filter feature-matching and edge alignment of a cylindrical model are described and comparative results presented using independent test data produced and ground-truthed for the EU CHIL project.

1 Introduction

A number of applications require the visual tracking of people, from automated visual surveillance systems to interfaces for human-computer interaction. In visual surveillance, position is an important cue for indexing and real-time event detection. In the automated indexing of meetings and seminars, one of the most important pieces of information to be extracted is the location of the participants, since this allows the roles of the participants to be understood, and for resources (such as steerable microphones or close-up cameras) to be directed at particular individuals, as well as permitting indexing by location.

In the past, we have principally applied our tracking technology in the 2D surveillance domain, but here present a use of the same algorithms in a meeting-mining scenario, and using a new 3D tracking system to combine results from different camera views. This algorithm is compared with three other algorithms for the same task, two described for the first time. We compare the strengths and weaknesses of the four approaches and demonstrate performance using the independent data collected and ground-truthed by the CHIL (Computers in the Human Interaction Loop[1]) EU consortium.

[1] `chil.server.de`

2 Related work

In recent years a large body of work has sprung up in multimodal sensing of environments and the interpretation of sensor data for human-computer interfaces and for indexing. "Meeting mining" has been one area of particular interest with systems to index and summarize meetings and presentations. Several systems to do this have exploited visual information, particularly for automatic camera control [7]. Other systems have used person tracking to guide biometric (face) acquisition [15]. Such systems may use an omni camera, or a PTZ camera colocated with a wide-angle camera, or even "virtual cameras" [12] made from selective sampling from a wide-angle, high-resolution camera. Calibrated multicamera systems have been used for person localization and more detailed tracking of heads, hands and full articulated body tracking [2].

Stereo tracking has been used by a number of systems, more usually focused on human-computer interaction with the user being located in a limited space in front of a screen. Notable among these is pFinder [13] that uses triangulation of coloured regions to determine the position of the head, hands and body of the user.

For the scenario considered in this paper, the principal task is to locate the speaker, both in 2D (position on the ground plane) and 3D (head location). The latter is most useful for steering sensors to acquire better information (telephoto cameras, or directional — beamforming or shotgun — microphones). The CHIL task envisages a multimodal approach to meeting mining problems, and we anticipate combining person tracking with acoustic source localization to locate all the people (particularly speakers) in a room. PTZ cameras can be directed to acquire images of areas of visual interest, particularly faces for face recognition, facial expression recognition and audio-visual speaker identification [8]. Images of people acquired through active camera control can also be used for understanding hand gestures or gait recognition.

3 Tracking

In this paper we present four different approaches that have been taken to tracking a person in indoor environments instrumented with multiple cameras with overlapping fields of view. This section describes those approaches. Section 4 then describes the CHIL evaluation data used to compare the systems, and Section 5 gives results for these trackers and other trackers that were evaluated on the same data.

3.1 Background subtraction tracker

The first method that we have used for tracking the speaker is a new 3D tracker built on top of the IBM Smart Surveillance Engine (SSE) 2D (image plane) tracker. In this case we apply the 2D tracker to each of the available views.

The 2D is based around a background-subtraction object detection system (described in [3]) which uses a multiple Gaussian colour model at each pixel. Objects are tracked in the image plane from frame to frame using the Colour-Field tracking method described in [9]. A preliminary 3D tracking system called the Face cataloger [4] applied this tracker independently in two, nearly orthogonal views, and used a head detection algorithm to determine the center of the head regardless of the view direction. The estimated head points from pairs of tracks in the 2D views were then triangulated to determine correspondences between tracks and to calculate a 3D position for each head center. This was then used to steer a calibrated pan-tilt-zoom camera to the head position to allow close-ups of the head and face to be captured. A face detection system guided a closed-loop control system to zoom in even closer if the face was visible from the steered camera, overcoming errors in localization and calibration in the fixed cameras. A wide-angle head pose algorithm giving absolute head orientation on heads as small as 8x8 pixels could be used to determine which PTZ camera to schedule to obtain a face image.

