TRACKING MOVING OBJECTS IN VIDEO USING ENHANCED MEAN SHIFT AND REGION-BASED MOTION FIELD

Tiesheng Wang [a], Irene Y.H. Gu [b], Mats Viberg [b], Zhongping Cao [b], Nuan Song [b]

[a] Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, China
[b] Department of Signals and Systems, Chalmers University of Technology, Sweden
tieshengw@sjtu.edu.cn, {irenegu, viberg}@chalmers.se

ABSTRACT

In this paper, we propose a scheme for moving object tracking from videos by combining mean shift and motion field statistics. For mean shift, we employ an enhanced spatial-range mean shift that enables a reduced number of over-segmentation. For motion statistics, we combine the optical flow and high-order moment to generate motion regions that are associated with moving objects (or object parts). Experiments have been conducted on several indoor and outdoor (color/gray-scale) image sequences ranging from simple to median complexity. To evaluate the performance, three objective criteria are applied in addition to the visual inspection. The results show that the proposed method is promising for moving object tracking in video, with an averaging detection rate of 95%. Further, the proposed scheme is compared with that using the conventional mean shift for the tracking, indicating a significantly reduction in false alarm ($\approx 30\%$).

1. INTRODUCTION

There has been an increasing interest in video object detection and tracking due to, for example, multimedia applications, MPEG video coding, virtual reality and video surveillance [1, 2]. Many methods have been developed for moving object tracking, for example, Mixture of Gaussians (MoG) [9] and Bayesian-based dynamic background modeling [2]. In the MoG, colors from a pixel in the background are described by multiple Gaussian distributions. This approach works relatively well for detecting foreground objects in simple or moderate background scenarios. For complex background, a Bayesian-based scheme, which includes dynamic learning and maintenance of complex background, is more robust. Mean shift is a statistical-based method seeking local modes from the kernel-based probability density estimates, and has been shown to be very robust in image segmentation. For example, [4] proposed a video segmentation scheme by employing a mean shift filter using 7D feature vectors including color, time, motion and position-related features. Each video volume is then considered as a collection of three feature vectors, and clustered to obtain a consistent moving object by using a hierarchical mean shift filter. Further extension was proposed in [5] where anisotropic mean shift kernels and spatio-temporal consistent motion are applied. In [8], mean shift was applied for blob-tracking of moving targets. The method includes estimating the color histogram for the target and Bhattacharyya coefficients for the candidate, followed by computing the distance between their distributions. However, the method only offers blob-tracking without segmentation of moving targets.

Motivated by the above, we introduce a mean shift-based moving object tracking scheme that offers both tracking and segmentation. In the proposed scheme, the enhanced mean shift, previously applied to 2D images for reducing over-segmentation [7], is now extended to the application of moving object tracking in videos through combining region-based motion fields.

2. SYSTEM DESCRIPTION

The proposed system, as shown by the block diagram in Fig.1, consists of an enhanced version of spatial-range mean shift filter for 2D image segmentation (Section 3), a candidate motion region detection scheme by combining segmentation results, the optical flow, and the 4th order moment of temporal variations (Section 4). Since both the optical flow and high-order moment provide useful but partial information on pixel motion statistics, motion region is estimated by fusing these two pieces of pixel-based information. Finally, tracking of foreground moving objects is performed by matching modes and optical flow direction of the detected object regions (Section 4). The proposed system is limited to tracking foreground moving objects that contain some level of global motion (i.e., excluding the cases for detecting partially moving objects). Further, videos are assumed to be captured by a non-moving camera.

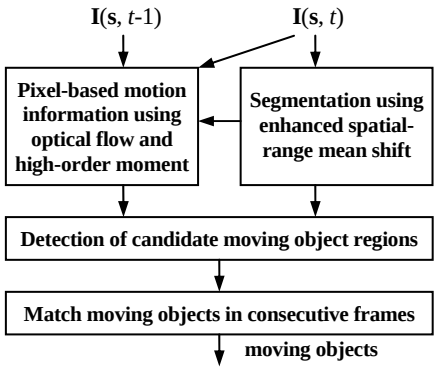

Fig.1 Block diagram of the proposed scheme for tracking of (global) moving objects from image sequences.

