Detection-Assisted Initialization, Adaptation and Fusion of Body Region Trackers for Robust Multiperson Tracking

Keni Bernardin, Alexander Elbs, Rainer Stiefelhagen
Interactive Systems Lab,
Institut für Theoretische Informatik
Universität Karlsruhe, 76131 Karlsruhe, Germany
keni@ira.uka.de, alex@segv.de, stiefel@ira.uka.de

Abstract

In this paper, we present a system for simultaneous tracking of multiple persons in a smartroom using multiple cameras. Robust person tracks are created, continuously adapted, and deleted by fusing cues from foreground segmentation maps and various appearance-based object detectors. Tracking is performed using color histograms which are automatically filtered and adaptated based on local image characteristics. Tracks from the various 2D views are merged to 3D position estimates by an intelligent fusion algorithm based on triangulation error reduction. The approach allows to robustly track moving, standing or sitting persons in cluttered environments and to successfully recover lost tracks at any point in the room. We also introduce a new set of metrics to measure multiple object tracking performance. Our system reaches a high tracking accuracy with average position errors of less than 17cm.

1. Introduction and Related Work

One of the major problems faced by indoor tracking systems is the lack of reliable features that allow to keep track of persons in natural, evolving and unconstrained scenarios. The most popular features in use are color features and foreground segmentation or movement features [2, 3, 4, 7, 8, 15], each with their advantages and drawbacks. Doing e.g. blob tracking on background subtraction maps is error-prone, as it requires a clean background and assumes only persons are moving. In real environments, the foreground blobs are often fragmented or merged with others, they depict only parts of occluded persons or are produced by shadows or displaced objects. When using color information to track people, the problem is to create appropriate color histograms or models. Generic color models are usually sensitive and environment-specific [5]. If no generic model is used, one must at some point decide which pixels in the image belong to a person to initialize a dedicated color histogram [4]. Moreover, the color model needs to be adapted regularly as the person's color varies with environmental conditions [8]. Assuming a manual segmentation of the body region is available, very robust color tracks can be achieved. But this is not practical for online systems, which is why many approaches rely on a semi-automatic initialization in special creation areas, or on the contours found by foreground segmenters [13, 15]. This, however, requires the cooperation of the users and/or a clean and relatively static background.

Here, we propose to use appearance based detectors for body regions to give initial hints for a person's location in the image. In combination with foreground segmentation maps, precise color histograms can be created and maintained, which allow for robust tracking.
Another aspect of the problem, when adressing dynamic scenarios with multiple freely moving people, is to estimate the number of people, to initiate the right number of tracks and maintain them through occlusion, and to avoid switching them when they overlap. Many approaches tackle this problem for single-view setups [1, 2, 7, 13]. Haritaoglu et al. [7] use temporal templates to recover tracks after an overlap. Tao et al. [13] use a particle filter approach to estimate optimal configurations of people to match the observation. Few approaches actually use multiple synchronized cameras to observe real 3D positions [4]. The here presented system relies on wide baseline views to achieve 3D tracking. It tackles the problem in two stages. It first realizes a robust tracking of entities in 2D views, and then uses the information about relative camera positions to derive 3D estimates, to solve issues relating to occlusion and overlap, and to cross-validate 2D tracks in the different camera views. It makes no assumptions about the environment, e.g. no special creation or deletion zones, about the consistency of a person's appearance, and constantly verifies the validity of its tracks, thus recovering from tracking errors automatically.

2. Multi-View Person Tracking System

The developed system is a 3D point tracker designed to track multiple persons using a variable number of fixed cameras installed at the room corners. It tracks up to three different body regions for each person on each camera image, using color histogram models. Hints from several types

of object detectors trained to recognize different body parts, as well as features gained from foreground segmentation maps are used to select appropriate pixels in the images and initialize robust color histograms for tracking. The information from several cameras is then fused to produce 3D hypotheses of the persons' positions. The fusion algorithm uses the multiple views to robustly handle occlusion or overlap, to recover lost tracks or to recognize and delete erroneous tracks. In the following, a detailed explanation of the system's components is given.

