

An Appearance-based Particle Filter for Visual Tracking in Smart Rooms

Oswald Lanz and Roberto Brunelli

Fondazione Bruno Kessler - IRST,
Via Sommarive 18,
38050 Povo di Trento,
Italy

Abstract. This paper presents a visual particle filter for tracking a variable number of humans interacting in indoor environments, using multiple cameras. It is built upon a 3-dimensional, descriptive appearance model which features (i) a 3D shape model assembled from simple body part elements and (ii) a fast while still reliable rendering procedure developed on a key view basis of previously acquired body part color histograms. A likelihood function is derived which, embedded in an occlusion-robust multibody tracker, allows for robust and ID persistent 3D tracking in cluttered environments. We describe both model rendering and target detection procedures in detail, and report a quantitative evaluation of the approach on the 'CLEAR'07 3D Person Tracking' corpus.

1 Introduction

People tracking from video is one of the key enabling technologies for recent applications in the field of Domotics, Ambient Intelligence, Surveillance, Traffic Analysis, Control and Automation and Human-Computer Interaction. While the demand for perceptual components able to provide detailed reports on human activity within the environment arises directly from the task that such applications aim to perform, they could in theory work on many different modalities such as audio, video, infra-red, active source triangulation, etc. The huge amount of information carried by images about the scene and the fact that no special devices need to be weared have made video-based solutions an appealing framework.

While the visual intelligence of animals can handle by far more complicated tasks, designing an algorithm that works in general circumstances still remains surprisingly difficult. After more than two decades of research [16] we still lack robust solutions that can handle many of the scientific challenges present in natural scenes such as occlusions, illumination changes and high-dimensional motion. Traditional approaches [9, 5, 6, 8] can work well under certain assumptions or when system parameters are tuned to the characteristics of the scene under analysis. Recently, generative approaches [11, 18] have gained increased

attention by the community because they allow to overcome some of the limitations of traditional approaches by explicitly modeling the observation process implemented in a camera. The basic idea is to generate a set of likely hypotheses and to compare the actual observation with a synthetic view of them *rendered* using explicit models of the different components of the scene such as the targets as well as the background. While powerful in principle because important features such as articulation constraints [7] or occlusions [14] may be included, their success rests on reliable yet efficient design of these models which in practice is often difficult. However, the potentials of such approaches have been demonstrated e.g. in a recent international evaluation campaign, CLEAR'06 [1], where the best performing 3D single person tracker has been a color-based particle filter [3]. This paper builds upon that work, extending the appearance model proposed there to a 3D model able to capture viewpoint-specific variations and proposing a new, fully automatic acquisition procedure.

1.1 Overview

The scenarios addressed in the CLEAR evaluation exhibit a number of peculiarities which turn the visual tracking problem into a scientific challenge. Among the most important ones we can cite:

- Variable number of targets. Actors are free to enter the monitored room and leave it during the recordings: determining and updating the number of targets to be tracked is part of the task.
- Dynamic background. Moving objects, door opening/closing, dynamic projections (displays, slide projectors, graphical interfaces, etc.) continuously change the appearance of scene constituents which are not of direct interest to the analysis. Such changes may be 'kept unknown' to the system, leading to a background scene which appears to be dynamic.
- Static targets. Comprehensive analysis requires that all tracked participants be continuously monitored, even if they that are not actively contributing to the current interaction. Such passive targets are difficult to handle by traditional approaches based on background suppression.
- Occlusions. As interactions arise among several targets who can move freely in a limited space, visual features of interest may disappear in the images according to the current constellation of the targets and the viewing direction considered. This causes loss of information in the data: analysis must consider all targets jointly to interpret data consistently.
- Changing lighting conditions. Room illumination may change significantly across the scene, and also over time, due to slide projections and non-uniform illumination.

All these issues are of major concern when using classical approaches based on the typical processing chain: background suppression - morphological noise filtering - blob classification. Robust solutions usually require the engineering

of complex cascades of low-level filters whose behaviour is difficult to understand and which require the tuning of many parameters to the particular scene conditions.

We address these problems by adopting a principled Bayesian approach which remains simple and whose performance is largely unaffected by most of the issues mentioned above. The core of the tracker is a likelihood function built upon a generative model of its visual appearance. Its purpose is to assign scores to hypothetical target poses by analysing a set of calibrated input images. Tracking is then formulated as a stochastic filtering problem, and solved by propagating the best-scored hypotheses according to a simple model of target dynamics. For the purpose of this paper, a hypothesis is specified in terms of the target’s 2D position on the floor, its horizontal body orientation, and a binary label for specifying a sitting or standing pose.

