Moving Object Localization in Thermal Imagery by Forward-backward MHI

Zhaozheng Yin and Robert Collins
Department of Computer Science and Engineering
The Pennsylvania State University, University Park, PA 16802
{zyin, rcollins}@cse.psu.edu

Abstract

Detecting moving objects automatically is a key component of an automatic visual surveillance and tracking system. In airborne thermal video, the moving objects may be small, color information is not available, and even intensity appearance may be camouflaged. Previous motion-based moving object detection approaches often use background subtraction, inter-frame difference or three-frame difference. In this paper, we describe a detection and localization method based on forward-backward motion history images (MHI). This method can accurately detect location and shape of moving objects for initializing a tracker. Using long and varied video sequences, we quantify the effectiveness of this method.

1 Introduction

Detecting moving objects in image sequences is a ubiquitous problem that plays an indispensable role in automatic surveillance and tracking. Detecting and localizing the object accurately is important for automatic tracking system initialization and recovery from tracking failure. For tracker initialization, it is necessary to first localize position and shape of the object and analyze its features. Later, if the tracker fails, the moving object detection module can localize moving objects globally in the image, and the tracking system can then associate the globally detected objects with previously tracked objects to restart the tracker.

When prior knowledge of moving object appearance and shape is not available, change detection or optical flow can still provide powerful motion-based cues for detecting and localizing objects, even when the objects move in a cluttered environment, or are partially occluded. There are three main approaches to pixel level change detection: background subtraction, inter-frame difference and three-frame difference. Background subtraction compares the current frame (Figure 1(b)) with a background image (Figure 1(c)) to locate the moving foreground objects (Figure 1(d)). This method can extract the shape of the object well provided that the static background model is available and adaptive to illumination change. Stauffer and Grimson [8] developed a probabilistic method for background subtraction. The background is adaptively updated by modeling each pixel as a Gaussian mixture model. However in airborne video captured by a moving camera, the stabilized background as shown in Figure 1(c) is costly to reconstruct at every frame.

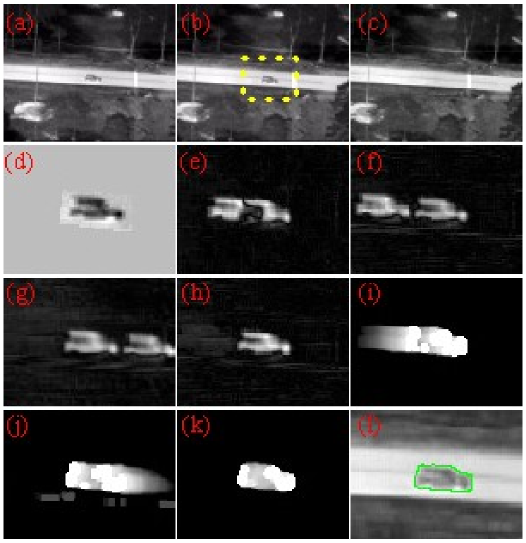

Figure 1: (d)-(l) are subimages of (b). (a) the 76th frame. (b) the 88th frame. (c) reconstructed background. (d) background subtraction between (b) and (c). (e) frame difference between the stabilized 82nd and 88th frames. (f) frame difference between the stabilized 76th and 88th frames. (g) frame difference between the stabilized 100th and 88th frames. (h) three-frame difference. (i) forward MHI. (j) backward MHI. (k) detected mask by forward-backward MHI. (l) detected shape by forward-backward MHI.

Interframe difference methods easily detect motion but do a poor job of localizing the object. If temporal distance between two differencing frames is small, only part of the object is detected (Figure 1(e)). If temporal distance is large, two object locations are detected - one where the object is, and one where it used to be (Figure 1(f)). Motivated by this problem, the three-frame difference approach uses future, current and previous frames to localize the object in the current frame [6]. Irani and Anandan [5] provide a unified approach to moving object detection both in 2D and 3D scenes, in which the object is detected by three-frame difference. Using future frames introduces a lag in the tracking system, but this lag is acceptable if the object is far away from the camera or moves slowly relative to the high capture rate of the camera. Figure 1(g) is the difference between current and future frames. Figure 1(h) gives the logical 'AND' result of Figure 1(f) and Figure 1(g). The location of the object is well detected, but the shape of the object is only coarsely detected.