Because of improvements in the underlying 2D tracking algorithm, for the CHIL task we designed a new 3D tracking algorithm, and dispensed with the head detector which had limitations when multiple targets were being tracked, and was found to be unnecessary in seminar situations where the head is almost always the highest point of the speaker's body. In this version of the tracker, the underlying 2D tracking algorithms are again taken, unmodified from their usual outdoor surveillance configuration. The 2D tracker is used only to provide a temporally-smoothed model of the objects observed in each view, together with each object's location, tracked through occlusions. The 2D track information is not used in the 3D engine, so temporal consistency is applied independently in 3D.

At each frame our 2D tracker is applied and we use the resulting 2D probabilistic models to determine the position of the head top. This is the point whose y coordinate is the top of the object model bounding box and whose x coordinate is that of the centroid of the upper sixth of the model. This assumes that people's projections in the camera views have the head uppermost, but are not necessarily vertical.

These 2D object points are considered as hypotheses for the top of the speaker head, and when coupled with the camera calibration information, each gives a 3D ray along which the speaker's head might lie. We seek validation for these hypotheses in other views by computing the shortest distance d_{r_1,r_2} between each pair (r_1, r_2) of such rays from different cameras. All such pairings are evaluated, sorted and compared to a distance threshold D (300mm), with the closest match being considered first. The following algorithm is used:

1. Start with a pool of ray hypotheses for each camera.
2. If $d_{r_1 r_2} < D$ create a 3D hypothesis at the midpoint $\mathbf{p}_{r_1,r_2}$ of the shortest inter-ray line.
3. Search the remaining cameras for other rays r_3 that pass this point within a distance of D, move the hypothesis to $\mathbf{p}_{r_1,r_2,r_3,...}$ the least squares fit for all rays and repeat for any remaining cameras.
4. Store the hypothesis $X_i(t)$, and remove from the pools the rays just used in this 3D hypothesis.
5. Repeat from step 2

Figure 1 shows the background subtraction results on simultaneous frames from four cameras, marked with 2D head-top candidates and the back-projected locations of a 3D candidate, as well as the current best track.

These 3D hypothesis points can be associated over time and concatenated into 3D tracks. Given the constraints of the problem (that we are only required to track the speaker, who is present throughout the video sequences) we have used a Viterbi decoding approach to finding the best track hypothesis through the sequence of 3D head-top hypotheses. The Viterbi decoder uses a beam search with N (typically 50) search hypotheses to search for the shortest path passing through head location hypotheses. A path P consists of a time-ordered sequence of n head locations. $P = \{X_{i_1}(t_1), X_{i_2}(t_2), ..., X_{i_n}(t_n)\}$ where $t1 < t_2 < \ldots < t_n$. The cost $C(H,t)$ of such a path at frame t is based on the length between consecutive points, with penalties for each frame that the path does not explain a hypothesis: C_E at the start or end of the track and C_I within a track.

$$C(H,t) = t_1 C_E + (t - t_n) C_E + \sum_k \|X_{i_k}(t_k) - X_{i_{(k-1)}}(t_(k-1))\| + (t_k - t_{k-1} - 1) C_I \quad (1)$$

Figure 1: Detection results for the BGS-based tracker on four views for one frame. Foreground blobs are shown in solid green. Candidate head-top points are shown as small orange circles. 3D hypothesis locations are shown back-projected as larger blue circles. The current Viterbi path is shown as a green line.

At each frame we update all of the paths, where each path can be retained with no additional evidence (with a penalty) or by adding on one of the 3D location hypotheses for that frame. This produces many more search hypotheses which are sorted by cost, and only the top N retained for the next frame. At the end of a sequence the single lowest cost path is retained as the "best" path through the 3D location hypotheses. For comparison with the ground truth we use linear interpolation between hypotheses (and point repetition at the start and end) to generate a position for each frame.

