Human Tracking by Fast Mean Shift Mode Seeking

C. Beleznai
Advanced Computer Vision GmbH - ACV, Vienna, Austria
Email: csaba.beleznai@acv.ac.at

B. Frühstück
Siemens AG Österreich, Programm- und Systementwicklung, Graz, Austria
Email: bernhard.fruehstueck@siemens.com

H. Bischof
Institute for Computer Graphics and Vision, Graz University of Technology, Graz, Austria
Email: bischof@icg.tu-graz.ac.at

***Abstract*— Change detection by background subtraction is a common approach to detect moving foreground. The resulting difference image is usually thresholded to obtain objects based on pixel connectedness and resulting blob objects are subsequently tracked. This paper proposes a detection approach not requiring the binarization of the difference image. Local density maxima in the difference image - usually representing moving objects - are outlined by a fast non-parametric mean shift clustering procedure. Object tracking is carried out by updating and propagating cluster parameters over time using the mode seeking property of the mean shift procedure. For occluding targets, a fast procedure determining the object configuration maximizing image likelihood is presented. Detection and tracking results are demonstrated for a crowded scene and evaluation of the proposed tracking framework is presented.**

***Index Terms*—automated visual surveillance, motion detection, mean shift clustering, human tracking, occlusion handling**

I. Introduction

Scenes of practical interest usually contain a large number of interacting targets under difficult imaging conditions. In such circumstances the task of reliable object detection and tracking becomes non-trivial and obtaining a meaningful high-level representation poses a challenging task.

Human detection and tracking systems proposed in recent years attempt to tackle increasingly complex scenarios. Motion detection is an essential part of automated visual surveillance systems; however, reliable segmentation of moving regions into individual objects of interest still represents a great challenge. For instance, blob-based motion segmentation in the presence of interacting targets typically generates objects which are under- or oversegmented (see Fig.1.a and Fig.1.b). Blob-based motion segmentation relies on thresholding, where the threshold is a sensitive parameter leading to an immediate decision whether a pixel belongs to a moving or non-moving region. Thresholding eliminates relevant information and motion segmentation errors are difficult to correct afterwards given the poor quality of binary images.

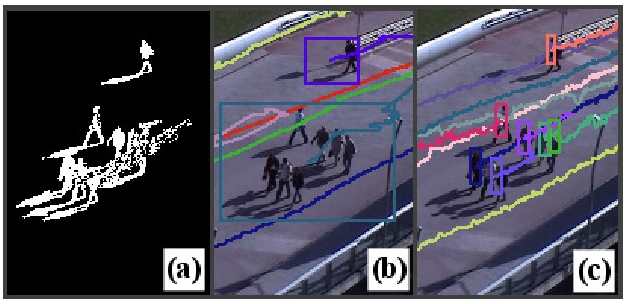

Figure 1. Blob analysis (a) typically produces undersegmented results for humans in groups leading to low detection rates and poor tracking results (b). Image (c) illustrates the proposed detection and tracking approach generating significantly improved results.

In this paper we propose a novel detection and tracking scheme directly operating on the difference image obtained by background subtraction. The method shows good tracking performance in crowded scenarios, even in the presence of a large overlap between objects. A fast variant of mean shift clustering is applied to delineate objects and mode seeking along the density gradient of the difference image is used to propagate and update object properties. Upon occluding objects the optimal spatial arrangement, i.e. the *object configuration* 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

Based on “Human Tracking by Mode Seeking”, by C. Beleznai, B. Frühstück, and H. Bischof which appeared in the Proceedings of the 4th International Symposium on Image and Signal Processing and Analysis, Zagreb, Croatia, September 2005. © 2005 IEEE.

shift procedure and on priors with respect to the number and size of involved humans.

The paper is organized as follows: section II describes related work. Section III provides a brief overview on the applied fast mean shift procedure using a uniform kernel. Section IV describes the mean shift clustering-based object detection technique. Section V gives details on the tracking algorithm based on mode propagation and describes the occlusion handling scheme using a simple human model. Section VI provides an algorithmic summary of the tracking system. Section VII presents detection and tracking results and their performance evaluation. Finally, the paper is concluded in section VIII.