2.1 Classifier Cascades and Foreground Segmentation

As discussed above, an important step for correct initialization of person color models is recognizing when and where a person appears in the image. To achieve this, a set of appearance-based object detectors is used to scan image regions of interest. These detectors are classifier cascades that build on haar-like features, as decribed in [9, 14]. For our implementation, the cascades were taken from the OpenCV [16] library. They are trained on a wide variety of example images and are not specific to the conditions of our smartroom. Experiments were conducted with three types of cascades: One to recognize whole silhouettes of standing persons (*full body*), one to recognize the upper body region of standing or sitting persons (*upper body*), and one to recognize frontal faces (*faceA*). Additionally, a second face cascade (*faceB*), trained specifically on images from our smartroom was used for comparison. Using these detectors, the image is scanned at different scales and bounding rectangles are obtained for regions likely to contain a body part (see Fig. 1(a)). By using such detectors, we avoid the drawbacks of manually defined creation/deletion zones and are able to initialize or recover a track at any place in the room.

The classifier cascades can produce false detections, and they do not deliver the exact bounding contours of persons.This is why an additional preprocessing step is taken: The image is segmented into foreground regions by using an adaptive background model, and only detection hits on foreground regions are considered valid. This combined approach offers two advantages: The cascades, on one hand, increase robustness to segmentation errors, as foreground regions not belonging to persons, such as moved chairs, doors, shadows, etc, are ignored. The foreground segments, on the other hand, help to decide which of the pixels inside a detected rectangle belong to a person, and which to the background. Knowing exactly which pixels belong to the detected person is useful to create accurate color histograms and improve color tracking performance.

2.2 Color Histogram Tracking

Whenever an object detector has found an upper body, full body or face in the image, a color histogram of the respective person region is constructed from the foreground pixels belonging to that region, and a track is initialized. The actual tracking is done in HSV color space by applying the meanshift algorithm [6] on histogram backprojection images.

As the person silhouette may well include colors also present in the surrounding background, the constructed histogram model H needs to be filtered to increase its discriminative properties. After normalization, we have $H(x) = P(x|Person)$. By applying Bayes' rule, we can obtain the likelihood ratio $\frac{P(Person|x)}{P(\neg Person|x)} \sim \frac{P(x|Person)}{P(x|\neg Person)} = \frac{H(x)}{H_N(x)}$, with H_N a histogram modeling non-person colors. This shows that to achieve accurate tracking, we need to divide H by an appropriately chosen background histogram.

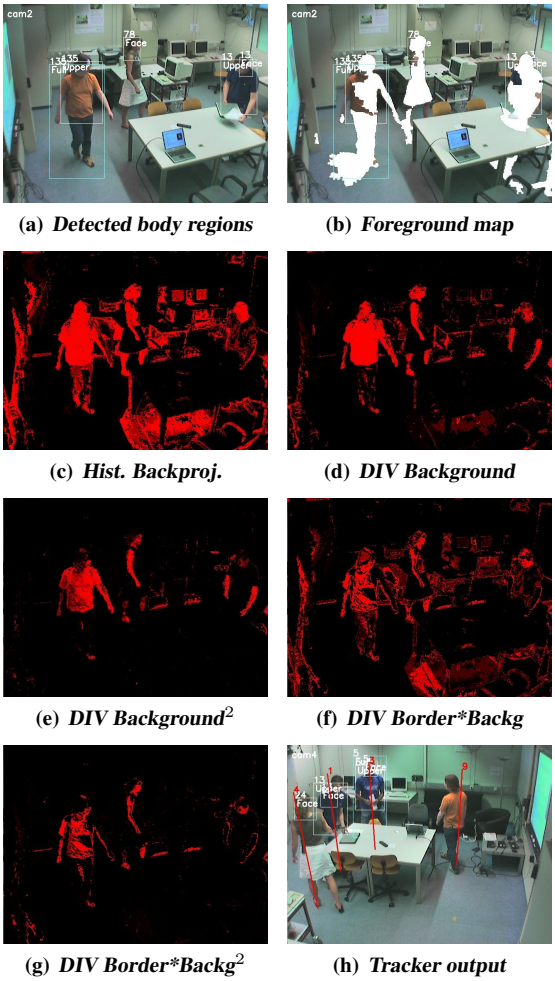
For the dividing histogram H_N, several variants can be used: the overall background histogram (as in [10]), the histogram of the region immediately surrounding the person (as in e.g. [11]), and other similar variants (see Fig. 1). The best choice strongly depends on the local image conditions at the time of initialization and is made at runtime by our system: For each resulting filtered histogram, the corresponding backprojection image is quickly scanned by a few steps of the meanshift algorithm and the choice which best keeps the track centered around the initially detected body region is adopted.

