Model-Based Occlusion Handling for Tracking in Crowded Scenes [1)]

Csaba Beleznai[1], *Bernhard Frühstück*[2], *Horst Bischof*[3] *and Walter Kropatsch*[4]

[1]Advanced Computer Vision GmbH - ACV, Vienna, Austria
csaba.beleznai@acv.ac.at
[2]Siemens AG Österreich, Programm- und Systementwicklung, Graz, Austria
[3]Inst. for Computer Graphics and Vision, Graz Univ. of Technology, Austria
[4]Pattern Recognition and Image Processing Group, Vienna Univ. of Technology, Austria

Abstract:
The task of reliable detection and tracking of multiple objects becomes highly complex for crowded scenarios. Data association is difficult to perform reliably in the presence of missing observations due to occlusions. We propose a novel real-time approach to segment and track multiple overlapping humans. The optimal segmentation solution is given by the maximum likelihood estimate in the joint-object space. The search for solution is guided by a fast mean shift procedure and relies on information on the number of humans involved in the occlusion which can be estimated using the tracking history. Results are presented for the task of human tracking in crowded scenes and evaluated in terms of tracking performance.

1 Introduction

Automated visual surveillance systems aim at obtaining a high-level representation for a given scene. To achieve this goal, object detection and tracking algorithms have to generate data providing a reliable basis for high-level functionalities. Realistic scenes, however, usually contain many interacting and occluding objects, leading to frequent detection and tracking failures.

Tracking systems proposed in recent years attempt to tackle increasingly complex scenarios and several approaches for occlusion handling have been suggested. Colour-based segmentation [5] in crowded scenes is usually of limited use since often colours are not sufficiently distinctive for different individuals. Silhouette analysis [6, 7] and stochastic segmentation from binary images [12] require a good segmentation quality in order to find landmark points such as heads or shoulders. Methods relying on particle filters [8], performing exploration of

[1)]This work has been carried out within the K plus Competence Center ADVANCED COMPUTER VISION. This work was funded from the K plus Program.

the solution space of possible human configurations are usually computationally prohibitive if the number of scene objects becomes large. Nevertheless, certain extensions of the standard particle filter approach, the mixture particle filter [10] and the boosted particle filter [9] seem to be promising algorithms for multi-target tracking.

In this paper we propose a novel occlusion handling scheme, which significantly improves the tracking performance even in the presence of a large overlap between objects. The optimal spatial arrangement, i.e. *configuration* of occluding humans is determined by searching for the maximum likelihood estimate in the space of joint-object configurations. The search employs a sampling scheme relying on the mean shift procedure and on priors with respect to the number and size of involved humans.

The paper is organized as follows: section 2 gives a brief overview on the applied human detection and tracking algorithms; describes the computation of a fast variant of the mean shift vector and presents the proposed model-based occlusion handling procedure in detail. Section 3 demonstrates tracking results employing the proposed occlusion handling scheme and evaluation of the improved tracking algorithm. Finally, the paper is concluded in section 4.

2 The tracking system

2.1 Human detection and tracking

A common technique to detect motion in a scene viewed by a stationary camera involves background modelling and subsequent change detection. We adopt a similar approach [3] and the obtained difference image is used to detect objects. The difference image can be thought as a mixture of clusters. Instead of thresholding, clustering is performed using a fast variant of the mean shift clustering procedure (see [1] for details). Local density maxima and associated basins of attraction [4] represent and delineate the object candidates. Each cluster also has an associated set of points $\{PX_1, ..., PX_n\}$, also referred to as *path-points*, defining the paths explored by the mean shift mode seeking steps. The set of path points is utilized to perform occlusion handling in a computationally efficient manner.

The object tracking algorithm relies on an incremental mode seeking computation over time. Further details on the tracking algorithm can be found in [1].