The paper is organized as follows. The next section presents a multi-view appearance model for tracking applications; a model acquisition procedure is presented in Sec. 4. Sec. 5 reports experiments aimed to validate the model and its acquisition procedure, and to demonstrate its suitability for tracking, while Sec. 6 presents the conclusions and discusses possible improvements.

2 Modeling visual appearance

Modeling human appearance in unconstrained situations is challenging and a research topic by itself [12, 17]. There is a trade-off between the level of detail one is interested in, the robustness to be achieved against non-modeled effects, and the computational load that can be afforded by the application. For the purpose of location tracking it is sufficient to resort to low-dimensional descriptions, which then marry well with the real-time demand common in many applications. To be usable with local search methods such as mean-shift [5] or particle filtering [2], the likelihood function built on top of it should render comparable hypotheses to comparable scores, thus be a smooth function of target position. At the same time, however, it should be able to absorb background clutter while remaining sufficiently discriminative to distinguish between different targets so as to maintain target identity.

Our approach addresses all these issues and is characterized by the use of two components:

- a coarse, volumetric, description of the shape of an upright standing person;
- body part- and viewpoint-based representation of target color, in form of head, torso and legs histograms.

The shape model is responsible for consistently mapping a hypothetical position into a triple of image patches where target head, torso and legs are expected to appear under that hypothesis. Using histograms to describe the appearance within these patches guarantees invariance to small spatial offsets and non-modeled articulated motion, and robustness to slight illumination changes. Part-based definition of the model highlights appearance independence that usually

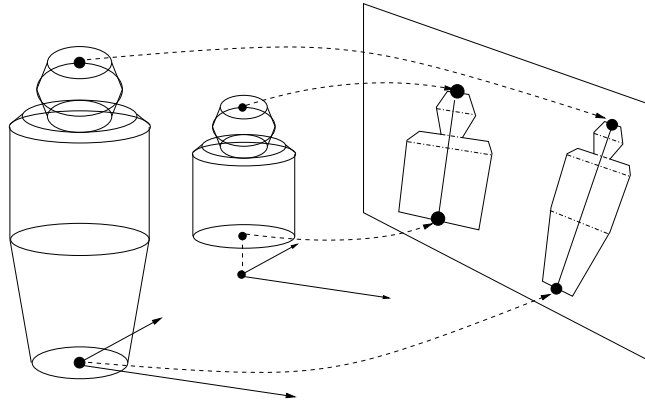


Fig. 1. 3D shape model of a standing and sitting person, and the approximate, but efficient, rendering implementation which maintains imaging artefacts such as perspective distortion and scaling.

exists between the different body parts due to skin color and clothing. This allows to discriminate between background clutter and the actual target, and between targets wearing different clothes or of different size.

2.1 Shape

A coarse, part-based 3D model identifying the scene volume covered by a person standing upright is adopted for shape, similar to the generalized-cylinder model proposed in [11]. This model is shown in Fig. 1 and is assembled from a set of horizontally elongated cone trunks. It can be calibrated to a specific target by adapting two parameters: target height and width. To obtain the image projection of this 3D model when placed in a specific 3D position \mathbf{x} of the scene we proceed as follow. Firstly, we compute a pair of 3D points which represent the center of feet and top of head of the model. In case of \mathbf{x} describing the 2D position with respect to the floor plane and h being the height of the target, these two points are simply given by \mathbf{x} enhanced by a third coordinate which has value 0 and h , respectively. These two points are then projected onto the camera reference frame by means of a calibrated camera model. The segment joining these two image points defines the axis around which the contour is drawn with piece-wise linear offset. This offset is further multiplied by a constant scale factor $s(o) \in [0.7 : 1.0]$ that accounts for profile width changing with the relative orientation o of the target to the camera. Precisely, $s(o) = 0.7 + 0.3|\cos o|$. For a sitting pose we set the height of the 3D hip centre to the typical height of a chair and ignore the leg trunks. This rendering procedure is fast and sufficiently accurate on near-horizontal views such as the ones captured by cameras placed at the corners of a room.