The CHIL data contains segments of data that start in mid-seminar. In practice the cameras would have been on continuously beforehand and we would have had the chance to build a background model over a period of time, some of which would have been when the room was at least partially empty. To overcome the lack of background models, we supply the system with a single image created by splicing frames from the development set together so as to remove the speaker. (Done automatically using the development set speaker location ground truth. The authors are grateful to Aristodemos Pnevmatikakis for supplying us with these images). In practice because of the large amount of audience movement in some sequences, the fixed camera geometry and the requirement that the system only track the speaker, we have defined "regions of disinterest" — hand drawn regions where detections are ignored, based on the position of the speaker in the development data.

3.2 Particle filter tracker

Our second tracker uses a radically different approach to the tracking of the speaker. This novel tracker was inspired by the particle filter tracker of Nickel *et al.* [6], and follows their approach of having a population of hypotheses as to the two-dimensional, ground-plane location of the speaker (which was the desired output for the evaluations). These hypotheses are treated as particles in a particle filtering system [16] whereby each hypothesis is evaluated for how well it matches the available data, and then the distribution is re-sampled at every frame to generate a new set of hypotheses. We propose a number of novel changes to their system, notably avoiding the use of background subtraction, and using a cylindrical model.

The system of Nickel *et al.* evaluates each hypothesis by constructing a cuboid centred on each hypothesis, and considered to approximate the visual appearance of the person. For each camera view, this cuboid is projected into the image, and the total foreground weight (using an unthresholded Stauffer & Grimson [11] style background subtraction algorithm) within the projected area, approximated for speed by a rectangle.

Our prime motivation for adopting the particle filtering approach was that it allowed us to avoid the use of background subtraction. Here for speed, simplicity and to avoid all the problems (particularly static objects and starting mid-seminar without a background model) associated with background subtraction, we used simple frame-differencing, where the principal data measurement was the absolute difference, $\delta I(x)$ in image frames at each pixel:

$$\delta I \quad = \quad |R_t - R_{t-1}| + |G_t - G_{t-1}| + |B_t - B_{t-1}|, \quad (3)$$

where $R_t(x), B_t(x), G_t(x)$ are the red, green and blue intensities of a pixel x at time t. This dispenses with initialization problems and moved-objects, but means that tracking is only carried out actively while the speaker is moving. Nevertheless, continuity constraints can be applied to maintain the speaker's location when no visual activity is detected.

Figure 2 shows the difference images obtained in all four views for one frame in the 2003 data set.

Now, unlike with background subtraction, we do not expect these pixels to be evenly distributed across the person's projected area, rather they should be at the sides of a person. More specifically they will lie along the leading and trailing edges of regions of uniform colour, which will be along the sides as a person walks, and the edges of their arms (also tending to lie close to the person's sides) when gesturing. So, we project bounding edges of the person (we use a vertical cylinder model and find the left and right edges in each camera view) and use a chamfer distance to weight more heavily difference pixels that are close to the hypothesized edge of the person. We use integral images (popularized

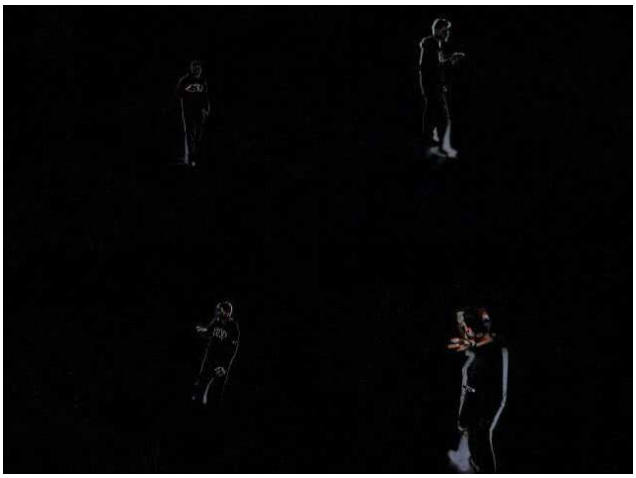

Figure 2: Colour absolute difference $\delta I(x)$ images for the four camera views shown in Figure 1.

for face detection by [5]) for speed in evaluating the chamfer distances. Figure 3 shows the evaluation of the chamfer distance. Here we assume people will have near-vertical projections, but the width of the chamfers makes the system robust to failures of this assumption.