3. SEGMENT VIDEO FRAMES USING ENHANCED SPATIAL-RANGE MEAN SHIFT

In this section, we briefly describe the enhanced mean shift previously proposed for 2D image segmentation [7], which is employed in this paper for moving object tracking purpose. The basic idea behind using an enhanced mean shift instead of a conventional mean shift is that less over-segmentation may lead to more accurate detection of motion regions. This is because the statistics of motion are estimated based on pixel-level processing. Further, motion in the image is most likely to be observed around the object boundaries along the motion direction. Reducing over-segmentation may mitigate the problem when converting pixel-based motion information to region-based one. Consequently, it may lead to a better tracking of moving objects.

Image segmentation using mean shift [3] can be considered as clustering pixels having similar local modes (maxima) from the kernel-based pdf estimates. Let the kernel density for a random vector $\mathbf{x}$ be estimated from a set of L-dimensional feature vector $S=\{\mathbf{x}_i, i=1,2,\ldots,n\}$as,

$$\hat{p}_K(\mathbf{x}) = \frac{c_k}{nh^L}\sum_{i=1}^{n} k\left(\left\|\frac{\mathbf{x}-\mathbf{x}_i}{h}\right\|^2\right) \qquad (1)$$

where $K(\mathbf{x})=c_k k(\|\mathbf{x}\|^2)$ is a radially symmetric kernel with L_2 distance measure, c_k is a normalization constant. The local modes of $\hat{p}_k(\mathbf{x})$ can be obtained by setting $\nabla\hat{p}_k(\mathbf{x})=0$, leading to $\frac{1}{2}h^2 c\frac{\nabla\hat{p}_k(\mathbf{x})}{\hat{p}_G(\mathbf{x})} = m_G(\mathbf{x})$, where K is the shadow kernel of kernel G, $G(\mathbf{x}) = c_g g(\|\mathbf{x}\|^2)$, $g(x) = -k'(x)$, $c=c_g/c_k$ is a constant, $m_G(\mathbf{x})$ is the mean shift defined as:

$$m_G(\mathbf{x}) = \sum_{i=1}^{n}\mathbf{x}_i g\left(\left\|\frac{\mathbf{x}-\mathbf{x}_i}{h}\right\|^2\right)\Bigg/\sum_{i=1}^{n} g\left(\left\|\frac{\mathbf{x}-\mathbf{x}_i}{h}\right\|^2\right) - \mathbf{x} \qquad (2)$$

A spatial-range mean shift filter can be considered as a nonlinear edge-preserving smoothing filter: when the differences of pixel intensities are small, the mean shift filter acts as a lowpass filter in a local image region. However, if the intensity differences are large (e.g. around edges), no filtering is applied to these pixels. In such a way, a joint spatial-range mean shift filter takes into account both the geometrical closeness and the photometric similarity in an image.

If one selects the feature vector as $\mathbf{x}=[\mathbf{x}^d\ \mathbf{x}^r]^T$ containing both the *domain feature* $\mathbf{x}^d=[s_x\ s_y]^T$ as pixel spatial positions, and the *range feature* $\mathbf{x}^r=\mathbf{I}(\mathbf{x}^d)$ as a function of domain feature (e.g., image intensity), it is easy to show that for a Gaussian kernel G, or its profile $g(\mathbf{x})$ in (2) becomes,

$$g\left(\|\mathbf{x}\|^2\right) = \exp\left(-\frac{\|\mathbf{x}^d\|^2}{2\sigma_d^2}\right)\exp\left(-\frac{\|\mathbf{I}(\mathbf{x}^d)\|^2}{2\sigma_r^2}\right) \qquad (3)$$

Or, $g(\|\mathbf{x}\|^2) = g_d(\|\mathbf{x}^d\|^2)g_r(\|\mathbf{I}(\mathbf{x}^d)\|^2)$, $h_d=2\sigma_d^2$, $h_r=2\sigma_r^2$ are the domain bandwidth and range kernel bandwidth, respectively. Under this choice, the mean shift filter is working in a joint spatial-range domain. A spatial-range mean shift filter is known to be closely related to a bilateral filter or nonlinear image diffusion [6].