II. Related Work

Human detection by blob analysis is used in many approaches [1]. However, inferring the position of individual humans from the binary segmentation results by shape analysis [2] or by stochastic segmentation [3] requires good segmentation quality in order to find landmark points such as heads or shoulders.

Blob-based analysis can be complemented by appearance [4] or color information [5], enabling a tracking system to better cope with occlusions. Color-based segmentation [6] in crowded scenes can be also used for tracking if colors are distinctive for different individuals.

Pece [7] proposed clustering in the difference image using mixtures of Gaussians and tracking by propagating cluster parameters. Due to the Gaussian assumption on the cluster shapes, interacting and occluding targets are often clustered together. Our approach also performs difference image clustering; however, without relying on specific assumptions with respect to the distribution of the data. Thus nearby density maxima, i.e. cluster centers are kept separate.

Color- or histogram based tracking [8] performs mode seeking along the gradient of a histogram similarity function. Our tracking approach adopts a similar mode seeking strategy, but mode seeking in our case is performed to track density maxima in the difference image.

In the context of multiple target tracking, particle filtering recently appeared as a promising technique [9]. It is capable to integrate different mechanisms, such as visual object recognition [10], color tracking and occlusion handling [11].

Our work proposes a simple, computationally efficient object detection and multi-target tracking framework, which can be also combined with existing detection and tracking techniques.

III. The Fast Mean Shift Procedure

Object detection is performed by delineating clusters in the difference image by the mean shift mode seeking procedure. The mean shift algorithm is a nonparametric technique to locate density extrema or modes of a given distribution by an iterative procedure [12]. 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 in [13]:

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

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 [14] in the form of:

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

and

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

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 (2) and (3)) are precomputed in a single pass (see [14] and [15] for details); 2. the expression in (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''.

IV. Mean Shift Clustering

The clustering step is facilitated by the use of a human size model $\{H(x), W(x)\}$, where H and W denote human height and width, respectively. This information is obtained by a simple calibration step.

The principal steps of mean shift clustering are performed analogously to the steps described in [16]:

1. The difference image intensity maximum is mapped to unit intensity and its entire range is scaled proportionally.

2. A sample set of n points $X_1 \ldots X_n$ is defined by locating local maxima - above a very low threshold T_1 - in the difference image.

The final result does not depend critically on T_1. A very low value just increases the run time and generates more outliers which can be eliminated during the mode tracking step.

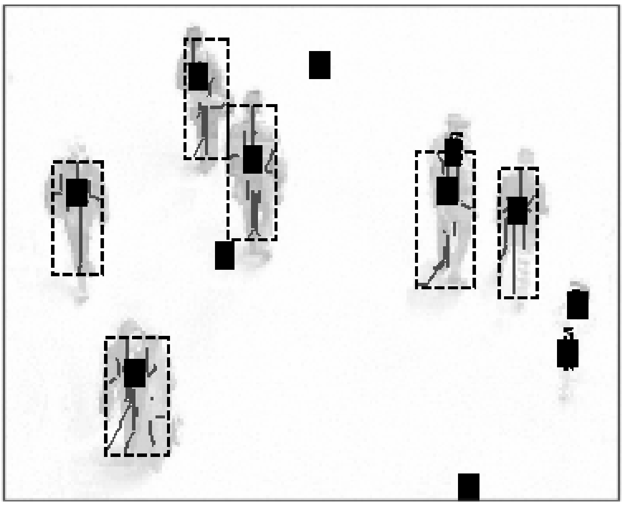

Figure 2. Example for fast mean shift based clustering in a difference image (shown inverted). Obtained clusters are delineated by rectangular basins of attraction (rectangles with dashed line). Cluster centers (black dots) and connected path point sets (shown as lines running towards cluster centers) are also shown. Note that some clusters are generated by noise and motion clutter.

3. The fast mean shift procedure is applied to the points of the sample set with a window size of (*$H(X_i)$*, *$W(X_i)$*) according to the local size model. The mean shift procedure converges to the nearest mode typically within 3-4 iterations.

The mode seeking process delineates a path between the initial point of the sample set and the detected local mode candidate. Each mean shift iteration defines a point on the path, what we denote as a path-point {*PX*}. Thus, each detected mode candidate location has an associated set of path-points {PX_1,…, PX_n}, not including the mode itself.