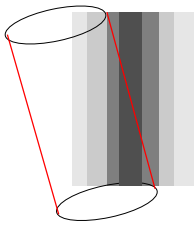

Figure 3: Diagram showing the stepped chamfer distance $w_p(x)$ (darker grey means lower distance) for pixels x near the sides of the projected cylindrical person model corresponding to particle p.

Each particle, p, is assigned a weight $\omega(p)$ by summing the weighted image differences across all v views available:

$$\omega(p) \quad = \quad \sum_v \sum_x \delta I(x) w_p(x) \qquad (4)$$

Thus particles in locations which are supported by the evidence (significant image difference at the edge of the hypothesized location in all views) are weighted more heavily.

As suggested by Nickel *et al.* we also evaluated a face detection system, counting face matches in the upper third of the projected person location as strong evidence for a person. However the face detector was not found to be sufficiently accurate or fast enough to be useful, and results are not presented here.

Figure 4 shows the four camera views of the meeting scenario with vertical green lines, the bottom of each showing the position of a hypothesis and the heights indicating relative weights.

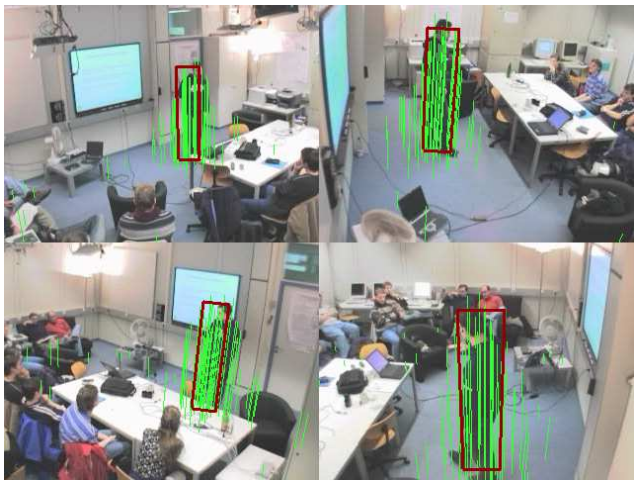

Figure 4: Four camera views showing the hypothesis positions and their relative weights (heights of green lines), as well as the projection of the mode's cylinder into each view as a red rectangle.

After each frame is processed, we attempt to find the mode of the distribution by applying a mean shift algorithm, starting from the mode found in the previous frame. The mode found is used as the speaker location for the current frame. Finally particles are resampled. We have tried a number of ways of doing this resampling: mixture-of Gaussians, mode-based, and perturbation.

The mixture-of-Gaussians method clusters the current samples to approximate the weight distribution with a mixture of Gaussians. A new sample set is then drawn from this distribution with a small probability of taking a particle from a uniform background distribution. All particles are constrained to lie in the area of the room. Mode-based resampling again draws a new set of particles and approximates the distribution with a fixed-variance Gaussian centred on the most recent mode, together with a background distribution. Finally perturbation perturbs the current set of particles randomly, and randomly removes some, adds others from a uniform distribution or creates copies of particles (with additional perturbation) drawn randomly, with a weighting to prefer those that scored highest.

Results are presented on the mode-based method which has been found to perform best.

3.3 Face detection-based tracker

The face detection-based tracker again uses a completely different approach to the problem of person localization, though it is an approach adopted by a number of other partners in the CHIL project. Here the tracker attempts to localize the face of the speaker and uses triangulations of 2D face positions to determine the 3D head position. This tracker is described in more detail elsewhere [14], but is built around

a FloatBoost face detection system which is an extension of that of Jones and Viola [5]. Aside from differences in the training method used, the system uses two face detectors — one for frontal faces and the other for side-views of faces. The face detectors are applied to moving regions of video frames, but once a face is detected, it is tracked using a mean-shift algorithm on the histogram of hues.