The essence of enhanced spatial-range mean shift is the edge-guided merging, leading to a reduced image over-segmentation. In order to do so, a convergence image map $\mathbf{M}(\mathbf{s},t)$ is created as a by-product during the mean shift. For a given position $\mathbf{s}$ in the map, the value of $\mathbf{M}(\mathbf{s},t)$ is the convergence number where the mean shift initiated from any pixels have finally converged to this pixel position. This map is found to contain rich information on whether a pixel is associated with an image edge, a homogeneous region, or immediate neighbourhood to an edge. For example, $\mathbf{M}(\mathbf{s},t)=0$ means the number of pixels that converged to $\mathbf{s}$ is zero, which is associated with image edges; $\mathbf{M}(\mathbf{s},t)=1$ means that only one pixel converged to $\mathbf{s}$, implying that $\mathbf{s}$ is within a homogenous area; while $\mathbf{M}(\mathbf{s},t)=i(i>1)$ means that i pixels converged to $\mathbf{s}$, indicating an immediate area next to edges. Fig.2 shows an example of the map.

Fig.2 The map image $\mathbf{M}(\mathbf{s},t)$ (right) for an original image (left). The map image contains the number that the mean shift converged to each pixel position. (*black*: zero number of convergence, indicating pixels on image edges; *white*: high number of convergence, indicating pixels next to edges; *gray*: median number of convergence, indicating pixels in smooth areas).

The enhanced mean shift contains a refined segmentation step by merging over-segmented regions, yet, avoiding merging any two regions which may across edges as indicated by the map. That is equivalent to defining pixels with high convergence numbers immediately next to the edges as the natural region boundaries. For each video frame $\mathbf{I}(\mathbf{s},t)$, the enhanced spatial-range mean shift results in a set of segmented regions $R_{i,t}$, $i=1,2,\ldots N$, and their modes.

4. DETECTING MOTION REGIONS FROM PIXEL-BASED MOTION INFORMATION

Based on the segmented regions from the enhanced spatial-range mean shift, motion information associated with each segmented region is then estimated. The basic idea is to first extract the motion information for each pixel, followed by converting the pixel-level information in each segmented region to the region-based one, which will be detailed below.

4.1 Optical Flow-Based Candidate Region Detection

Optical flow is a useful tool for pixel-wise estimation of image velocity field, assuming that a given image $\mathbf{I}(s_x,s_y,t)$ is smoothing function, or the neighbouring pixels in the image have similar brightness. Further, a constraint is imposed to image motion by assuming that image intensity remains constant along the motion trajectory, i.e.,

$$\mathbf{I}_{s_x}u + \mathbf{I}_{s_y}v + \mathbf{I}_t = 0 \qquad (4)$$

where $\mathbf{v}=(u,v)=(ds_x/dt, ds_y/dt)$ is the optical flow vector, $\mathbf{I}_t = d\mathbf{I}/dt$, $\mathbf{I}_{s_x} = \partial\mathbf{I}/\partial s_x$, $\mathbf{I}_{s_y} = \partial\mathbf{I}/\partial s_y$. In our system, the

optical flow vector (u,v) is computed using the Gauss-Seidel iteration to a pair of consecutive image pixels.

One may observe that along the boundaries of a moving object, optical flow vectors point to the directions of motion, and their magnitudes are proportional to the motion velocities. Fig.3 shows an example of the optical flow field from 2 consecutive frames in the "car parking" sequence, where large optical flow values are shown on the object boundaries that are perpendicular to the motion direction. We define a segmented region as a candidate motion region if a certain percentage (T_1) of pixels within the region whose optical flow magnitudes exceed a pre-selected threshold ε_1 .