When the mean shift offset vector is computed according to (1), the area sum (i.e. sum of pixel intensities) within the kernel (denominator of the expression in (1)) is also obtained. The set of area sum magnitudes {S_1,…,S_n} is useful to have since it provides information on the magnitude of the local density and as we will see later, it can be used in the occlusion handling step evaluating a given spatial configuration of kernels.

4. Given the finite size of the mean shift convergence criterion, detected mode candidate locations - obtained for the same peak of underlying density - might slightly deviate. Detected mode candidates are linked based on spatial proximity: all detected modes within a window of the size (*W*, *H*) are grouped together and a cluster center *Y* is obtained by taking the mean of linked candidate coordinates. Path-point sets belonging to grouped mode candidates are also merged, such as the sets of area sum magnitudes. The merged set of path points is used to delineate the cluster: a bounding box representation of the basin of attraction is obtained by determining the spatial extrema of path points in *x*- and *y*-directions.

The above clustering process yields following information for a given cluster *i*: a cluster center Y_i, a set of path-points $\{PX_1^i,\ldots,PX_k^i\}$, the basin of attraction boundaries in form of a bounding box and a set of area sum magnitudes $\{S_1^i,\ldots,S_k^i\}$. An illustrative example depicting these constructs is shown in Fig 2.

At this stage the detected clusters can be considered as object candidates and probable outliers can be eliminated by imposing size constraints on the size of the attraction basins.

V. Cluster Tracking

A. Tracking by Mode Seeking

Moving objects of a scene usually represent moving local density maxima in the corresponding sequence of difference images. The mode seeking property of the mean shift procedure implies that a mode can be pursued by a repetitive mean shift procedure: for each mode displacement in the difference image - assuming that the interframe displacement is much smaller than the kernel size - the mode location can be repeatedly found.

The principal advantages of this tracking strategy are: 1. the data association problem is solved implicitly, since the mode seeking procedure is guided to the nearby mode along the steepest density gradient; 2. it represents a simple and computationally efficient technique, because only a few fast mean shift iterations are sufficient to re-detect the object. Furthermore, the mode seeking process can be easily complemented by an underlying motion model.

The disadvantage of the above strategy is that such a sequential mode seeking assumes the spatial distinctiveness of available modes. When several density maxima are in spatial proximity - such as in a difference image of a crowded scene containing humans occluding each other -, the distribution locally might become strongly non-Gaussian and mode candidates tend to exhibit coalescence, leading to the breakdown of affected tracking processes.

If objects in the original image and corresponding difference image density extrema are spatially well-separated, the mean shift mode seeking procedure can reliably track them. In the initial frame the entire difference image is evaluated by the fast mean shift clustering algorithm as described in section IV. The obtained cluster centers are then used in subsequent frames as the points of a new sample set *X'*. Starting from these points the fast mean shift procedure is carried out (using the locally-scaled uniform kernel of height *$H(x)$* and width *$W(x)$*) locating the nearby mode candidate which corresponds to the new location of the moving object, i.e. the new cluster center. For spatially isolated mode candidates we do not compute additional cluster parameters, such as basin of attraction or path-points, since we assume that the underlying distribution varies only slightly with respect to its shape.

In the following, the different cases of the tracking framework are described and a possible solution for coalescing mode candidates is presented.

B. Occlusion Handling

If several objects meet and form a group, occlusion - partial or complete - between the objects might take place. Such an event generates overlapping or very close density extrema in the difference image. Typically one specific mode attracts all or most of the nearby mode seeking procedures. We denote this phenomenon as "*mode candidate stealing*". The occurrence of mode candidate stealing can be easily detected, since two or more mode candidates appear in close proximity.

Typically, before moving objects form a group, they can be tracked separately, as described in subsection *A*. After each tracking step, it is examined whether at least another cluster center exists within a window of (*0.5H(x), 0.5W(x)*) around a detected cluster center. If this is the case, mode candidate stealing has occurred implying that the local configuration of humans cannot be obtained by mode seeking.

In such situations we employ a Bayesian approach - similarly to the technique described in [3] - to find the local optimum 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 to the kernel used in the mean shift procedure. All parameters (height, width and orientation) of the rectangular region are known.