2.2 Fast Mean Shift Computation

The mean shift algorithm is a nonparametric technique to locate density extrema or modes of a given distribution by an iterative procedure [4]. Starting from a location x the local mean

shift vector represents an offset to x', which is a translation towards the nearest mode along the direction of maximum increase in the underlying density function. The local density is estimated within the local neighborhood of a kernel by kernel density estimation where at a data point a kernel weights $K(a)$ are combined with weights $I(a)$ associated with the data. Fast computation of the new location vector x' can be performed as [1]:

$$x' = \frac{\sum_a K''(a-x)ii_x(a)}{\sum_a K''(a-x)ii(a)} \tag{1}$$

where K'' represents the second derivative of the kernel K, differentiated with respect to each dimension of the image space, i.e. the x- and y-coordinates.

The functions ii_x and ii are the double integrals, i.e. two-dimensional integral images [11] in the form of:

$$ii_x(x) = \sum_{x_i<x} I(x_i)x_i \tag{2}$$

and

$$ii(x) = \sum_{x_i<x} I(x_i) \tag{3}$$

If the kernel K is uniform with bounded support, its second derivative becomes sparse containing only four impulse functions at its corners. Thus, evaluating a convolution takes only the summation of four corner values in the given integral image.

To compute the mean shift vector at location x, the following steps are performed: (1) three integral images (defined in Eq.2 and Eq.3) are precomputed in a single pass (see [11] and [2] for details); (2) the expression in Eq.1 is evaluated using only ten arithmetic operations and twelve array accesses. The number of operations is independent of the kernel size, given the sparse structure of K''.

2.3 Occlusion handling

If several objects meet and form a group, occlusion - partial or complete - between the objects might take place. Typically, before moving objects form a group, they can be tracked separately. Occlusion is detected by the data association algorithm. In such situations we employ a probabilistic approach - similarly to the technique described in [12] - to find the optimal configuration of humans best explaining the difference image data I. This task can be stated as a model-based segmentation problem.

We employ a very simple human shape model, a rectangular region. This region is equivalent

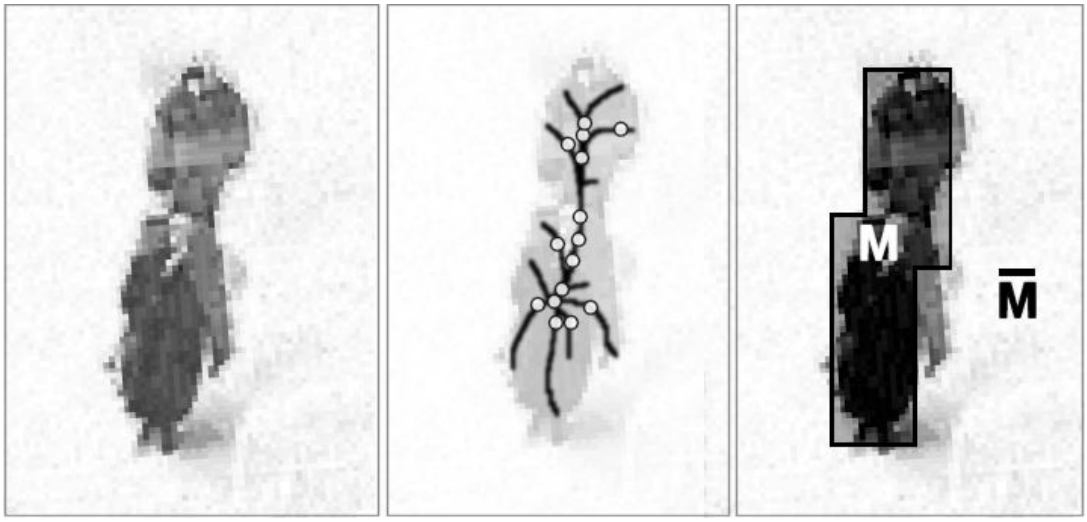

Figure 1: Example frame illustrating the approach to search for the most probable configuration of two humans in the presence of occlusion. Left: Occlusion between two humans shown in the inverted difference image. Center: Mode seeking is performed starting from a set of sample points. Obtained path points (shown as dots) represent possible locations of a human. Right: the optimal configuration of two objects for the given image region.