To further ensure consistent tracking, the color histogram for a track is adapted every time a classifier cascade produces a valid detection hit on that track. This also serves as confirmation that the track still represents a person. The adapted histogram is obtained by a weighted combination of the old and new values: $H_a = (1 - \alpha)H_{old} + \alpha H_{new}$. H_a is however only used if it represents an improvement over H_{old}. This is done by judging the discriminative properties of the resulting backprojection maps. Only if the likelihood ratio of pixels inside the detection rectangle to pixels inside the region immediately surrounding it is greater for H_a than for H_{old}, the adaptation is made.

Tracks that are not confirmed by a detection hit for a certain period of time are deleted, as they are most likely erroneous.

2.3 Combining Multiple 2D Body Region Tracks

For each upper/full body or face track, the actual body center of the corresponding person in the image is estimated. If only one type of track is availabe, this point serves as 2D hypothesis for the person position. If however many types of tracks (e.g. an upper body and a face track) are available for the same person, a more robust, combined hypothesis is produced. We recognize that different tracks belong to the same person by thresholding the distance between the corresponding body centers. If it is small enough, the tracks are associated. On the other hand, if it becomes too large, it is a sign that an error occured (either one of the tracks is erroneous, or tracks belonging to two different persons were wrongfully matched), and the tracks are deleted. New detector hits must then reinitialize the tracking of those body parts. Tracking up to three different body regions for



(a) *Detected body regions* (b) *Foreground map*

(c) *Hist. Backproj.* (d) *DIV Background*

(e) *DIV Background*^{2} (f) *DIV Border*Backg*

(g) *DIV Border*Backg*^{2} (h) *Tracker output*

Figure 1. Color histogram creation, filtering and tracking. *a)* **Face, upper and full body detections (***rectangles***) in one camera view.** *b)* **Foreground segmentation (in** *white***). Only foreground pixels inside the rectangles are used.** *c)* **Histogram backprojection for the upper body track of the leftmost person.** *d), e), f) and g)* **Effects of different types of histogram division.** *Background***: Overall background histogram.** *Border***: Histogram of the background region immediately surrounding the detected rectangle.** *h)* **Tracker output as seen from another view**

a person increases robustness against occlusions, lighting variations, and other tracking errors.

2.4 Fusion and Generation of 3D Hypotheses

The 2D hypotheses produced for every camera view are triangulated by an intelligent 3D tracking algorithm, to produce 3D position estimates. For this, the triangulation error, as described in [12], is used.

With a variant of the RANSAC algorithm, likely correspondences between 2D tracks are first established. When the triangulation error between a set of 2D hypotheses is small enough, they are associated to form a 3D track. Likewise, when it exceeds a certain threshold, the 2D hypothesis which contributes most to the error is dissociated again, and the 3D track is maintained using the remaining hypotheses.

Once a 3D estimate for a person's position has been computed, it is further used to continuously validate 2D tracks and to initiate color histogram tracking in camera views where the person has not yet been detected. It is also used to robustly handle occlusions: When a certain amount of overlap is detected between two 2D tracks in a camera image, the 3D position serves to determine which of the tracks is furthest from the camera. This track is then deactivated, its position is predicted for the duration of the occlusion, and it is reactivated again when there is no more overlap.

The developed tracker draws its strength from the intelligent fusion of several camera views and body region tracks. It initializes its tracks automatically, constantly adapts its color models and verifies the validity of its tracks using special object detectors. It is capable of tracking several people, regardless if they are sitting, walking or standing still, in a cluttered environment with uneven lighting conditions. Fig. 1(h) shows a sample tracker output.

3 Evaluation of Tracker Performance

Defining good measures to express the characteristics of a system for continuous tracking of multiple objects is not a straightforward task. Various measures exist and there is no consensus in the literature on the best set to use. Here, we propose a small expressive set of metrics and show a systematic procedure for their calculation.

Assuming that for every time frame t a multiple object tracker outputs a set of hypotheses $\{h_1 \dots h_m\}$ for a set of visible objects $\{o_1 \dots o_n\}$, we define the procedure to evaluate its performance as follows:

Let the correspondence between an object o_i and a hypothesis h_j be valid only if their distance $dist_{i,j}$ does not exceed a certain threshold T, and let $M_t = \{(o_i, h_j)\}$ be a dynamic mapping of object-hypothesis pairs.

Let $M_0 = \{\}$. For every time frame t,

1. For every mapping (o_i, h_j) in M_{t-1}, verify if it is still valid. If o_i and h_j still exist and if $dist_{i,j} < T$, make the correspondence between o_i and h_j for frame t.