2 Related work

In the frame difference method, the choice of temporal distance between frames is a tricky question. It depends on the size and speed of the moving object. Furthermore, the background subtraction, inter-frame difference and three-frame difference only tell us where the motion is. Strehl and Aggarwal [9] resort to gray level edges to segment the object based on the detected motion. Paragios and Deriche [7] present a moving object detection and tracking approach by geodesic active contour and level sets. This boundary-based approach uses the motion detection boundary by applying the edge detector on the interframe difference.

In contrast to the above methods, the motion history image (MHI) provides more motion properties, such as direction of motion. Bobick and Davis [1] use MHI as part of a temporal template to represent and recognize human movement. MHI is computed as a scalar-valued image where intensity is a function of recency of motion. An extension to the original MHI framework is to compute normal optical flow (motion flow orthogonal to object boundaries) from MHI by Bradski and Davis [2]. Wixson [10] presented another integration approach which integrates frame-by-frame optical flow over time. The consistency of direction is used as a filter. In the W4 system, Haritaoglu et.al [4] use a change history map to update the background model. Another related work is developed by Halevi and Weinshall [3] to track multi-body non-rigid motion. Their algorithm is based on a disturbance map, which is obtained by linearly subtracting the temporal average of the previous frames from the new frame.

In this paper we proposed a moving object localization approach based on MHI. Similar to the work of Bobick and Davis [1], the motion images generated by inter-frame differencing are combined with a linear decay term. From previous frames to the current frame, we get the forward MHI as shown in Figure 1(i). The trail gradient in the MHI indicates the direction of object motion in the image. Similarly, we construct a backward MHI from the future frames to the current frame as shown in Figure 1(j). Again we assume the lag introduced into the tracking system is acceptable. Combining the two MHIs, we obtained the object mask (Figure 1(k)) and shape (Figure 1(l)). Comparing to the previous approaches, this method does not require adaptive background reconstruction, it provides more motion information than the three-frame difference method, and it can recover the shape of the moving object better. Our approach is also suited for moving cameras, because we do stabilization of adjacent frames in time and propagate locally. This reduces the correspondence errors of stabilization across larger temporal distance. The details of our approach are discussed in Section 3. Section 4 presents the implementation results, which are evaluated on several thermal video sequences. Finally we make a brief conclusion in Section 5.

3 Object localization by MHI

Motion history images combine object movement information over an image sub-sequence. The old object motion, which was obtained from frame difference among images far away from the current instant, fades away due to the decay term. In general, MHI shows the cumulative object motion with a gradient trail. Our MHI based approach has three main modules:

1. *Preprocessing module.* The previous frame at time instant $\tau - \Delta$, $I(\tau - \Delta)$, is stabilized into the coordinate of the frame at time τ, $I(\tau)$. Both of these two frames are normalized before the differencing.

2. *MHI generation module.* The MHI at time τ, $H_F(\tau)$, is a function of the MHI at time $\tau - 1$, $H_F(\tau - 1)$, and the motion image at time τ, $D_F(\tau)$.

3. *Object localization module.* The forward MHI at the current instant t, $H_F(t)$, is computed recursively from previous time instant $t - (L - 1)$ to t. $H_F(t)$ is combined with the backward MHI at the current instant t, $H_B(t)$, to determine the moving object mask in the current image $I(t)$. The backward MHI has the same generation process except that τ is reduced recursively from $t + (L - 1)$ to t.