to the kernel used in the mean shift procedure. All parameters (height, width and orientation) of the rectangular region are known due to calibration. The tracking algorithm provides prior information on the number N of objects involved in the group formation. The search for θ_N^* - the most probable configuration consisting of N objects - in the space of possible configurations Θ_N becomes a maximum likelihood estimation problem:

$$\theta_N^* = argmax_{\Theta_N} P(I|\theta_N) \tag{4}$$

The unknown parameters are the locations of the humans $\{x_i, y_i\}_{i=1..N}$ in the occluded state.

When occlusion between the tracked objects is detected, we perform the following procedure:

1. A new sample set of points by locating local maxima is generated within a local image region spanned by the spatial extrema of involved object windows.

2. Starting from these points fast mean shift procedure is carried out until convergence (see Figure 1).

3. Sampling is guided by the path of mean shift procedures and path points, PX are used to hypothesize object locations. The mean shift algorithm has advantageous properties supporting this strategy: (1) the mean shift kernel becomes quickly centered on relevant data; (2) local plateaus or ridges on the density surface are distinguished by a large number of path points.

4. The likelihoods for individual human hypotheses are not independent, since inter-occlusion

between humans might be present. Therefore the joint likelihood for multiple humans has to be formulated.

A hypothesized configuration θ_N divides the difference image into two image regions: pixels explained by the configuration and pixels outside of the configuration. If M_i is the image region occupied by the i_{th} model, the union of image regions $M = \bigcup_{i=1}^{N} M_i$ defines a mask containing all pixels explained by the configuration. Accordingly, $\overline{M}$ denotes the complementary region outside of the models (see Figure 1). The local image region R around the occluding objects is given by $R = M \bigcup \overline{M}$.

A configuration maximizing the likelihood should fulfill following criteria: (1) maximizing the sum of difference image intensities within the model region M, while (2) minimizing the sum of difference image intensities in $\overline{M}$, outside of the models. A log-likelihood function expressing this balance between the two quantities can be formulated as:

$$\begin{aligned} ln\ P(I|\theta) \ &\propto \ A \sum_{x \in M} I(x) - (1 - A) \sum_{x \in \overline{M}} I(x) \\ &\propto \ A \sum_{x \in M} I(x) - \sum_{x \in R} I(x) \quad , \end{aligned} \tag{5}$$

using the complementarity between M and $\overline{M}$ and the experimentally determined weight A.

The above quantity is evaluated for the configuration θ_N. Fast evaluation of the likelihood expression of Eq. 5 can be performed as follows:

The sum of pixel intensities within the kernel centered at the i_{th} path point, i.e. the area sum S_i is obtained during the mean shift procedure. The first term of Eq. 5 can be computed by: (1) taking the sum of area sums at the sampled locations and (2) correcting for possible overlaps between hypothesized models.

Since the models are represented by rectangular regions with sides parallel to the image border, the overlap regions can be easily computed. The maximum number of possible overlaps between N objects is $\frac{N(N-1)}{2}$. Then, the sum of pixel intensities in the region covered by models (first term of Eq. 5) can be computed as:

$$\sum_{x \in M} I(x) \ = \ \sum_{i=1}^{N} S_i \ - \ \sum_{x \in V} I(x) \ , \tag{6}$$

where V denotes the union of overlapping regions. The union of overlapping regions is determined by examining the intersections between all overlap regions. Since pairwise overlaps span rectangular regions, therefore - using the integral images defined in Eq. 3 - the sum of pixel intensities within an overlap region can be obtained by three arithmetic operations. The second term of Eq. 5 - representing the sum of pixel intensities in the entire region R - is needed to be computed only once using the integral image ii.