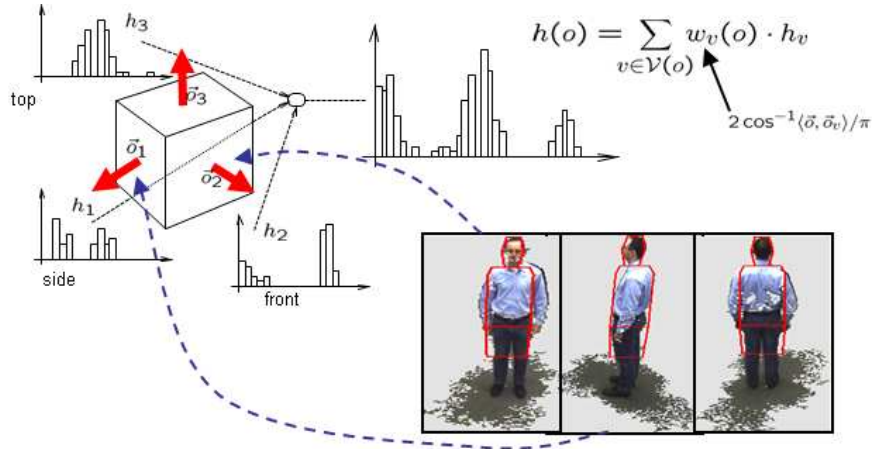


Fig. 2. Histogram synthesis. The cube faces represent pre-acquired top, front and side view of the object as seen from a new viewpoint. A weight is assigned to each view according to its amount of visible area (which is proportional to the cosine angular offset to the new views). New object appearance is then generated as a weighted interpolation of these histograms.

2.2 Color

The projected silhouette is further decomposed into three body parts: head, torso and legs. Within these parts, the appearance of the target is described by one color histogram per part. In our implementation we quantize the RGB color space uniformly in $8 \times 8 \times 8$ bins (thus we have histograms of size 512). We follow a view-based rendering approach to synthesize the appearance of a given hypothesis, based on the idea of embedding the high-dimensional manifold of target appearance in a small, linear histogram subspace. The basic idea is to record a set of key views of the target, to extract the corresponding descriptions (i.e. color histograms) for each body part, and to generate the descriptions for a new view by interpolation from these key views. How to acquire such key views for a target will be discussed later. The histogram rendering procedure for a specific pose works for each body part as depicted in Fig. 2. When generating the appearance for a new hypothesis, the object orientation o with respect to the camera is taken into account. The set of neighboring model views $\mathcal{V}(o)$ is identified and contributes to the new histogram $h(o)$ according to

$$h(o) = \sum_{v \in \mathcal{V}(o)} w_v(o) \cdot h_v. \tag{1}$$

Interpolation weights $w_v(o)$ sum up to 1 over the set $\mathcal{V}(o)$ and account linearly for the angular offset from the current viewing direction

$$w(o) \propto 2/\pi \cdot \cos^{-1}\langle \mathbf{o}, \mathbf{o}_v \rangle. \tag{2}$$

Here $\langle \mathbf{o}, \mathbf{o}_v \rangle$ denotes the scalar product between the 3D vectors pointing in the direction of pose orientation \mathbf{o} and key view orientation \mathbf{o}_v . This method supports histogram rendering from any viewing orientation, including top-down views. In setups where only lateral views are available (typically cameras placed in the corners of a room) it is sufficient to acquire key view histograms for side views only (like those seen in Figure 3), and interpolation for a new view can be achieved on the two closest ones. We have used this faster rendering strategy in the implementation of the tracker.

3 Tracking multiple targets with known appearance

An occlusion robust particle filter [14] is implemented to jointly track the locations of a number of targets. The basic idea behind particle filtering is to propagate a sample-based representation of the posterior probability density over all possible target locations. Such posterior is obtained recursively by propagating the previous estimate to current time and updating it by integrating the evidence contained in the current image.

