Detecting and Tracking People and their Body Parts in Infrared

Kai Jüngling and Michael Arens

Fraunhofer IOSB
Fraunhofer Institute of Optronics, System Technologies and Image Exploitation
Gutleuthausstr. 1, 76275 Ettlingen, Germany

ABSTRACT

In most of today's surveillance tasks, people's actions are the focus of attention. A prerequisite for action interpretation is a stable tracking of people to build meaningful trajectories. Specifically in surveillance applications, not only trajectories on agent level are of interest, but also interpretation on the level of limbs provides important information when it comes to more sophisticated action recognition tasks. In this paper, we present an integrated approach to detect and track people and their body parts in thermal imagery. For that, we introduce a generic detection and tracking strategy that employs only local image features and thus works independently of underlying video data specifics like color information – making it applicable to both, visible and infrared data. In addition, we show how this approach serves to detect a person's body parts and extract trajectories which can be input for further interpretation purposes.

Keywords: Visual surveillance, person tracking, limb tracking

1. INTRODUCTION

Object detection and – more specifically – person detection and tracking have been subject to extensive research over the past decades. This results from as versatile application areas as video surveillance, thread assessment in military applications, human computer interaction, and driver assistance systems. Specifically in the area of automated surveillance, person detection and tracking is of large interest since it is a prerequisite for automated situation assessment and interpretation. Systems like[1] or[2] (in case of infrared data), that rely on background modeling techniques in order to detect people as foreground regions have two major drawbacks. First, they are not able to reliably distinguish between object classes, and second, they are not applicable from a moving camera without employing other techniques for motion compensation. Both problems can be tackled by using a dedicated person detector. Lately, a lot of work has been proposed in this direction. Most of it[3–6] focuses on the visible spectrum while some work[7,8] specifically addresses infrared. Some work has been proposed to build a tracking based on these object detectors[4,8–10] but only little of these approaches[8] are applicable for the more challenging task of tracking in infrared since most of them rely on sensor specific features like color or rich texture which is not available in infrared. An extensive review of the whole field of pedestrian detection and tracking is beyond the scope of this paper and can be found in.[11]

In this work, we seize on the task of detecting and tracking multiple pedestrians in real world environments from a monocular, possibly moving infrared camera and by that pursue our work presented in.[12,13] We show that our local feature based tracking strategy, previously applied to data in the visible spectrum,[13] is capable of the more difficult task of tracking people in infrared from a moving camera, too. In addition to just track people as a compound, we introduce a technique that allows for tracking a person's limbs on the basis of the person tracking results. In what follows, section 2 gives a short overview of the tracking approach and presents an evaluation of person tracking in three infrared image sequences. Section 3 introduces our limb tracking approach and evaluates this limb tracking for an exemplary sequence.

Further author information:
Kai Jüngling: E-mail: kai.juengling@iosb.fraunhofer.de
Michael Arens: E-mail: michael.arens@iosb.fraunhofer.de

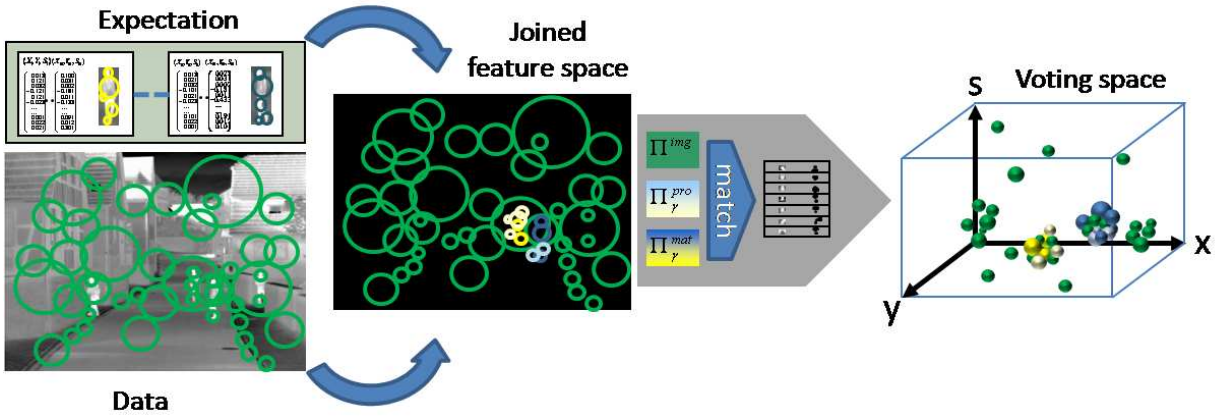

Figure 1. Coupling of expectation and data for tracking.