3.4 Edge-based body tracker

A fourth method for tracking the speaker is to use a 3D, model-based tracker that we have developed for articulated body tracking [10]. This uses a 3-dimensional graphical model of a human (with limb parts being represented by cylinders connected in kinematic chains) which is aligned to the video data in multiple views using a nonlinear optimization approach to minimize the distance between model edges and observed image edges. For the CHIL evaluation, no articulations are required, but the optimization framework is directly applicable to tracking a simple 2 degree of freedom cylindrical torso model (like that used in the particle filter tracker). More complex articulated models can be fit by the same framework to detect joint angles, for instance to detect hand gestures.

In previous work we have used a variety of features to fit the model to the video data, but here we use image edges, with edge domain background subtraction to remove background edges which often dominate over the foreground edges. Again, as in the particle filter tracker, the parameter space is the 2-dimensional ground plane, with a fixed height for the speaker, that does not need to be accurate.

The cylindrical model is iteratively aligned to the observed image edges, assuming an approximate fit from the previous frame. The model is matched to the foreground region of the current frame by the following procedure:

1. The (vertical) model edges are projected into each view, and sampled at regular intervals (e.g. every pixel).

2. From each of these sample points, a search is conducted perpendicular to the model edge, to find the maximum gradient within a search window.

3. These displacements, transformed according to the pose of the body part and the camera view are aggregated into a single nonlinear equation which can be solved by iterative least squares to determine the displacements of all the joints minimizing the displacements to the found gradient maxima, as described by Bregler and Malik [1].

The 2D body tracking required by the CHIL project is much simpler, but we can apply the full articulated body tracker with the trivial model, and a corresponding speed increase because there are no kinematic chains and only two degrees of freedom.

Figure 5: A single camera view showing the projected cylinder model (green) and found edge pixels (white).

The main limitation of this approach is that it needs initialization and reinitialization if it should lose track. In the system evaluated here, no reinitialization is carried out, and initialization of the model in the first frame is done by hand. In future work we plan to augment this method by initializing and re-initializing it with detection from one of the other methods.

4 Data

The data on which we have analysed the tracker performance was collected as part of the CHIL project — a consortium of European Union institutions. All the initial data were collected by Karlsruhe University in their smart meeting room (see Figure 6), and consists of video from four calibrated, static cameras mounted in the corners of the 5.9m by 7.1m room. The speaker is visible from all cameras in nearly all frames. The speaker stands in front of an audience partially visible from each of the cameras. A projector is used in the seminar, whose images change and sometimes the speaker obstructs the beam of the projector.

640 by 480 pixel video data is captured at 15 frames per second from each camera. For many of our experiments, the video is downsampled to 320 by 240 before use. Synchronized acoustic data is also available from a variety of microphone arrays, for use in acoustic speaker localization, source separation, speaker identification and speech recognition, as well as the multimodal equivalents for all of these tasks. In this work only the video data has been used. The data are divided into “dry-run” data recorded in 2003 and “evaluation” data recorded in 2004, all of which were used in a consortium-wide evaluation in December 2004. The seminars of each dataset are divided into segments for development and test. Results are presented on the test segments only.

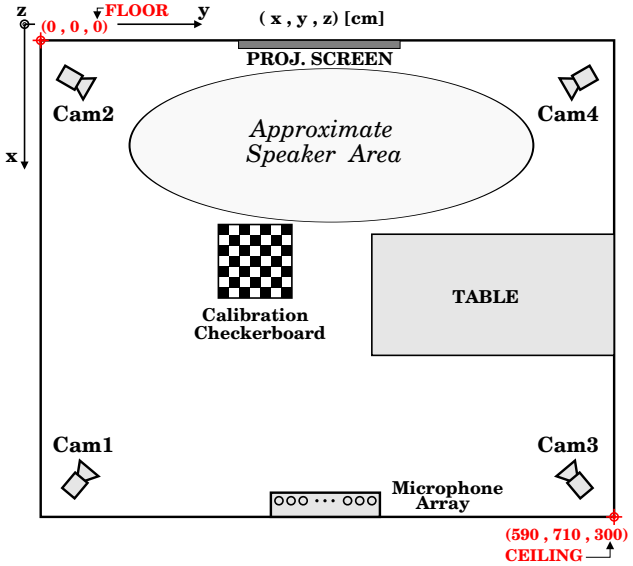

Figure 6: Plan of the meeting room used for collection of the test data used.