3.1 Likelihood

The likelihood function used in the update step is built upon the presented appearance model and computes as follows. To score a given pose hypothesis \mathbf{x} on a new input image, we design a likelihood function based upon the proposed shape and colour model. To do so, hypothetical body parts are identified within the image by means of the shape model rendered for \mathbf{x} . Candidate colour histograms are then extracted from these areas. To assign a score to \mathbf{x} , these histograms are compared with the interpolated histograms of the model rendered for that pose, using a similarity measure derived from Bhattacharyya-coefficient based distance. If a^h, a^t, a^l are the area of body part projections and h_z^h, h_z^t, h_z^l and h_m^h, h_m^t, h_m^l denote normalised extracted and modelled histograms respectively, the assigned likelihood is

$$\exp \left\{ - (a^h d^2(h_z^h, h_m^h) + a^t d^2(h_z^t, h_m^t) + a^l d^2(h_z^l, h_m^l)) / 2\sigma^2(a^h + a^t + a^l) \right\} \quad (3)$$

with histogram distance d given by (index i scans all colour bins)

$$d^2(h, k) = 1 - \sum_i \sqrt{h_i k_i}. \quad (4)$$

Parameter σ can be used to control the selectivity of this function and is set empirically to 0.12.

3.2 Dynamics

For the purpose of prediction, the pose vector \mathbf{x} is enhanced with the 2D velocity of the target on the floor plane. A pose hypothesis, or particle, is thus embodied by an 5-dimensional vector, plus one binary dimension for sitting/standing.

The particle set representing the probabilistic estimate at time t is projected to time $t + 1$ by a first order autoregressive process: each particle’s location is linearly propagated along its velocity according to the time elapsed, and zero-mean Gaussian noise is added to all components to account for nonlinear behaviour. After prediction, particle likelihoods are computed on the different views available and particle weights are assigned as the product of their likelihoods over the different views. Weighted re-sampling is applied in a straightforward manner.

4 Target detection, model acquisition, and target release

The performance of an appearance-based tracker depends strongly on the quality of the model it searches for in the video. In constrained scenarios, such as the monitoring of meetings or lectures in a dedicated room, it may be possible to acquire accurate models for each participant a priori in an enrollment phase, through the target visiting a *hot spot* area. Stable and robust tracking can then be achieved in spite of occlusions and cluttered environments, as demonstrated live during the IST 2006 event [?].

4.1 Target detection

To cope with unconstrained scenarios recorded in the CLEAR07 corpus, we built upon the detection procedure proposed in [3] to account for sitting and standing pose, occlusions, and limited visual coverage of each camera. Furthermore, the procedure has been integrated with the tracking. Again, a Bayesian approach is adopted. The tracking domain is collapsed into a regular 2D grid, with each grid corner representing the possible location of a new target. At each position, a number of hypotheses are created for different target heights and widths, body orientations, and pose. Given a set of input images capturing the current room configuration, a likelihood is assigned to each such hypothesis, calculated as the product of responses computed independently on each single view. Relevant features in each view derive from:

- coverage of foreground blobs extracted using an adaptive background model,
- match of projected model contour and observed edges, and
- horizontal color symmetry of the upper body.

If the largest likelihood response on the detection grid is above a given threshold, a new target is detected. Body part histograms are then extracted from all unoccluded views by rendering the shape model for that configuration, and stored as the key views the target’s appearance model (annotated with the relative viewing angles derived from body orientation), together with the target’s physical width and height. Appearance model, shape parameters and current position are transmitted to the tracker which allocates a new track to it.

To speed-up the acquisition process, a multi-resolution approach is implemented. At the coarse level, a subset of configurations (undersampled grid, one height/width level only) are tested on all available views (including the ceiling

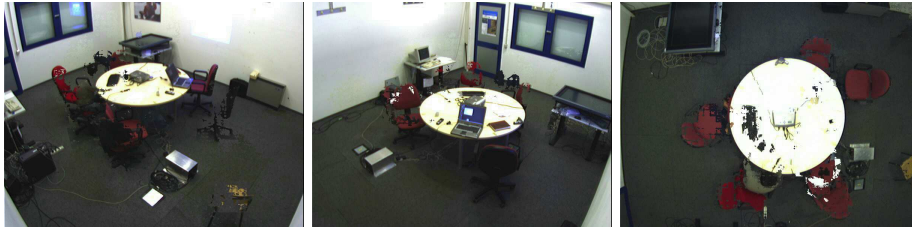


Fig. 3. Updated background models at the end of sequence ITC_20060922B.B. In spite of moved objects (chairs and laptop) and persistent foreground blobs (static targets) the models still describe the background scene at an acceptable accuracy.

camera), considering only foreground blob coverage. The neighborhood of hypotheses with significant likelihood are analyzed at the finer level, and tested at full resolution by considering all three visual features. To avoid re-acquisition of already tracked targets, the neighborhood of their currently estimated location is removed from the set of detection hypotheses before each iteration.