2. PERSON TRACKING

Our person tracking approach is built on our former work on infrared person detection presented in.[12] For tracking, we seize on our work on people tracking in the visible spectrum[13] and employ the principal approach to track people in infrared. The principal idea of our tracking approach is to integrate tracking directly into the ISM based detection approach and thus pursue person tracking solely based on the local image features (we use SIFT[14] in this work) employed for detection. This makes it perfectly suited for tracking in infrared, since no other features like color or rich texture are available here. Fig. 1 gives an overview of this tracking strategy. For a new image of an image sequence, SIFT features are extracted. Object hypotheses which are already known in the system at this point in time are matched with the new image features to determine where expectation and new data concur. This information is then integrated into the detection by assigning different classes to the features ((1) Image features without a match in the hypothesis feature set, (2) image features with a match in the hypothesis feature set, (3) hypothesis features without an according image feature) and by labeling the features with the according hypothesis ID in case of feature type 2 and 3. This feature set is matched with the codebook to generate the voting space. The voting space now includes the additional information on feature-hypothesis affiliation which can be employed to conduct tracking directly in object detection. For that, an independent mean shift search is started for every known hypothesis in particular. This search only includes votes labeled with the specific ID and native image feature votes. New objects are detected using the standard maximum search as described in.[12] For more detailed information on this tracking strategy we refer to our previous work.[13]

2.1 Evaluation

In this section, we evaluate our infrared person tracking approach with two main aspects. First, we show how tracking improves detection performance by comparing it to the results of standalone detection.[12] The second aspect is tracking performance (correct identity preservation in the system) itself. An example for the principal performance of our tracking approach in difficult situations is shown in Fig. 2 where two people move past each other. We see that our tracking approach is able to consistently maintain identities in this situation even when one person is mainly occluded and the recording camera is moving as in this case.

For quantitative tracking evaluation, we use the same metrics as in[13] and the image sequences (2 and 3) from[12] to compare tracking results with standalone detection. This comparison is shown in Fig. 3. It reveals the significant improvement of detection stability and accuracy in tracking. The performance of tracking itself is shown in Table 1. Here, we evaluate an additional sequence (1) that includes more challenges for tracking itself since this sequence was taken from a moving vehicle and people themselves are running.

The results in Table 1 show the nearly perfect performance in sequence 2 with a Multiple Object Tracking Accuracy (MOTA, see[13] for details) of 0.93 and no mismatch. This tracking performance might have been expected since, as we see in the second row of Fig. 4, this is a rather easy sequence for tracking since only

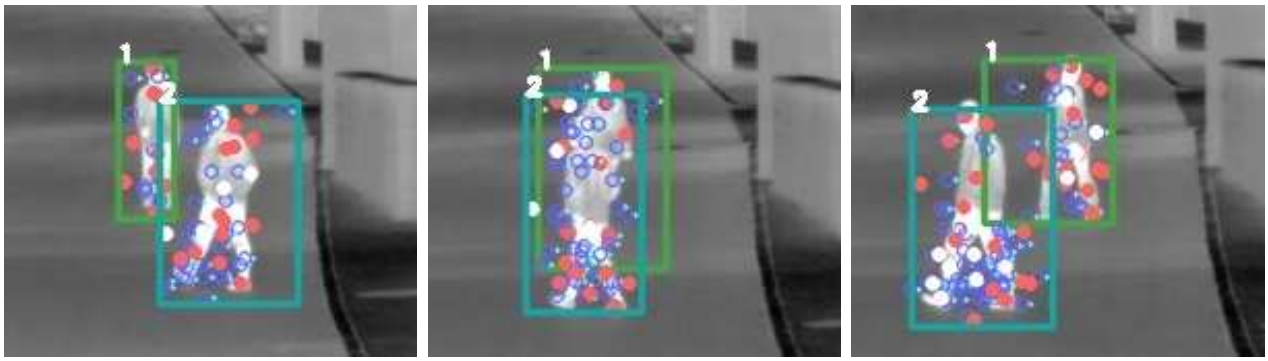

Figure 2. Results of our tracking approach in a situation where two people move past each other.