The tracking algorithm provides prior information on the number of objects involved in the group formation. Thus, the number of models *N* needed to explain the data is also available.

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:

$$\theta_N^* = \arg\max_{\theta_N} P(I|\theta_N)\,. \qquad (4)$$

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

When we detect mode candidate stealing, we perform the following steps to find the optimum local configuration of humans:

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 iterations are carried out until convergence (see Fig. 3, center).

3. The mean shift algorithm has the property that it runs along the path of the maximum increase in the underlying density. Furthermore, the magnitude of the mean shift offset correlates with the local magnitude of the density gradient. These properties have following implications: 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, *PX* (see section IV).

Our sampling procedure is guided by the path of mean shift runs. We use the path points as a candidate set of possible object locations. We also make use of area sum magnitudes *S*, which are available at these locations, obtained as a "by-product" of mean shift computation. This strategy significantly reduces the search space and facilitates the fast evaluation of a given configuration.

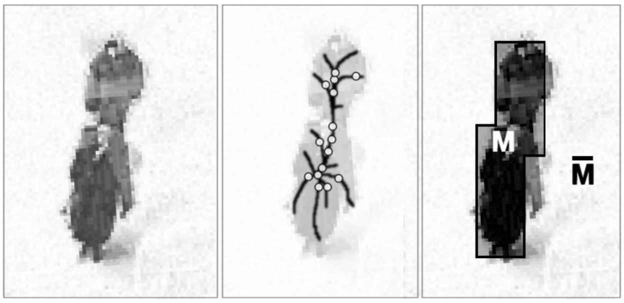

Figure 3. Example illustrating the approach searching for the most probable configuration of humans in the presence of occlusion. Left: Occlusion between two humans shown in the inverted difference image. Center: Mean shift mode seeking is performed starting from a set of sample points. Obtained path points (shown as dots) represent possible locations of a human. Right: the found optimum configuration of the two humans for the given image regions.

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 Fig. 3). The local image region *R* around the occluding objects is given by $R = M \cup \overline{M}$.

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

$$\ln P(I|\theta) \propto a_1 \sum_{x\in M} I(x) - a_2 \sum_{x\in \overline{M}} I(x)$$

$$\propto A \sum_{x\in M} I(x) - \sum_{x\in R} I(x)\,, \qquad (5)$$

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

5. The above quantity is evaluated for the configuration θ_N . Fast evaluation of the likelihood expression of (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 (1) 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 in (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 image defined in (3) - the sum of pixel intensities within an overlap region can be obtained by three arithmetic operations.

The second term of (5) - representing the sum of pixel intensities in the entire region R - is needed to be computed only once using the integral image *ii*.

6. Generally, in our scenarios the number of occluding humans is given by a small number; usually 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.

The models of the best configuration are associated - using a nearest neighbor criterion - with the predicted cluster centers and trajectories are updated accordingly.

When using a blob-based detection system, occlusions between objects often generate object merging, rendering the tracking task difficult. The presented approach provides N measurements even in the case of complete occlusion between N objects, due to the use of N as prior.

If the prior information on N is incorrect - due to detection or tracking failures - the error is propagated further, over the duration of occlusion events. This problematic issue is not handled in the present approach.

C. Appearance of New Objects

New objects are detected using a simple scheme. For all previously detected objects, the difference image is reset to zero intensity within the local kernel. The residual difference image is analyzed again for the existence of clusters, as described in section IV.

Coalescence of a newly-created cluster with a nearby cluster indicates that the appearing object is identical with an existing object: in such cases the appearing object is deleted.

D. Object Disappearance

If the mean shift offset for a tracked object remains zero over a given time period, three possibilities exist:

1. the object has come to a full stop,
2. the object is generated by noise or clutter,
3. the object has disappeared.

To evaluate such a case, the difference image region within the object kernel is examined. A new set of sample points is generated by selecting local maxima and a clustering step according section IV is carried out. The basin of attraction of the cluster is delineated. If the dimensions of the basin of attraction significantly deviate from the local scaling of a human, the object is deleted.