The ground truth for the video data consists of the hand-marked position of the centre of the head in each view, together with a bounding box of the face when the face is visible. Ground truth is available for every 10th frame. These results were triangulated to produce a ground truth 3D head position distributed to all consortium members. The calibration parameters were not distributed, but the calibration images were, so the experiments here were carried out using calibration from our own calibration tool.

The data for each seminar are split into development (segments 3,4) and evaluation sets (segments 1 & 2). Results are presented averaged across all the evaluation data for each of the 2003 and 2004 data sets. The 2004 test set consists of ten segments totalling 4674 groundtruthed frames.

5 Results

The results are presented in table 1 for the four systems described above.

Method	2003 Error (mm)	2004 Error (mm)
BGS	218	438
Face	228	441
Particle Filter		788
Edge tracker		442

Table 1: Average error (distance from ground truth point) for the four methods described above on 2003 and 2004 evaluation data sets.

The edge tracker is comparable to the BGS and face-based trackers, but its behaviour is significantly different. Nine of the ten sequences are tracked very well with a mean error of 294mm. In the tenth sequence (Seminar `2004-11-11_C`, segment 1) which starts with the speaker out of the field of view of one camera and almost out of sight of another, tracking fails after about 200 frames and leads to a large average error.

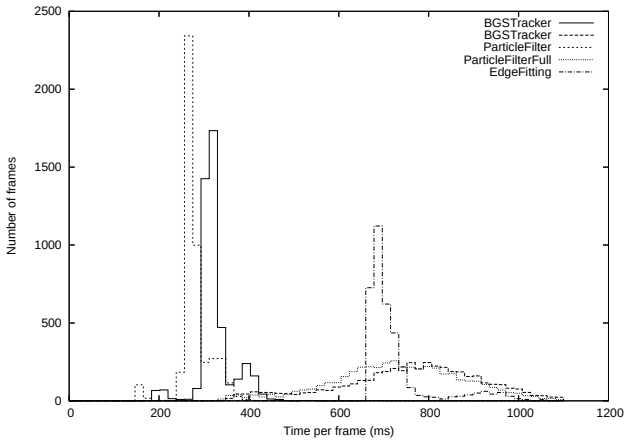

Figure 7: Histogram of execution times per frame (on a single 2.4GHz Pentium processor) on the first sequence of Seminar A, 11/11/2004. Frame times are shown for three of the trackers, with full (640x480) or half (320x240) frames.

Figure 7 shows a comparison of execution times on a 2.4GHz Pentium machine, using full or half resolution. The face detection based tracker runs at 5fps on a 2.8GHz Pentium machine. By comparison, the visual system of Nickel *et al.* runs at 5.6fps on quarter resolution frames (except for face detection) on a 3.0GHz machine, and report an average error of 363mm, reduced to 305mm when augmented with acoustic data.

6 Conclusions

The different trackers each have their strengths and weaknesses. The particle filtering approach is attractive because it does not rely on the construction of, and differencing from, a background model, which can lead to problems when there are many moving objects and the speaker is often static. However, use of frame differencing only is noisy and subject to distraction, for instance when a static speaker raises an arm. It should be robust to errors, and does not fall into local minima, but can easily be mislead for a short while if the speaker is static and there is other activity in the scene. Short-term changes (like lighting or slide changes) have short-lived effects on the tracker. There is potential for extending this approach to track multiple targets, though occlusion handling is difficult without making the evaluation

much more complex and the feature space much larger (two dimensions per candidate). The balance between applying particles to tracking existing hypotheses and exploring the space for new hypotheses is tricky if the number of particles is to be kept within "real-time" limits, and the mechanism for achieving this balance needs more study to achieve good tracking results.