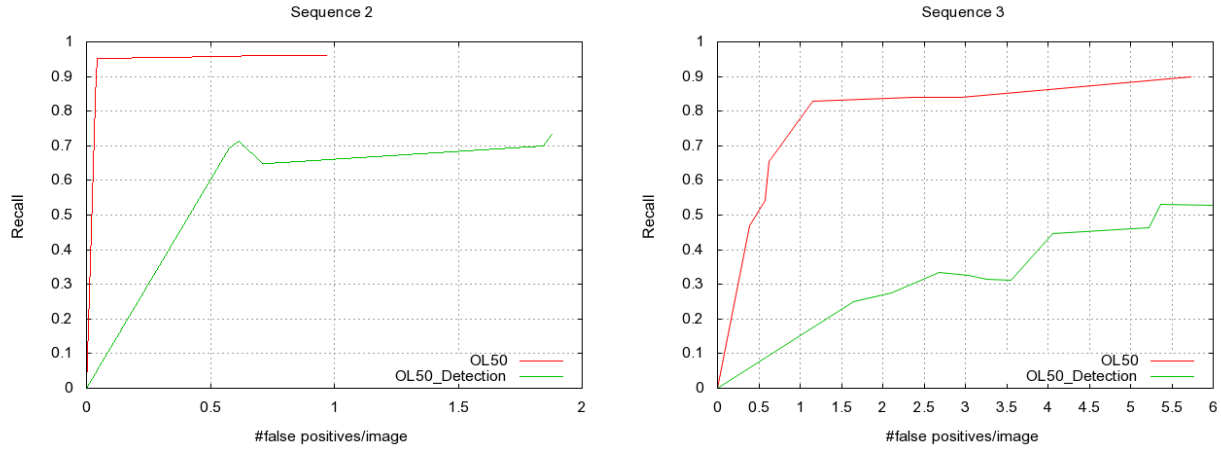

Figure 3. Recall/false positive curves for sequence 2 (left) and sequence 3 (right). Each chart contains two graphs which refer to the performance of tracking (red) and standalone detection (green) regarding the 50% bounding box overlap criterion.

two people move around in the scene. In the more challenging scenario in sequence 3, tracking shows a good performance too with a MOTA of 0.72. The main challenge for tracking here is identity maintenance for the 4 persons in the back of the scene which appear at a very low scale. Sequence 1 is the most challenging one for tracking because the camera and people are moving. Even here, we accomplish reasonable results with a MOTA of 0.76. Sample results of these three sequences are shown in Fig. 4.

Table 1. Tracking results for sequences 1-3.

Sequence	**1**	**2**	**3**
Frames	417	400	201
Objects (#ids)	1119 (8)	763(2)	1471 (8)
Miss rate	0.16	0.05	0.16
False positive rate	0.08	0.02	0.12
Mismatch rate	0.001 (2)	0	0.001 (2)
MOTA	**0.76**	**0.93**	**0.72**

Figure 4. Example tracking results of sequence 1 (top row), 2 (second row) and 3 (bottom row).

3. LIMB TRACKING

In,[12] we showed how the ISM based person detection approach can be employed to detect a person's body parts. In this section, we introduce a limb tracking approach that builds on this limb classification approach and employs feature trajectories which are built by feature propagation (see section 2) in person tracking. The goal of this approach is to build semantically annotated feature trajectories based on information of two levels: the signal level by employing feature tracks from person tracking and the semantic level by using the limb information to support and label low level trajectories. By that, both levels benefit from that fusion. Feature track failures (or instability over a longer time interval) on the descriptor level may be compensated by the higher level limb tracks. Failures or misclassifications on the limb level can be compensated by the signal level tracks since this level provides richer information.