2. For all objects for which no correspondence was made yet, try to find a matching hypothesis. Allow only one to one matches. Start by matching the pair with the minimal distance and then go on until the threshold T is exceeded or there are no more pairs to match. If a correspondence (o_i, h_k) is made that contradicts a mapping (o_i, h_j) in M_{t-1}, replace (o_i, h_j) with (o_i, h_k) in M_t. Count this as a mismatch error and let mme_t be the number of mismatch errors for frame t.

3. After the first two steps, a set of matching pairs for the current time frame is known. Let c_t be the number of matches found for time t. For each of theses matches, calculate the distance d_t^i between the object o_i and its corresponding hypothesis.

4. All remaining hypotheses are considered false positives. Similarly, all remaining objects are considered misses. Let fp_t and m_t be the number of false positives and misses respectively for frame t. Let also g_t be the number of objects present at time t.

5. Repeat the procedure from step 1 for the next time frame. Note that since for the initial frame, the set of mappings M_0 is empty, all correspondences made are initial and no mismatch errors occur.

Based on the matching strategy described above, two very intuitive metrics can be defined: The *M*ultiple *O*bject *T*racking *P*recision ($MOTP$), which shows the tracker's ability to estimate precise object positions, and the *M*ultiple *O*bject *T*racking *A*ccuracy ($MOTA$), which expresses its performance at estimating the number of objects, and at keeping consistent trajectories:

$$MOTP = \frac{\sum_{i,t} d_{i,t}}{\sum_t c_t} \tag{1}$$

$$MOTA = 1 - \frac{\sum_t (m_t + fp_t + mme_t)}{\sum_t g_t} \tag{2}$$

The $MOTA$ can be seen as composed by 3 error ratios:

$$\overline{m} = \frac{\sum_t m_t}{\sum_t g_t}, \quad \overline{fp} = \frac{\sum_t fp_t}{\sum_t g_t}, \quad \overline{mme} = \frac{\sum_t mme_t}{\sum_t g_t},$$

the ratio of misses, false positives and mismatches in the sequence, computed over the total number of objects present in all frames.

4. Experiments

The here described tracking system was evaluated on a set of recordings made in the smartroom. The recordings were made using four synchronized SONY DFW-V500 color firewire cameras at 15 fps, they depict various situations involving 3-4 users interacting in the room and have a length of approx. 5min. These sequences were manually labeled by marking the persons' head centroid in every camera view. These points were then triangulated and projected to the ground to serve as ground truth reference for person positions. Several combinations of classifier cascades for person detection were tested. In addition, an extra test run was made using a simulated cascade (*truth*): The ground truth was used to generate very accurate "fake" detector hits every 15th frame. This experiment served to measure how well the tracking algorithm performs assuming we have very reliable and regular hints for person positions. The results are depicted in table 1.

Table 1. Results for various detector types

Detector	$MOTP$	$\overline{m}$	$\overline{fp}$	$\overline{mme}$	$MOTA$
full	151mm	41.9%	5.1%	0.5%	52.5%
upper	185mm	5.6%	39.4%	1.5%	53.6%
faceA	196mm	39.4%	15.9%	1.2%	43.6%
faceB	200mm	17.1%	72.0%	4.1%	6.9%
full+upper	**168mm**	**5.6%**	**36.1%**	**2.4%**	**55.9%**
full+faceA	161mm	30.1%	33.6%	1.9%	34.5%
upper+faceA	187mm	4.6%	48.8%	2.5%	44.1%
full+up.+f.A	205mm	11.9%	56.6%	2.6%	28.9%
truth	168mm	5.7%	0.5%	5.4%	88.3%

As could be expected, the performance of the system is tightly coupled with the quality of the detectors used. Using only one type of detector often caused high error rates. Using full body detectors alone produced many misses, as sitting persons were never detected. The same happens when using *faceA*, as the detector was not trained to recognize very low resolution faces (approx. 20x20 pixels). The face detector *faceB*, specially adapted to the room, was able to find more faces, but also has a huge false positive rate, which leads to the generation of many false person tracks as well as a higher amount of mismatches. The best combination was posed by the *full+upper* body cascades producing slightly better results (56% accuracy) than when using *upper* alone. When using more than 2 types of detectors, the cumulative effect of false positives caused the system performance to degrade again. For comparison, a system using the simulated, highly accurate *truth* cascade produced almost no misses or false alarms, with an accuracy of 88%. Almost all trackers were fairly able at producing precise position estimates with average errors below 20cm.