3.1 Preprocessing

In airborne video, the background is moving over time due to the moving camera. Before using the frame difference to get motion images, we need to stabilize the frames first. If the camera is static, this step can be skipped. Two-frame background motion estimation is achieved by fitting a global parametric motion model (affine or projective) to sparse optic flow. Sparse flow is computed by matching Harris corners between frames using normalized cross correlation. Given a set of potential corner correspondences

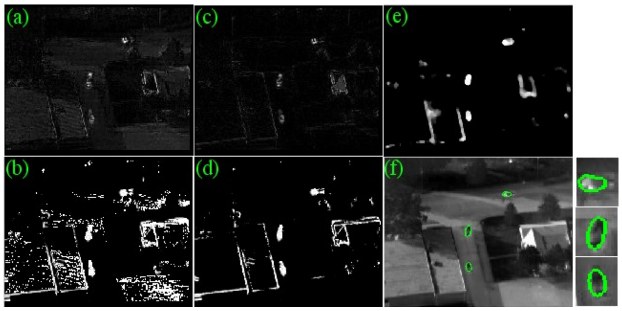

Figure 2: $t = 1030, \tau = 1032, \Delta = 3$ (a) frame difference between the τ and $\tau + \Delta$ frames without normalization. (b) $H_B(\tau)$ without normalization (c) frame difference with normalization. (d) $H_B(\tau)$with normalization (e) combining $H_F(t)$ with $H_B(t)$ (f) detected object contours at current instant t

across two frames, we use a Random Sample Consensus (RANSAC) procedure to robustly estimate global affine flow from observed displacement vectors. The largest set of inliers returned from the RANSAC procedure is then used to fit either a 6 parameter affine or 8 parameter planar projective transformation. Using $P^{\tau}_{\tau-\Delta}$ to represent the affine motion from frame $\tau - \Delta$ to frame τ, we perform the warping as:

$$I^{'}(\tau - \Delta) = P^{\tau}_{\tau-\Delta} \times I(\tau - \Delta) \tag{1}$$

To avoid large corner correspondence error between two frames, we incrementally compute the transformation matrix step by step as Eq (2). In practice we do not choose a large Δ since that will cause big cumulative error even using Eq (2).

$$P^{\tau}_{\tau-\Delta} = P^{\tau}_{\tau-1} \times P^{\tau-1}_{\tau-2} \times \cdots \times P^{\tau-\Delta+1}_{\tau-\Delta} \tag{2}$$

Another notorious problem in airborne video is rapid change in pixel intensities when the camera sensor has automatic gain change control. Especially in thermal videos, when very hot or cold objects appear, the gray value of each pixel changes greatly as the camera rapidly adjusts its gain to avoid saturation. The changing illumination makes the intensity-based frame difference method inadequate for obtaining accurate motion. Yalcin et.al [11] have proposed an intensity-clipped affine model of camera sensor gain. In this paper, we use a simplified normalization method:

$$I^{'}(\tau) = \frac{I(\tau) - \overline{I(\tau)}}{std(I(\tau))} \tag{3}$$

where $\overline{I(\tau)}$ represents the mean value of the image, $std(I(\tau))$ stands for the standard deviation of the image.

After the normalization step, the pixel value can be negative. This will not affect the frame difference result in the next module, and we do not need to scale the pixel value into the range of 0 to 255. In the thermal video of Figure 2, there is a big gain change between the 1032nd frame and the following images. The motion image generated by frame difference will be polluted if there is no normalization (Figure 2(a)). Thus the MHI at time τ will also be degraded (Figure 2(b)). Figure 2(c-d) gives the motion image and MHI with the normalization computed from Eq 3 for comparison. Figure 2(e-f) shows the final localization results with the normalization.

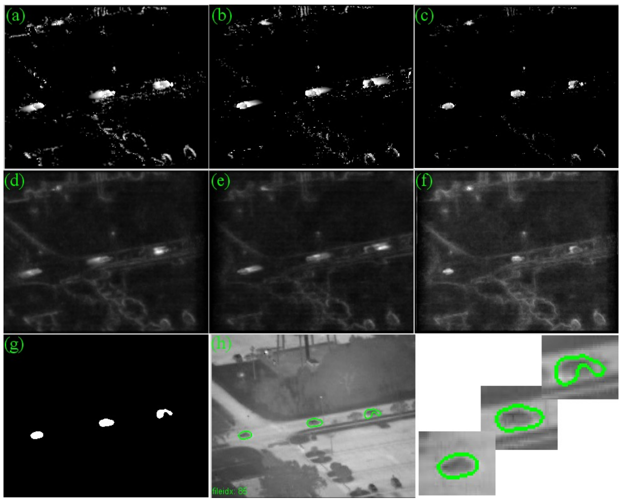

Figure 3: (a) forward MHI. (b) backward MHI. (c) combination of (a) and (b) (d) forward MEI. (e) backward MEI. (f) combination of (d) and (e). (g) post-processed (c). (h) detected object contours, the upper-right vehicle in the scene is partial occluded by a tree