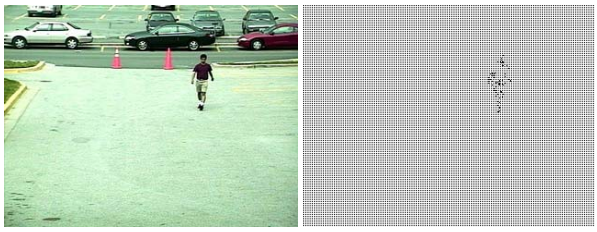

Fig.3. The optical flow field for an image from 'car parking' video frame #66. *Left*: the original image; *Right*: the optical flow field.

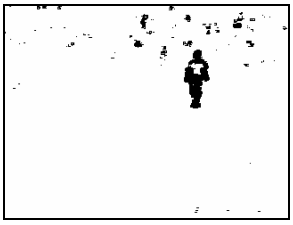

Fig.4. Detected changes based on the 4th order moments of temporal variations, from two consecutive images in the video 'car parking'.

It is worth mentioning that, since the optical flow values are small in the interior object area as well as some parts of the boundaries, this approach alone is often not sufficient for detecting regions in an entire moving object. It appears that some interior object regions may not be included as the candidate motion regions since an object is usually segmented into a few regions. Further, the detected regions may include undesirable regions corresponding to the shadow of objects.

4.2 Moment-based Candidate Region Detection

Another way to estimate motion statistics is through detecting changes across two consecutive image frames. Pixel-based changes can be estimated by computing the 4th order sample moments of temporal variations in consecutive image frames [10] as follows,

$$\hat{\mathbf{m}}(\mathbf{s},t)=\frac{1}{(2l+1)^2}\sum_{i,j=-l}^{l}\left(\mathbf{d}(s_x+i,s_y+j,t)-\bar{\mathbf{d}}(s_x,s_y,t)\right)^4 \qquad (5)$$

where $\mathbf{d}(s_x,s_y,t)=\mathbf{I}(s_x,s_y,t)-\mathbf{I}(s_x,s_y,t\text{-}1)$ is the pixel-wise difference of image values in two consecutive frames, and $\bar{\mathbf{d}}(s_x,s_y)$ is the average value within a $(2l+1)\times(2l+1)$ window centered at (s_x,s_y), (3×3window was used in our tests).

The above 4th order moments can be used to detect image changes along the temporal direction. Small image intensity variations (e.g. due to illumination changes) can be further differentiated from the relatively large object motions. This pixel-based change information is then converted to region-based motion detection by employing a simple voting method: for each segmented region, if a certain percentage (T_2) of pixels within the region whose 4th order moment values exceed a pre-specified threshold ε_2 , the region is defined as a candidate motion region.

It is worth noting that even though the high order moment is sensitive to image noise, the false alarm of detecting a candidate region due to image noise is relatively small, since a large percentage of pixels within a region that are coincidently affected by the noise is relatively small. However, it is also worth mentioning that using this approach alone can result in many false alarm regions that are not associated with the moving objects. Fig.4 shows an example where the detected candidate regions from the 4th order moment also include static background regions whose intensities changed through image frames. This is because the detected changes can be caused by different reasons, such as: (a) intensity changes caused by a globally moving object; (b) intensity changes of a static object/background; and (c) intensity changes caused by a local movement in a small part of a large static object (e.g. eye and lip movement). For tracking global moving objects, we are only interested in the case (a). It is also worth mentioning that carefully selecting the threshold can often reduce the shadow areas of objects.

4.3 Fusing Candidate Regions

As mentioned in Sections 4.1 and 4.2, each of the above two methods is shown to yield useful but only partial information on the foreground moving objects. Therefore, the candidate regions detected from these two methods are then fused. If a segmented region is both selected by the optical flow and the 4th order moment as a candidate motion region, then the region is finally selected as a motion region. Fig.5 shows an example of the region-based motion field fusion process.