4.2 Detection likelihood

It remains to show how the three components of single-view responses are computed. A background model in Y-Cb-Cr color space is maintained for each camera [4]. It is bootstrapped from a set of empty room images, and updated at pixel-level by temporal filtering: $B_t[i, j] = \alpha B_{t-1}[i, j] + (1 - \alpha) I_t[i, j]$. Parameter α may vary with the framerate, but is kept constant for the evaluation. An image mask is computed at each iteration to update the model on stable background image regions only. It accounts for persistent motion (accumulated frame-by-frame difference + morphological filtering) and image regions supported by the current tracks (identified by rendering the shape model for the current estimates), which are then neglected in the background update (see Fig.3). The coverage likelihood term for a given hypothesis is computed as the percentage of projected silhouette area currently covered by foreground blobs (simple Y-Cb-Cr thresholding). The contour likelihood is computed as follows. Canny edge detection is performed on the input image to extract candidate edges. Such edges are validated through connected component analysis and morphological noise cleaning. To each detection hypothesis a score is assigned according to the amount of overlap between measured and expected edges. To compute this term, the shape model is rendered and its contour matched with the validated edge map. The log-value of the resulting score is composed of two terms. The first one is computed as the sum over model contour pixels u of their Euclidean distance to the nearest validated edge, normalized w.r.t. the length of the contour under consideration. The second term penalizes edges whose orientation does not match with the contour normal $N(u)$. To limit the influence of missing edges, the contribution of contour pixels that exceed 30% of projected body width are set to a maximum value. Fast scoring can be achieved by limiting edge analysis to

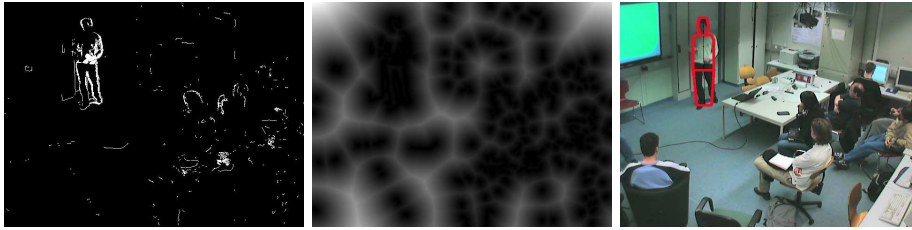


Fig. 4. Validated motion edges, their distance transform, and the projected 3D shape exhibiting the highest contour likelihood. Note that detection is unaffected by the presence of other people in the room.

image regions-of-interest (ROI) which can be determined by pre-rendering of the detection hypotheses. Within such ROIs, Euclidean distances are pre-compiled into lookup tables obtained by means of the Distance Transform (DT) of edge images (see Fig. 4). If $\Delta(\mathbf{x})$ denotes projected body width, $\mathcal{C}(\mathbf{x})$ describes its contour, $D(u)$ is the DT of the validated edge map and $\mathbf{G}(u)$ is the gradient vector of the closest edge in u , the likelihood assigned is

$$\exp\left(-\frac{1}{\text{length}\{\mathcal{C}(\mathbf{x})\}} \int_{\mathcal{C}(\mathbf{x})} \min\left\{1, \frac{D(u)}{0.3\Delta(\mathbf{x})}\right\} - 0.5\langle \mathbf{N}(u), \mathbf{G}(u) \rangle du\right). \quad (5)$$

Similar shape likelihoods have been proposed for tracking in [10] and for detection in [19]. In [10] edges are searched for only along a small predefined set of contour normals. While this formulation is computationally cheaper when the number of hypotheses to be tested is limited, the overhead introduced through the computation of DT in our approach becomes negligible for a large search task such as high-resolution target detection. In addition, our measure is more accurate as it considers continuous contours. The third likelihood term, color symmetry of the upper body, is computed as the L_1 distance between left and right part RGB histograms extracted from the current image.

4.3 Target release

To decide whether to propagate an existing track or not, views of the ceiling camera are analyzed. Foreground blobs are extracted by considering significant differences to the background model (Y-Cb-Cr thresholding + morphological filtering). If the neighborhood of a track remains unsupported for a number of subsequent iterations (1 second) the track is terminated.