Generally, in our scenarios two, rarely three objects form an occluded group. Typically 5-12 path points are used for hypothesizing object locations, thus in the worst case, evaluation of a couple of thousand configurations is necessary. All the configurations are evaluated and the best configuration is taken. The models of the best configuration are associated - using a nearest neighbor criterion - with the predicted cluster centers and trajectories are updated accordingly.

3 Results and discussion

Figure 2: Tracking results obtained using the proposed occlusion handling scheme. (a) Example frame showing tracking results for Sequence A containing people moving along relatively straight trajectories. (b) Obtained tracks for a scene of Sequence B viewed by a top-mount camera. Humans in this scene perform more irregular movements.

Two video sequences depicting crowded scenes were used to evaluate the proposed occlusion handling scheme. Sequence A (Figure 2, left) consists of 1676 frames with an image resolution of 360-by-288 pixels. Sequence B (Figure 2, right) depicts a scene viewed from the top consisting 731 frames with a resolution of 360-by-288 pixels. Tracking results using the described occlusion handling approach are superimposed. Stable detection and tracking results are obtained. If the occlusion between objects involves more than two persons, the humans switch positions and the duration of occlusion is long, tracking errors still might appear. These errors are mainly due to association errors generated by the simple nearest neighbor-based association rule.

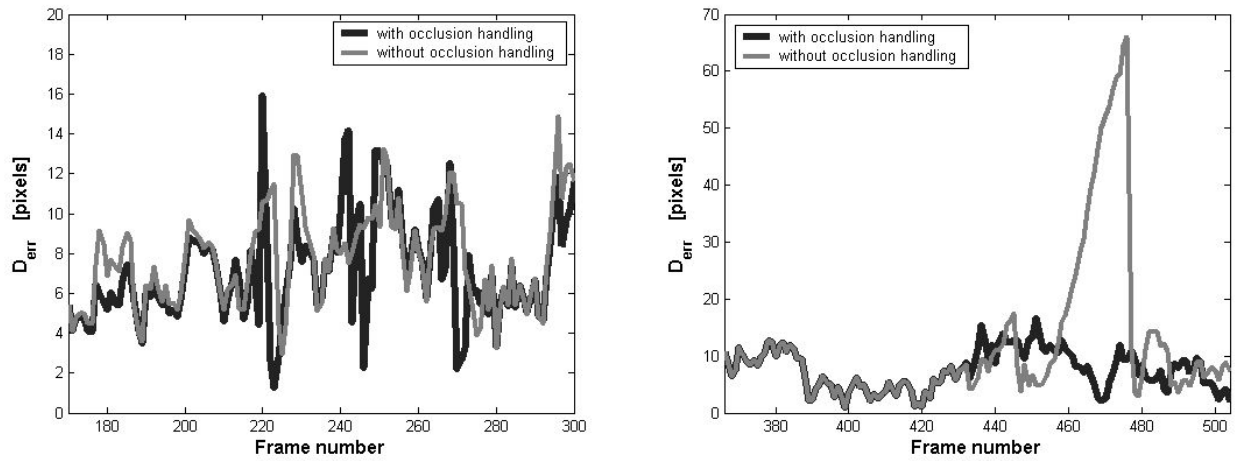

Figure 3: Spatial tracking errors computed relative to ground truth. Error measures were computed for a trajectory of a human undergoing frequent occlusions, where the trajectory is obtained by tracking with (black line) and without (gray line) occlusion handling. Left and right plots show error measures obtained for a track of Sequence A and Sequence B, respectively.

Quantitative evaluation of the tracking performance was carried out. The centroid positions of humans were determined manually for both sequences. Tracking results were obtained using the described occlusion handling technique. To assess the efficiency of the presented occlusion handling scheme, we also applied the detection and tracking framework to both sequences without performing occlusion handling. In this case, upon occlusion between targets - detected by a simple overlap criterion -, measurement update by mode seeking [1] was not carried out and the tracked object was only guided by its motion model.