Fig.5 Example of detected moving object for the 'car parking' sequences (frame #66-67). *Left*: optical flow-based candidate regions; *Middle*: the 4th order moment-based candidate regions. *Right*: detected moving regions after fusion.

4.4 Linking Objects through Frames

Finally, detected moving objects (or regions) in consecutive frames are linked by matching their modes as well as their optical flow directions. If two objects have similar modes and are consistent with the directions of optical flow, then these two objects are linked through image frames and assigned as a tracked object.

5. EXPERIMENTAL RESULTS

The proposed scheme has been tested for a range of indoor and outdoor image sequences. Fig.6 shows several randomly selected frames of tracked moving objects from 3 outdoor

videos 'car parking', 'rain' , 'running' and one indoor video 'hall' by using the proposed scheme. Table 1 includes the parameters used for the tests, where h_d and h_r were band-widths for mean shift filtering, h_{r2} for enhanced segmentation, and $\varepsilon_1 = \varepsilon_2 = 10^{-5}$ were used in all tests.

video	h_d	h_r	h_{r2}	T_1	T_2
hall	5	0.08	0.12	0.25	0.1
rain	4	0.05	0.10	0.25	0.1
car park	5	0.12	0.15	0.25	0.1
running	5	0.14	0.18	0.25	0.1

Table 1. Parameters used for tracking moving objects from videos in Fig.5 using the proposed scheme.

From the tracking results, one can observe that moving objects are correctly detected and tracked in most cases, and that the detected regions from motion field fusion have removed most false alarm regions for the detection. One can observe that occasionally there are some small holes in the interior of objects. This is probably due to over-segmentation. We also observed that sometimes the detected areas are bigger than actual objects. For example, the areas may include the shadows of a walking person or a moving car in addition to the desired object itself, or include extra areas due to errors caused from 2D segmentation. We also observed that sometimes fusion resulted in less shadow areas for a moving object. This is because the 4th order moment-based method often excludes these regions despite that the optical flow-based method often picks up the shadow regions. If an object is moving towards/away from the camera, the method does work properly. Overall, the proposed method is shown to yield reasonably good tracking results.

Performance Evaluation: To further evaluate the tracking performance, three objective criteria were applied. Criterion-1 is based on the detection rate P_D and the false alarm rate P_F for each image frame. P_D is defined as the pixels from the tracked moving object(s) that fall within the true moving object area (i.e. the ground truth of moving object), while P_F is defined as the pixels in the tracked moving object(s), however should belong to the background according to the ground truth. Criterion-2 is a similarity measure defined as $S(A,B)=(A \cap B)/(A \cup B)$, where A is the detected object regions, B the manually marked ground truth, and the range of S is between 0 and 1. The larger the value of S (i.e., close to 1.0), the more similar the detected regions compared with the ground truth. Criterion-3 is a distortion measure defined as $d(A,B) = \dfrac{\sum_{(s_x,s_y)} A(s_x,s_y) \oplus B(s_x,s_y)}{\sum_{(s_x,s_y)} B(s_x,s_y)}$, where $\oplus$ denotes the 'XOR' operator, pixels in A and B are set to be binary values. The smaller the distortion value (close to 0.0), the more accurate the detected objects as compared with the ground truth. The ground truths used for our evaluations (50 frames for the video 'running' and 31 frames for the video 'rain') were manually labeled. Table 2 includes the results from the above 3 criteria. From the table, one can observe that the averaging detection rate is rather satisfactory. The averaging false alarm rate is less satisfactory especially for the video 'raining'. A further visual inspection to the tracking results indicates that this was mainly caused by including shadow as part of the moving objects.