5. Conclusions and Future Work

In this paper, we have presented an approach for the tracking of multiple persons in cluttered, natural indoor scenes using multiple wide baseline camera views. The approach relies on person detection hints supplied sporadically by various types of appearance-based detectors and on foreground segmentation maps to create, filter and adapt robust color histograms for fast tracking. The cues supplied by several 2D tracks in several camera views are intelligently merged by a 3D fusion algorithm that handles occlusions, detects and deletes spurious tracks and produces 3D position estimates. Results have shown that good tracking

results can be achieved, even when using general person detectors, not specialized or trained for use in our room. Our system was evaluated using a new set of metrics for multiple object tracking introduced here. With a combination of upper and full body trackers, a tracking accuracy of 56% and precision of 17cm could be achieved.

6 Acknowledgement

The work presented here was partly funded by the *European Union* (EU) under the integrated project CHIL, *Computers in the Human Interaction Loop* (Grant number IST-506909).

References

[1] T. Darrell, G. Gordon, M. Harville and J. Woodfill, *"Integrated Person Tracking Using Stereo, Color, and Pattern Detection"*, International Journal of Computer Vision 37(2), 2000.

[2] W. Niu, L. Jiao, D. Han, and Y. Wang, *"Real-Time Multi-Person Tracking in Video Surveillance"*, Pacific Rim Multimedia Conf., Singapore, 2003.

[3] R. Y. Khalaf and S. S. Intille, *"Improving Multiple People Tracking using Temporal Consistency"*, MIT Dept. of Architecture House_n Project Technical Report, 2001.

[4] A. Mittal and L. S. Davis, *"M2Tracker: A Multi-View Approach to Segmenting and Tracking People in a Cluttered Scene Using Region-Based Stereo"*, European Conference on Computer Vision, 2002, pp. 18-33.

[5] N. Checka, K. Wilson, V. Rangarajan, T. Darrell, *"A Probabilistic Framework for Multi-modal Multi-Person Tracking"*, Workshop on Multi-Object Tracking (CVPR), 2003.

[6] D. Comaniciu and P. Meer, *"Mean Shift: A Robust Approach Toward Feature Space Analysis"*. IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 24, No. 5, May 2002.

[7] I. Haritaoglu, D. Harwood and L. S. Davis, *"W4: Who? When? Where? What? A Real Time System for Detecting and Tracking People"*. Third Face and Gesture Recognition Conference, pp. 222–227, 1998.

[8] Y. Raja, S. J. McKenna, S. Gong, *"Tracking and Segmenting People in Varying Lighting Conditions using Colour"*. 3rd. Conf. on Face & Gesture Rec., pp. 228, 1998.

[9] R. Lienhart and J. Maydt, *"An Extended Set of Haar-like Features for Rapid Object Detection"*. IEEE ICIP, Sep. 2002.

[10] K. Nickel and R. Stiefelhagen, *"Detection and Tracking of 3D-Pointing Gestures for Human-Robot-Interaction"*. Third IEEE Int. Conference on Humanoid Robots, Karlsruhe, Germany, Oct. 2003.

[11] R. Collins, Y. Liu, and M. Leordeanu, *"On-Line Selection of Discriminative Tracking Features"*. IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 27, No. 10, October, 2005.

[12] Dirk Focken, Rainer Stiefelhagen, *"Towards Vision-Based 3-D People Tracking in a Smart Room"*, IEEE International Conference on Multimodal Interfaces, Pittsburgh, PA, USA, October 14-16, 2002, pp. 400-405.

[13] H. Tao, H. Sawhney and R. Kumar, *"A Sampling Algorithm for Tracking Multiple Objects"*. International Workshop on Vision Algorithms: Theory and Practice, pp. 53–68, 1999.

[14] P. Viola and M. Jones, *"Rapid Object Detection using a Boosted Cascade of Simple Features"*. IEEE Int. Conf. on Computer Vision and Pattern Recognition, 2001.

[15] C. Wren, A. Azarbayejani, T. Darrell, A. Pentland, *"Pfinder: Real-Time Tracking of the Human Body"*. IEEE Trans. on Pattern Analysis and Machine Intelligence, vol 19, no 7, July 1997.

[16] OpenCV - Open Computer Vision Library, http://sourceforge.net/projects/opencvlibrary/