5 Evaluation results

The complete system implements a one-pass algorithm with no pre- and post-processing. For the tracking of an acquired target only the four corner cameras

are used. The top-down view of the scene is used solely to trigger target detection and model acquisition, and to terminate unsupported tracks. The system output is computed as the expectation over the current, weighted particle set. No manual intervention is needed, no parameters have been changed to process sequences from different sites.

The sequences have been processed at a real time factor of 1.5 on a Intel Xeon 2.3 Ghz DualCore biprocessor, using up to 80% of the processing capacity. Most of the time, with 4-5 tracked targets, it works at about 60-70% of CPU load, including reading and uncompressing images. The tracker works at a variable frame rate, according to the uncertainty present in the data. It implements a mechanism to adapt the number of particles needed to consistently represent the probabilistic estimates [15]. It is therefore not possible to express the efficiency in cputime/frame. The slowest part is detection, which requires the system to run slower than real time. Again, the detection processing time cannot be specified by a constant cputime/frame factor, as it depends linearly on the size of foreground blobs detected in the ceiling camera which are not covered by any of the active tracks. Thus, detection time varies with complex factors such as the effectiveness of background model adaptation and the tracking performance of already acquired targets. It ranges from a few milliseconds (when all targets are tracked) to up to several seconds (when background model adaptation failed and no tracks are active). Typically, a new target is detected in less than half a second. The tracking is easily accommodated in real time.

Table 1 reports the performance of the proposed approach on the 'CLEAR'07 3D Person Tracking' corpus (see [13] for a presentation of the error metrics). Some images displaying tracker output are shown in Fig. ???. We throw the following conclusions.

- **Significant challenges are posed to detection algorithms.** The corpus contains sequences with very different dynamics, ranging from empty-room segments (UPC_) to sequences which start with all people sat down (IBM_), to situations in which 5 people enter the scene almost simultaneously (UKA_). Target detection in such situations is challenging due to untracked occlusions and limited visual coverage. The acquisition of poor appearance models then directly impacts on the overall tracking performance.
- **Occlusions are not a major concern for the tracking.** The frequency of visual occlusions in the corpus is low. Most of the time people are sat down, thus corner cameras, as well as the ceiling camera, observe no or only small occlusions. There is no significant gain in pursuing occlusion reasoning at the cost of more intense computations, as implemented in our system [14].
- **Sensing geometry varies significantly with the different sites.** Fig. ??? shows images from two sites. For AIT setup the ceiling camera delivers highly informative and undistorted views of the scene; corner cameras have very narrow field-of-views, where targets are visible only partially, many times even only in one camera. This makes the target model acquisition very difficult. On the other hand, IBM room has good coverage from the corner

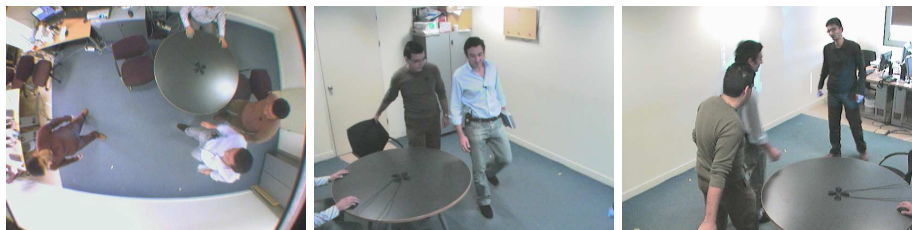


Fig. 5. Visual coverage in AIT sequences: the ceiling camera provides rich information; corner cameras have narrow FoVs with limited overlap.

cameras, but the ceiling view exhibits significant distortion at the peripheral regions. Also, the cameras’ color responses are very different.

In summary, the performance of the presented system is bounded by the limited detection efficiency. The system builds upon a robust multi target tracker, recently extended to handle online detection and model acquisition. As a consequence of this, there is space for significant improvements, especially in target model adaptation and non-instantaneous acquisition. Fully integrated use of the ceiling camera is also expected to provide measurable improvements, and may require a more flexible management of target models.