Comparison: The tracking results from the enhanced mean shift-based method were then compared with those from the corresponding conventional mean shift-based tracking. The results from the conventional mean shift-based tracking are also included in Table 2. Comparing the results from the two methods, it shows that the enhanced mean shift-based object tracking produces better performance than the conventional mean shift-based tracking. Further, the false alarm rate is significantly reduced (approximately 30%).

video	method	$\bar{P}_D$ (%)	$\bar{P}_F$ (%)	$\bar{S}$	$\bar{d}$
running	proposed	95.50	12.24	0.88	0.17
	MS-based	95.56	17.59	0.81	0.22
rain	Proposed	94.38	17.72	0.81	0.23
	MS-based	93.85	26.23	0.76	0.32

Table 2. Evaluation results for the proposed scheme and from the conventional mean shift (MS)-based tracking, using 3 objective criteria. Where: $\bar{P}_D$: averaging detection rate; $\bar{P}_F$: averaging false alarm rate; $\bar{S}$: averaging similarity measure; and $\bar{d}$: averaging distortion measure. The averaging was applied to 50 frames of the 'running' video and 31 frames of the 'rain' video.

6. CONCLUSIONS

The proposed scheme, which combines the enhanced mean shift with estimated motion fields, has been tested for tracking moving objects in videos. The test results have shown that each of the two motion detection methods has provided useful but partial information on foreground moving objects. The fusion of these two pieces of information has significantly improved the final results both in terms of tracked object areas and of reduced false alarm on the detected objects. Visual inspection and evaluation have shown that the proposed scheme is effective for moving object tracking from videos. Comparisons have shown that employing the enhanced mean shift for tracking has significantly reduced the false alarm (approximately 30%), and hence mitigated the tracking errors. Overall, the proposed scheme has yielded improved performance over the conventional mean shift-based tracking. Comparing with the mean shift-based blob-tracking in [8], the proposed scheme is able to provide simultaneous moving object tracking and segmentation.

ACKNOWLEDGEMENTS

This work was partly supported by Asia-Swedish Research Links Program in Sweden under Sida/VR grant number 348-2005-6095.

REFERENCES

[1] I.Koprinska, S.Carrato,"Temporal video segmentation: a survey", signal processing, image communication, vol.16, pp477-500, 2001.
[2] L.Li, W.Huang, I.Y.H.Gu, Q.Tian, "Statistical Modeling of Complex Backgrounds for Foreground Object Detection", IEEE Trans. Image Processing, vol. 13, No.11, pp.1459-1472, 2004.

[3] D. Comaniciu, P.Meer, "Mean shift: a robust approach toward feature space analysis". IEEE Trans. PAMI, 603-619, 2002.
[4] D.DeMenthon, R.Megret, "Spatial-temporal segmentation of video by hierarchical mean shift analysis", in Proc. of Statistical methods in video processing workshop(SMVP'02), Denmark, 2002.
[5] J.Wang, B.Thiesson, Y.Xu, M.Cohen,"Image and video segmentation by anisotropic kernel mean shift", in Proc. ECCV, 2004.
[6] D. Barash, "A Fundamental Relationship between Bilateral Filtering, Adaptive Smoothing and the Nonlinear Diffusion Equation", IEEE Trans. PAMI, vol.24, No.6, pp.844-847, 2002.
[7] N.Song, I.Y.H.Gu, Z.Cao, M.Viberg, "Enhanced spatial-range mean shift color image segmentation by using convergence frequency and position", in Proc. of EUSIPCO 2006, Florence, Italy.
[8] D.Comaniciu, et. al, "Real-Time Tracking of Non-Rigid Objects using Mean Shift", in Proc. IEEE Conf. CVPR, pp.142-149, 2000.
[9] C. Stauffer, W. Grimson, "Learning patterns of activity using real-time tracking", IEEE Trans. PAMI, 22(8), pp747–757, 2000.
[10] A.Neri, et.al, "Automatic moving object and background separation", Signal Processing, vol. 66, pp. 219-232, 1998.