VI. Algorithmic Summary

The algorithm of cluster center tracking proceeds according to the following main steps:

1. Integral images (2) and (3) are computed in a single pass.
2. Performing clustering in the initial frame of the difference image. Cluster attributes (cluster center, basin of attraction, set of path points and area sums) are determined.
3. Inter-frame cluster center tracking by fast mean shift procedure. A linear motion model is applied.
4. Testing for coalescence between mode candidates. If mode candidate stealing is detected, the most probable configuration is searched using the number of involved objects as priors.
5. Testing for appearance of new objects
6. Testing for disappearance of objects.

The cluster center tracking algorithm performs fast cluster center propagation for spatially isolated objects, and - in the case of occlusions - a computationally efficient scheme proposes the optimum configuration of occluding objects.

VII. Results and Discussion

Background differencing was carried out applying a motion detection technique [17] using an adaptive background model. One sequence (4600 frames, resolution: 360×288 pixels) depicting a scene of walking people was selected for evaluation. The humans in the scene cast shadows and motion clutter in form of a moving flag and moving vegetation is present (see Fig. 4).

The sequence was processed by the proposed tracking approach (Fig. 4.b) and also by a common blob-based tracking algorithm (Fig. 4.a). Blob-based detection was based on the method described in [17]. The blob tracking algorithm generated a new trajectory hypothesis each time when a new blob object appeared in the image. Blob objects and existing track hypotheses were matched by computing the overlap between their bounding boxes. A linear motion model was assumed. During occlusions hypothesized trajectories were solely guided by the motion model.

Tracking results obtained for the proposed tracker show that it copes well with occlusions and shadows. Trajectories remain stable for all the targets, as it can be seen in Fig. 4.b. Occlusions between two and rarely between three persons are resolved successfully using the proposed model-based occlusion handling scheme. Shadows are detected as isolated mode candidates. In the subsequent cluster delineation step (see section IV) these

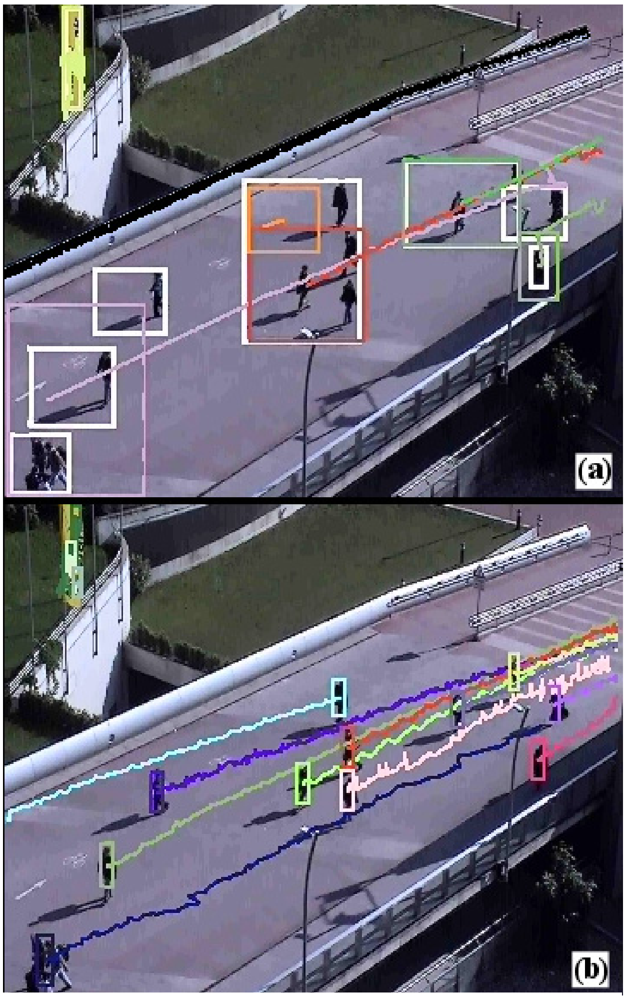

Figure 4. Tracking results in the case of a (a) blob-based tracking approach and (b) for the proposed tracking scheme.

mode candidates are eliminated based on the extent of the corresponding basins of attraction. Note, that shadows in this sequence are oriented nearly horizontally, thus their elimination based on geometric constraints works well. In cases where the shape and orientation of shadows is similar to those of the humans in the scene, shadows are detected and tracked as valid mode candidates.