	MOTP	MISSRATE	FALSEPOS	MISSMATCH	MOTA
overall	141 mm	20.7 %	18.6 %	518 (1.1 %)	59.6 %
best case	107 mm	3.2 %	6.2 %	12 (0.1 %)	89.8 %
worse case	212 mm	31.2 %	76.6 %	12 (0.1 %)	-13.7 %
avg. AIT	209 mm	20.1 %	26.6 %	128 (1.9 %)	51.4 %
avg. IBM	126 mm	11.9 %	10.8 %	92 (0.8 %)	76.4 %
avg. ITC	123 mm	23.4 %	15.3 %	55 (0.6 %)	60.7 %
avg. UKA	124 mm	25.4 %	19.8 %	203 (1.8 %)	53.0 %
avg. UPC	145 mm	24.5 %	22.7 %	133 (1.2 %)	51.7 %

Table 1. Tracking scores on the CLEAR07 corpus VPT task: overall, best case is IBM_20060815_B, worse case is AIT_20061020D_B

6 Conclusion and future work

In this paper we described a multi-view appearance model for tracking humans. It is derived according to imaging principles using a 3-dimensional part-based shape model of a standing person. Within each part the color appearance is generated by weighted histogram interpolation to account for relative orientation, considering the set of closest key views. To acquire such key views, a target is

detected within a hot spot region by exhaustive search driven by a model-based contour likelihood. The resulting appearance model is discriminative and fast, supporting real-time tracking of several people while maintaining their identity. In future work we want to face the problem of online model acquisition and adaptation, taking this work as a starting point. This is particularly challenging in the multi-target context due to occlusion and target similarity.

7 Acknowledgments

Research partly funded by the European Union under project IP 506909: CHIL - Computers in the Human Interaction Loop. Acknowledgements go to Paul Chippendale for help on background subtraction, and Francesco Tobia for data preparation and providing software tools.

References

1. CLEAR 2006 evaluation campaign. April 5-6 2006. [Online]: <http://www.clear-evaluation.org/>.
2. S. Arulampalam, A. Maskell, N. Gordon, and T. Clapp. A tutorial on particle filters for on-line non-linear/non-gaussian bayesian tracking. *IEEE Trans. Signal Processing*, 50(2), 2002.
3. R. Brunelli, A. Brutti, P. Chippendale, O. Lanz, M. Omologo, P. Svaizer, and F. Tobia. A generative approach to audio-visual person tracking. In *CLEAR'06 Evaluation Campaign Workshop*, 2006.
4. P. Chippendale. Towards automatic body language annotation. In *IEEE International Conference on Automatic Face and Gesture Recognition (FG06)*, 2006.
5. D. Comaniciu, V. Ramesh, and P. Meer. Real-time tracking of non-rigid objects using mean-shift. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2000.
6. R. Cucchiara, C. Grana, G. Tardini, and R. Vezzani. Probabilistic people tracking for occlusion handling. In *Int. Conf. Pattern Recognition*, 2004.
7. J. Deutscher, A. Blake, and I. Reid. Articulated body motion capture by annealed particle filtering. In *Int. Conf. Computer Vision and Pattern Recognition*, 2000.
8. S. Dockstader and A.M. Tekalp. Multiple camera tracking of interacting and occluded human motion. *Proc. of the IEEE*, 89(10), 2001.
9. I. Haritaoglu, D. Harwood, and L.S. Davis. W4: Real-time surveillance of people and their activities. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 22(8), 2000.
10. M. Isard and A. Blake. Condensation – conditional density propagation for visual tracking. *Int. Journal of Computer Vision*, 29, 1998.
11. M. Isard and J. MacCormick. BraMBLe: A bayesian multiple-blob tracker. In *Int. Conf. Computer Vision*, 2003.
12. A.D. Jepson, D.J. Fleet, and T. El-Maraghi. Robust online appearance models for visual tracking. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 25(10), 2003.
13. A. Elbs K. Bernardin and R. Stiefelhagen. Multiple object tracking performance metrics and evaluation in a smart room environment. In *IEEE International Workshop on Visual Surveillance (VS2006)*, 2006.

14. O. Lanz. Approximate bayesian multibody tracking. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 28(9), 2006.
15. O. Lanz. An information theoretic rule for sample size adaptation in particle filtering. In *International Conference on Image Analysis and Processing (ICIAP07)*, 2007.
16. T.B. Moeslund and E. Granum. A survey of computer vision-based human motion capture. *Computer Vision and Image Understanding*, 81(3), 2001.
17. H. Sidenbladh and M.J. Black. Learning the statistics of people in images and video. *Int. Journal of Computer Vision*, 54(1), 2003.
18. T. Zhao and R. Nevatia. Tracking multiple humans in complex situations. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 26(9), 2004.
19. T. Zhao and R. Nevatia. Tracking multiple humans in crowded environment. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2004.