For limb tracking, a number of unique limb identifiers e. g. "left leg", "head", "right foot" is declared and mapped to the assignable limb categories, e. g. "foot". The person tracking provides feature tracks over time and the limb classification approach described in[12] is applied to assign limb class probabilities to features. A track process for a certain limb is started if a minimum stability demand is met for a feature, e. g. a feature with a certain tracking ID should have been classified as a certain limb three succeeding times. If the limb category contains a limb presently not tracked, a new track is initialized. E. g. a new track for "left leg" could be initialized if "left leg" is currently not tracked. If more than one limb of the limb category is not tracked, a random one is initialized. Starting from these initial limb/feature combinations, existing limb trajectories are

Table 2. Limb classification results.

	head	**torso**	**leg**	**foot**	**fp/img**
Temporally isolated	0.62	0.2	0.55	0.64	0.76
Tracking	0.96	0.72	0.6	0.98	0.45

successively updated with new features based on three different criteria: (1) consistency of signal level tracks, (2) consistency of limb semantics and (3) spatiotemporal consistency. For that, the match (ranging from 0 to 1) of a new image feature f_k with an existing limb trajectory γ_i is based on a measure that integrates these three criteria:

$$p(\gamma_i|f_k) = \frac{\omega_T\beta_T(f_k,\gamma_i) + \omega_L\beta_L(f_k,\gamma_i) + \omega_S\beta_S(f_k,\gamma_i)}{\omega_T + \omega_L + \omega_S}. \quad (1)$$

Where the signal level match β_T is 0 if the tracking feature ID does not match and 1 if the ID does match. The limb match β_L is based on the limb classification for this feature and β_S is the spatial match of last limb position and feature position. This spatial position is modeled in the reference system of the person (the offset to person center which is determined by the ISM). By that, this approach is applicable despite camera motion. Both β_L and β_S are deduced from the euclidean distance but are converted to a match measure by:

$$\beta_{L|S} = 1.0 - (dist_{L|S}/MaxDist_{L|S}). \quad (2)$$

Since the limbs are not independent, the assignment problem between limb tracks and new features is solved by bipartite graph matching. This results in the overall best assignment considering all dependencies between limbs. The final assignment decision is based on these results but also considers the absolute match between limb track and feature. Here, we require a minimum match. If a limb track is not updated with new evidence at a certain point in time, the limb reliability is adapted recursively whereby the limb reliability is the sum of all single limb matches in a track, normalized with the lifetime of the limb track. By that, limb tracks with a stable history are retained longer since limb reliability decreases slower. Tracks with a low reliability are discarded and may be started again if the requirements are met.

3.1 Evaluation

For evaluation we pick a subsequence of sequence 1 and evaluate the limb trajectories and classification accuracy of a single person which is moving orthogonal to the recording camera. We evaluate the limbs "head", "foot", "leg" and "torso". Comparison of temporally isolated classification results with tracking classification results are shown in Table 2. These results show the improvement in limb classification per se and indicate how stable the limb tracks are. Limb tracking involves additional difficulties than just classify limbs correctly. Here, the link between a limb identifier and the real limb should be consistent over the whole sequence, e. g. the identifier "left leg" should be assigned to the person's same leg over the whole sequence. In our experiments, we have only two permutation for both "leg" and "foot". Additionally, the combination of leg and foot has to be consistent, this is not always the case in our experiments. Improvement could be made by the inclusion of additional model knowledge into the limb tracking approach. This could include the dependencies between limbs, for instance represented by a graph. Sample results of the limb tracking are shown in Fig. 5.

4. CONCLUSION

We presented an approach of person tracking in infrared that is based solely on local image features. The evaluation of this person tracking approach in different image sequences, including sequences taken from a moving camera, showed good performance specifically regarding identity maintenance during occlusions. We introduced a limb tracking that builds on this tracking and is able to track a person's limbs even under camera motion. This approach only includes little model knowledge and thus could be easily adapted for object component tracking in general. For the specific case of tracking a person's limbs, we expect the inclusion of additional model knowledge to improve the results.