Blob-based tracking (Fig. 4.a) yields only trajectory segments and trajectories representing the motion of humans in groups. Blob detection relies on connected component analysis, which leads to poor object segmentation quality in the case of overlapping humans and/or shadows. Segmented blob boundaries are shown in Fig. 4.a as white bounding boxes. Tracking failures arise from the under-segmented objects and due to the lack of measurement update during occlusions.

In order to quantitatively assess the detection performance, detection results were compared to a ground truth. As ground truth, the bounding boxes of humans were determined manually for a number of frames (see Table I). Humans with more than 50% visible parts were considered as valid objects. Correct detection was assumed when the centroid of the detected human was inside of the ground truth bounding box. A one-to-one mapping between detections and ground truth data was enforced.

TABLE I.
HUMAN DETECTION PERFORMANCE FOR A BLOB-BASED AND FOR THE PROPOSED APPROACH

Performance measures	Blob-based approach	Proposed approach
Detection rate	38%	94%
False alarm rate	32%	37%
Mean spatial deviation between detected humans and ground truth	31%	14%
Number of evaluated frames	3990	
Total number of valid humans in the ground truth	24883	

Tracking by mode seeking achieves a high detection rate of 94%, while generating a false alarm rate of 37% (see Table I). The high detection rate is due to the model-driven clustering and occlusion handling providing accurate locations for humans even during occlusions. The high false alarm rate is generated by the permanent motion clutter caused by flag and vegetation movements. Clusters of these moving regions are not distinguishable from humans by the proposed method.

The blob-based detection approach produces poor results for the test sequence given the frequent occurrence of undersegmented groups and humans with shadows. The amount of false alarms is slightly lower, since large connected moving regions count as a single detected blob, whereas the model-based approach explains them as a group of objects.

The mean spatial deviation (see Table I) of detection results was evaluated by computing the distance in the image space between the ground truth centroids and the centroids of matching detections. This distance is normalized by the local height model *H(x)*. As it can be seen from the table, blob detection locations are highly inaccurate, since detections are off by ca. 30% of the human height. The large amount of spatial errors is again due to undersegmented objects. Mean shift based detection produces smaller errors implying that detected objects coincide well spatially with ground truth objects.

In order to evaluate the tracking performance, a particular ground truth trajectory undergoing several occlusions was selected. This ground truth was compared to the trajectories generated by the blob-based and proposed tracking scheme. Trajectories obtained for the two different tracking algorithms are shown in Fig. 5 together with the ground truth trajectory.

Errors in terms of the spatial distance measured in pixels with respect to the ground truth trajectory are shown in Fig. 6 for the blob-based and for the proposed tracking approaches. The target tracked by mean shift mode seeking remains close to the ground truth target position for the entire track duration. The trajectory obtained by the blob-based approach, however, deviates significantly from the ground truth trajectory. In this case, the tracked target is defined most of the time by a group

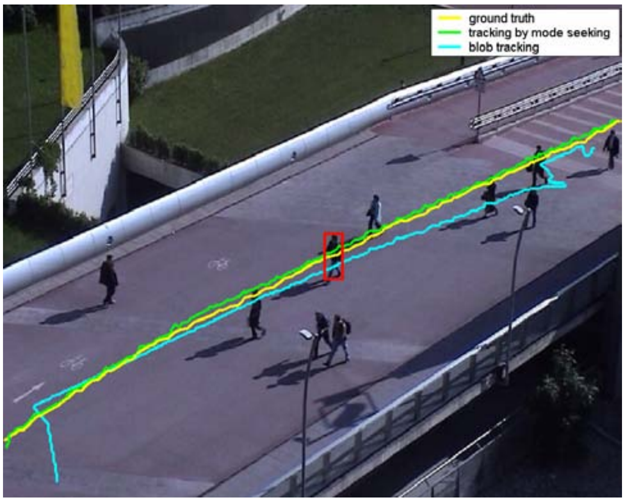

Figure 5. Trajectories illustrating the ground truth trajectory and the trajectories obtained by blob-tracking and by mode-seeking for a particular human (marked by a rectangle).