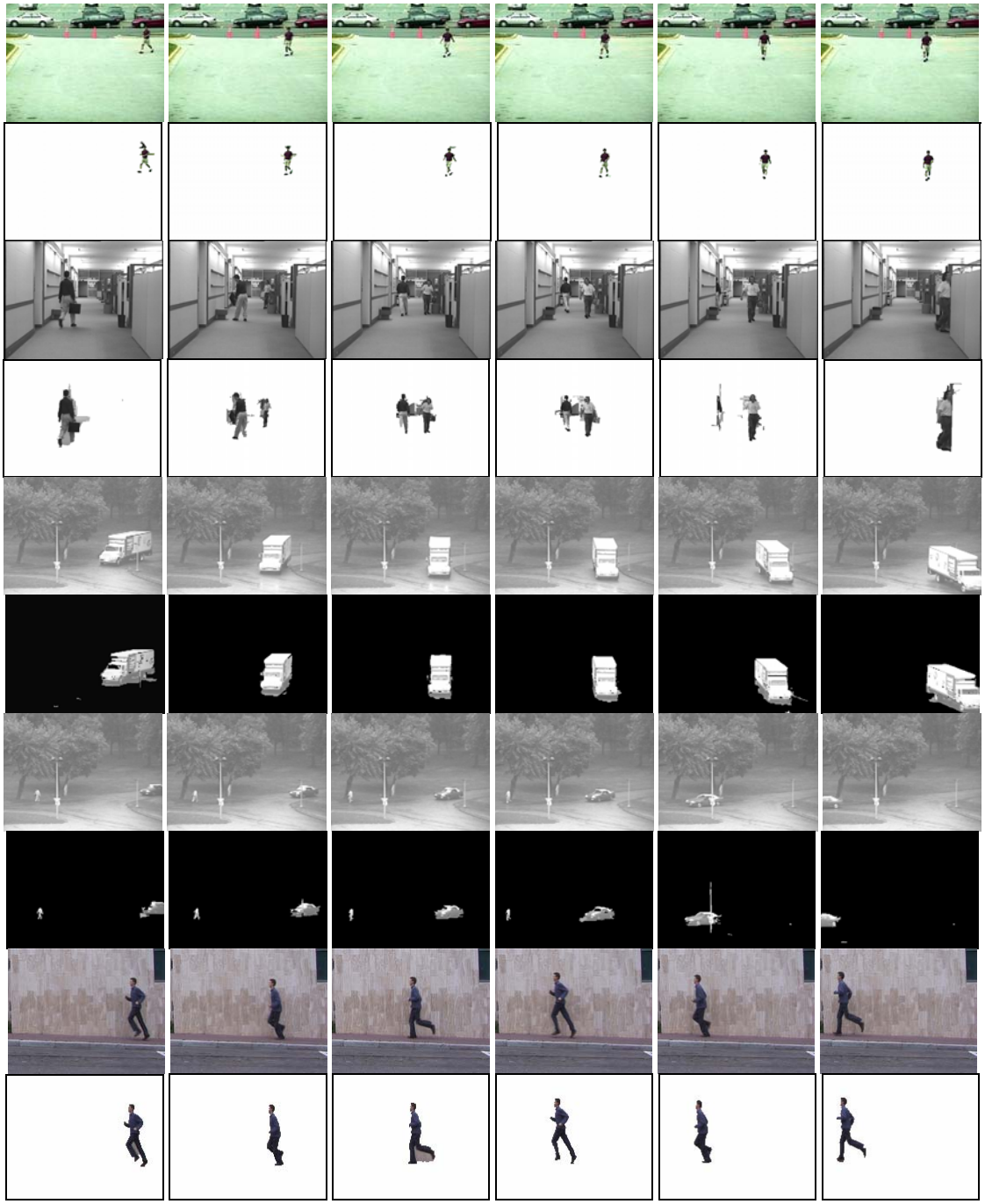

Fig.6. Results of moving object tracking from videos using the proposed scheme. Rows 1-2: video 'car parking': Images and tracked objects in frames # 97,120,126,136,145,156; Rows 3-4: video 'hall': original images and tracked objects in #17,31,60,72,82,103; Rows 5-8: video 'rain': original images and tracked objects in # 91,97,99,100,102, 106, 198, 200, 203, 205, 211, 214; Rows 9-10: video 'running': original images and tracked objects in #6,15,23,30,38,45.