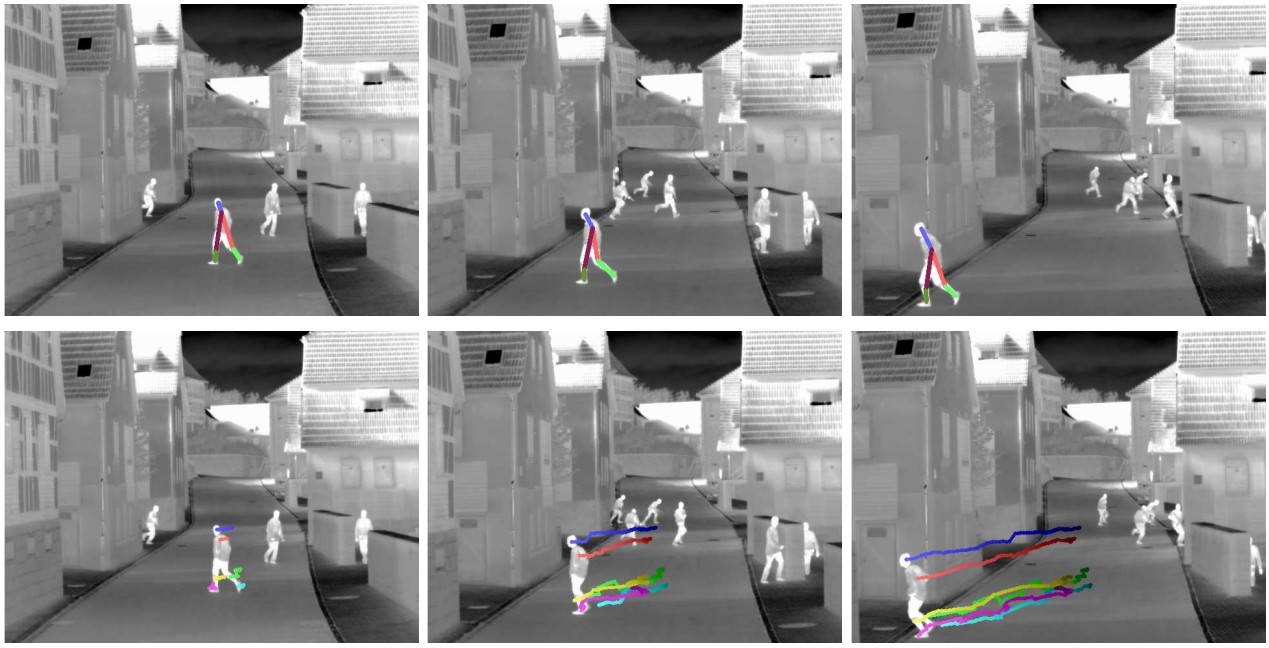

Figure 5. Limb based person skeleton (top-row) and limb trajectories (bottom-row).

REFERENCES

[1] Haritaoglu, I., Harwood, D., and Davis, L., "W4s: A real-time system for detecting and tracking people in 2.5 d," in [*Proc. ECCV*], 877–893 (1998).

[2] Davis, J. and Sharma, V., "Robust background-subtraction for person detection in thermal imagery," in [*Proc. CVPRW*], 128–135 (2004).

[3] Viola, P. and Jones, M., "Rapid object detection using a boosted cascade of simple features," in [*Proc. CVPR*], **1**, 511–518 (2001).

[4] Wu, B. and Nevatia, R., "Detection and tracking of multiple, partially occluded humans by bayesian combination of edgelet based part detectors," *IJCV* **75**, 247–266 (November 2007).

[5] Leibe, B., Leonardis, A., and Schiele, B., "Robust object detection with interleaved categorization and segmentation," *IJCV* **77**(1-3), 259–289 (2008).

[6] Dalal, N. and Triggs, B., "Histograms of oriented gradients for human detection," in [*Proc. CVPR*], 886–893 (2005).

[7] Suard, F., Rakotomamonjy, A., Bensrhair, A., and Broggi, A., "Pedestrian detection using infrared images and histograms of oriented gradients," in [*Proc. IV*], 206–212 (2006).

[8] Xu, F. and Fujimura, K., "Pedestrian detection and tracking with night vision," in [*Proc. IV*], 21–30 (2002).

[9] Leibe, B., Schindler, K., Cornelis, N., and Gool, L. V., "Coupled object detection and tracking from static cameras and moving vehicles," *IEEE PAMI* **30**(10), 1683–1698 (2008).

[10] Gavrila, D. and Munder, S., "Multi-cue pedestrian detection and tracking from a moving vehicle," *IJCV* **73**(1), 41–59 (2007).

[11] Hu, W., Tan, T., Wang, L., and Maybank, S., "A survey on visual surveillance of object motion and behaviors," *IEEE SMC* **34**(3), 334–352 (2004).

[12] Jüngling, K. and Arens, M., "Feature based person detection beyond the visible spectrum," in [*Proc. CVPRW*], 30–37 (2009).

[13] Jüngling, K. and Arens, M., "Detection and tracking of objects with direct integration of perception and expectation," in [*Proc. ICCVW*], 1129–1136 (2009).

[14] Lowe, D. G., "Distinctive image features from scale-invariant keypoints," *IJCV* **60**(2), 91–110 (2004).