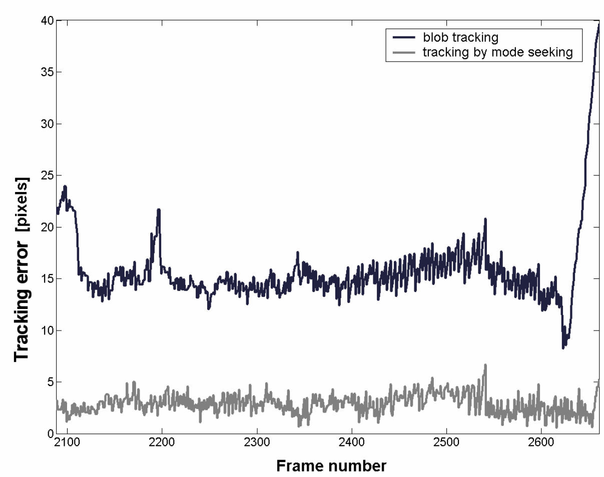

Figure 6. The spatial tracking error (with respect to a manually-determined ground truth trajectory) for blob-based tracking (gray line) and for tracking using mode seeking (dark line). The trajectories are shown in Fig. 5.

of people leading to a permanent offset in the centroid position of the target.

Tracking results for another image sequence (resolution: 360×288 pixels) are shown in Fig.7. This scene contains occasional occlusions between scene objects and humans. The proposed tracking approach can not cope with short-term occlusions. In such cases trajectories terminate upon occlusion and reinitialize after occlusion.

The proposed tracking approach runs in real-time (8-12 fps) on a 2.5 GHz PC for all of the presented sequences.

VIII. Conclusions

This paper presents a novel approach to detect and track humans in real-time in crowded scenes based on a fast variant of the mean shift procedure. We demonstrate how mean shift-based clustering - relying on a kernel of predefined size - can efficiently delineate objects corresponding to local density maxima in the difference image. Furthermore we show, how detected mode candidates can be tracked using the mode seeking property of the mean shift algorithm. A computationally efficient strategy to locate occluding humans is presented. Stable tracking results in a challenging scene depicting frequent occlusions are achieved showing significantly improved results when compared to a blob-based human detector.

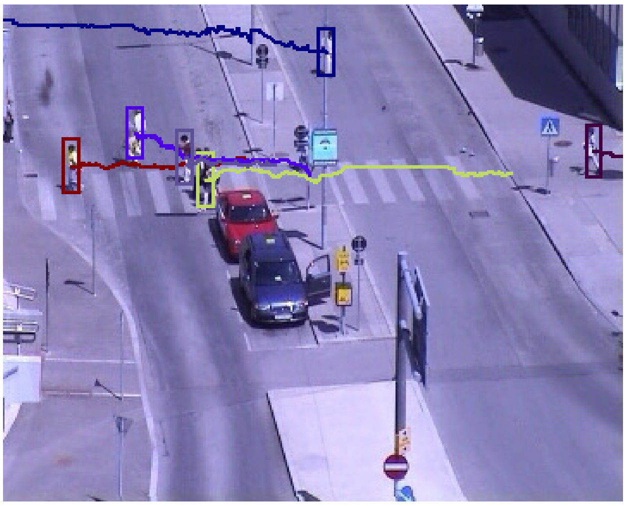

Figure 7. Example frame showing tracking results for another sequence, where occasional short-term occlusions by scene objects occur.

Acknowledgment

This work has been carried out within the Kplus Competence Center Advanced Computer Vision. This work was funded from the K plus Program.

References

[1] I. Haritaoglu, D. Harwood and L. S. Davis, “W4: Real-Time Surveillance of People and Their Activities,” *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(8), pp. 809-830, 2000.

[2] Y. Kuno, T. Watanabe, Y. Shimosakoda and S. Nakagawa, “Automated Detection of Human for Visual Surveillance System,” *Int. Conf. on Pattern Recognition*, C92.2, August 1996.

[3] T. Zhao and R. Nevatia, “Bayesian Human Segmentation in Crowded Situations,” *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 459-466, June 2003.

[4] A. W. Senior, “Tracking with Probabilistic Appearance Models,” *ECCV Workshop on Performance Evaluation of Tracking and Surveillance Systems*, pp. 48-55, June 2002.