The background subtraction method relies on having and maintaining a good background model. Currently we do not exploit the 2D tracking capabilities of the SSE. The face tracking system relies on face detection, which is not perfect, and cannot be guaranteed with fewer than four cameras, but here works well, and indeed leads to the best system we have hitherto reported on the CHIL data. The two results are complementary and it would seem that they could be combined into a more accurate hybrid tracker.

Finally, the edge-alignment technique works very well once initialized, but does not recover from tracking failures. A robust initialization and re-initialization procedure would make this system attractive, because of its accuracy and its easy expansion to many more degrees of freedom. This system can also be applied to the tracking of multiple objects using the ambiguity resolution procedure that we have described [10]. A combination, using the particle filtering approach for detection and initialization and the edge-alignment for tracking may be feasible.

References

[1] C. Bregler. Learning and recognizing human dynamics in video sequences. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Juan, Puerto Rico*, pages 568–574, 1997.

[2] C. Bregler and J. Malik. Tracking people with twists and exponential maps. In *Conference on Computer Vision and Pattern Recognition*, 1998.

[3] J. Connell, A.W. Senior, A. Hampapur, Y.-L. Tian, L. Brown, and S. Pankanti. Detection and tracking in the IBM PeopleVision system. In *IEEE ICME*, June 2004.

[4] A. Hampapur, S. Pankanti, A. Senior, Y.L. Tian, L. Brown, and R. Bolle. Face cataloger: Multi-scale imaging for relating identity to location. In *IEEE Advanced Video and Signal-Based Surveillance*, pages 13–20, Miami, July 2003. IEEE Computer Society.

[5] P. Jones and M. Viola. Robust real time object detection. In *Workshop on Statistical and Computational Theories of Vision*, 2001.

[6] Kai Nickel, Tobias Gehrig, Rainer Stiefelhagen, and John McDonough. A joint particle filter for audio-visual speaker tracking. In *International Conference on Multimodal Interfaces ICMI 05*, 2005.

[7] C. Pinhanez and A. Bobick. Intelligent studios: Using computer vision to control TV cameras. In *Workshop on Entertainment and AI/Alife*, pages 69–76, August 1995.

[8] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A.W. Senior. Recent advances in the automatic recognition of audiovisual speech. *Proceedings of the IEEE*, 2003.

[9] A Senior. Tracking with probabilistic appearance models. In *Third International workshop on Performance Evaluation of Tracking and Surveillance systems*, pages 48–55, June 2002. ECCV workshop.

[10] A.W. Senior. Real-time articulated human body tracking using silhouette information. In *IEEE Workshop on Visual Surveillance/PETS*, Nice, October 2003.

[11] C. Stauffer and W. E. L. Grimson. Adaptive background mixture models for real-time tracking. In *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Fort Collins, CO, June 23-25*, pages 246–252, 1999.

[12] M. N. Wallick, Y. Rui, and L. He. A portable solution for automatic lecture room camera management. In *ICME*. IEEE, 2004.

[13] C.R. Wren, A. Azarbayejani, T. Darrell, and A. Pentland. Pfinder: Real-time tracking of the human body. *IEEE Trans. Pattern Analysis and Machine Intelligence*.

[14] Z. Zhang, G. Potamianos, A. Senior, S. Chu, and T. Huang. A joint system for person tracking and face detection. In *submitted to ICMI*, 2005.

[15] Xuhui Zhou, Robert T. Collins, Takeo Kanade, and Peter Metes. A master-slave system to acquire biometric imagery of humans at distance. In *First ACM SIGMM International Workshop on Video Surveillance*, 2003.

[16] D. Zotkin, R. Duraiswami, and L. Davis. Multimodal 3-D tracking and event detection via the particle filter. In *IEEE Workshop on Detection and Recognition of Events in Video*, pages 20–27, 2001.