A single trajectory of a human undergoing several occlusions was selected in both sequences. The trajectory data obtained with and without occlusion handling was compared to the ground truth trajectory. The tracking error in term of spatial distance relative to the ground truth trajectory was computed for every frame. Tracking errors obtained for a selected trajectory in Sequence A and B are shown in Figure 3. As it can be seen from the left plot for Sequence A, tracking results obtained with and without occlusion handling exhibit similar performance. Due to the smooth movement of humans in this scene, even a simple first-order motion model estimates the object positions in occlusion events of short duration quite successfully. Sequence B contains humans moving along strongly curved trajectories, therefore tracking without occlusion handling leads to a failure where the tracked object is lost and a new track is initiated (see large peak in the right plot of Figure 3). Occlusion handling in such cases estimates the local configuration of humans successfully and the track remains on the target.

The proposed method is implemented in C++ and runs in real time on a 2.5 GHz PC for all of the presented sequences.

4 Conclusion

A simple and efficient scheme is proposed for segmenting occluded objects in a crowded scene. The presented technique performs well and stable tracking over occlusions is obtained. The configuration of occluding targets, maximizing the image likelihood, is determined efficiently using a sampling step guided by mean shift mode seeking and exploiting the use of integral images for fast integration of image intensities over rectangular image regions. Real-time tracking performance is achieved and demonstrated for some difficult scenarios.

References

[1] C. Beleznai, B. Frühstück, and H. Bischof. Tracking multiple humans using fast mean shift mode seeking. In *IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, pages 25–32, Breckenridge, USA, January 2005.

[2] C. Beleznai, B. Frühstück, H. Bischof, and W. Kropatsch. Detecting humans in groups using a fast mean shift procedure. In *Proc. of the 28th Workshop of the Austrian Association for Pattern Recognition (OEAGM/AAPR)*, pages 71–78, Hagenberg, Austria, 2004.

[3] R. Collins, A. Lipton, T. Kanade, H. Fujiyoshi, D. Duggins, Y. Tsin, D. Tolliver, N. Enomoto, and O. Hasegawa. A system for video surveillance and monitoring: VSAM final report. In *Technical Report CMU-RI-TR-00-12*, Robotics Institute, Carnegie Mellon University, 2000.

[4] D. Comaniciu and P. Meer. Mean shift analysis and applications. In *IEEE Int. Conf. Computer Vision*, pages 1197–1203, Kerkyra, Greece, 1999.

[5] A. Elgammal, R. Duraiswami, and L. S. Davis. Efficient nonparametric adaptive color modeling using fast gauss transform. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, volume 2, pages 563–570, December 2001.

[6] I. Haritaoglu, D. Harwood, and L. S. Davis. W4: Real-time surveillance of people and their activities. *IEEE Trans. on PAMI*, 22(8):809–830, 2000.

[7] Y. Kuno, T. Watanabe, Y. Shimosakoda, and S. Nakagawa. Automated detection of human for visual surveillance system. In *Int. Conf. on Pattern Recognition*, page C92.2, Vienna, Austria, August 1996.

[8] S. Maskell, M. Rollason, N. Gordon, and D. Salmond. Efficient particle filtering for multiple target tracking with application to tracking in structured images. *IVC*, 21(9):931–939, September 2003.

[9] K. Okuma, A. Taleghani, N. De Freitas, J. J. Little, and D. G. Lowe. A boosted particle filter: Multitarget detection and tracking. In *European Conference on Computer Vision*, volume 1, pages 28–39, 2004.

[10] J. Vermaak, A. Doucet, and P. Perez. Multi-modality through mixture tracking. In *Proc. IEEE International Conference on Computer Vision*, volume 2, pages 1110–1116, 2003.

[11] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, volume 1, pages 511–518, Kauai, Hawai, 2001.

[12] T. Zhao and R. Nevatia. Bayesian human segmentation in crowded situations. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 459–466, Madison, USA, June 2003.