[5] T. Yang, Q. Pan and J. Li, “Real-time multiple objects tracking with occlusion handling in dynamic scenes,” *IEEE Conference on Computer Vision and Pattern Recognition*, Vol. 1, pp. 970-975, June 2005.

[6] A. Elgammal, R. Duraiswami and L. S. Davis, “Efficient nonparametric adaptive color modeling using fast gauss transform,” *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 563–570, December 2001.

[7] A. E. C. Pece, “Tracking by Cluster Analysis of Image Differences,” *Proc. 8th Int. Symposium on Intelligent Robotic Systems*, July 2000.

[8] D. Comaniciu, V. Ramesh and P. Meer, “Kernel-Based Object Tracking,” *IEEE Trans. Pattern Anal. Mach. Intell.*, 25(5), pp. 564-575, 2003.

[9] S. Maskell, M. Rollason, N. Gordon and D. Salmond, “Efficient Multitarget Tracking using Particle Filters,” *Journal Image and Vision Computing*, 21(10), pp. 931-939, September 2003.

[10] 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*, May 2004.

[11] Y. Cai, N. de Freitas and J.J. Little, „Robust Visual Tracking for Multiple Targets,“ *In European Conference on Computer Vision*, May 2006.

[12] D. Comaniciu and P. Meer, “Mean Shift Analysis and Applications,” *IEEE Int. Conf. Computer Vision*, pp. 1197-1203, 1999.

[13] C. Beleznai, B. Frühstück and H. Bischof, “Tracking Multiple Humans using Fast Mean Shift Mode Seeking,” *IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, pp. 25-32, January 2005.

[14] P. Viola and M. Jones, “Rapid Object Detection using a Boosted Cascade of Simple Features,” *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, vol. 1, pp. 511-518, 2001.

[15] C. Beleznai, B. Frühstück, H. Bischof and W. Kropatsch, „Detecting Humans in Groups using a Fast Mean Shift Procedure,“ *Proc. of the 28th Workshop of the Austrian Association for Pattern Recognition*, pp. 71-78, 2004.

[16] D. Comaniciu and P. Meer, “Mean Shift: A Robust Approach Toward Feature Space Analysis,” *IEEE Trans. Pattern Anal. Mach. Intell.*, 24(5), pp. 603-619, 2002.

[17] 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,” *Technical Report CMU-RI-TR-00-12*, Robotics Institute, Carnegie Mellon University, 2000.

Csaba Beleznai graduated from the Technical University of Ilmenau in 1994. He received his Ph.D. degree in physics from the Claude Bernard University, Lyon in 1999.

C. Beleznai joined the K+ Competence Center “Advanced Computer Vision” in 2000 as a research scientist. Since 2005 he is Key-researcher at the competence center responsible for research activities in the area of “Surveillance and tracking”. His research interests include surveillance, visual object recognition and statistical methods in computer vision.

Bernhard Frühstück received his M.S. degree in telematics from the Graz University of Technology in 2000.

B. Frühstück joined Siemens PSE, Graz in 2000 as a research staff member working in the field of biometrics and image processing. His primary interests include image processing related to biometric and industrial applications.

Horst Bischof received his M.S. and Ph.D. degree in computer science from the Vienna University of Technology in 1990 and 1993, respectively. In 1998 he got his Habilitation (venia docendi) for applied computer science. Currently he is Professor for Computer Vision at the Institute for Computer Graphics and Vision at Graz University of Technology, Austria.

H. Bischof is Key-researcher at the K+ Competence Center “Advanced Computer Vision” where he is responsible for research projects in the area on “Statistical methods and learning”. He is member of the scientific board of the K+ centers VrVis (Virtual reality and visualization) and Know (Knowledge management). He is vice-president of the Austrian Association for Pattern Recognition. The research interests include, learning and adaptive methods for computer vision, object recognition, surveillance, robotics and medical vision, where Horst Bischof has published more than 260 reviewed scientific papers.

Horst Bischof is program co-chair of ECCV 2006 to be held in Graz. He was co-chairman of international conferences (ICANN 2001, DAGM 1994), and local organizer for ICPR 1996. Currently he is Associate Editor for the journals Pattern Recognition, and Computer and Informatics.