3.2 Motion History Image Generation

A single motion image computed by inter-frame difference shows where motion (change) exists, but noise may also be above threshold. Furthermore it is hard to choose a suitable frame difference distance Δ. One method of integrating motion images over time is the Motion Energy Image (MEI), computed as[1]:

$$E(t) = \sum_{t}^{t \pm (L-1)} D(\tau) \tag{4}$$

where $'-'$ means forward MEI, $'+'$ means backward MEI, L is the length of the time period, $D(\tau)$ is the absolute frame difference with difference distance Δ:

$$D(\tau) = |I(\tau) - I(\tau \pm \Delta)| \tag{5}$$

$I(\tau)$ and $I(\tau \pm \Delta)$ are stabilized and normalized images.

One drawback of MEI is that all the motion caused by the noise will also be accumulated. The MEI is blurred due to the summation of all the noisy motion within the time period. Thus it is hard to distinguish the objects from the background. As shown in Figure 3, there is much more noise ex-

[1] Originally Bobick and Davis [1] uses the logical 'AND' of the binary difference images to compute $E(t)$

isting in the MEI than in the MHI. Instead of only showing all the existing motion during a period of time, MHI keeps a record of how the historic motion evolves with the current motion image. By incorporating a temporal decay term, the forward MHI is computed as[2]:

$$H_F(x,y,\tau) = \begin{cases} max(0, P_{\tau-1}^{\tau} H_F(x,y,\tau-1) - d) & \text{if } D(x,y,\tau) < T \\ 255 & \text{if } D(x,y,\tau) \geq T \end{cases} \tag{6}$$

where $P_{\tau-1}^{\tau}$ is the warping matrix from frame $\tau-1$ to frame τ, d is the decay term and T is a threshold. The pixel value calculated above is within $[0, 255]$, so the decay term d is also defined within $[0, 255]$. For example, we can define $d = 255/L$. Without loss of generality, we can also scale the pixel value into other ranges like $[0, 1]$.

The forward MHI, $H_F(t)$, is a function of the previous forward MHI, $H_F(t-1)$ and current motion image $D_F(t)$. This satisfies the Markovian assumption that no other old motion images need to be stored. Compared to MEI, the recent moving pixels in MHI are highlighted while the old moving pixels in MHI are darker. As a benefit, the impulse noise in the old motion image decays away while the persistent motion generated by the moving object is preserved. Similarly we can compute the backward motion history image $H_B(\tau)$. Figure 4 gives an example of the MHI generation process. The initial backward and forward MHIs are set to zero:

$$\begin{aligned} H_F(t-(L-1)) &= 0 \\ H_B(t+(L-1)) &= 0 \end{aligned} \tag{7}$$

3.3 Object localization

After we get the forward and backward MHI, $H_F(t)$ and $H_B(t)$, we perform median filtering to smooth the MHIs and remove the salt-pepper noise. Alternatively, a Gaussian filter can be used. The forward-backward motion history masks are combined by

$$\text{Mask}(t) = min(medfilt(H_F(t)), medfilt(H_B(t))) \tag{8}$$

where $medfilt$ stands for the median smoothing filter. For objects moving in a constant direction, the 'min' operator in Eq (8) serves to suppress the gradient trail behind the object in the forward MHI, and the gradient trail ahead of the object in the backward MHI, yielding strong response only for pixels within the current object boundary.

Figure 5 provides an example in which some cars move close together and in two directions while four people are running together. The mask generated by Eq 8 is shown

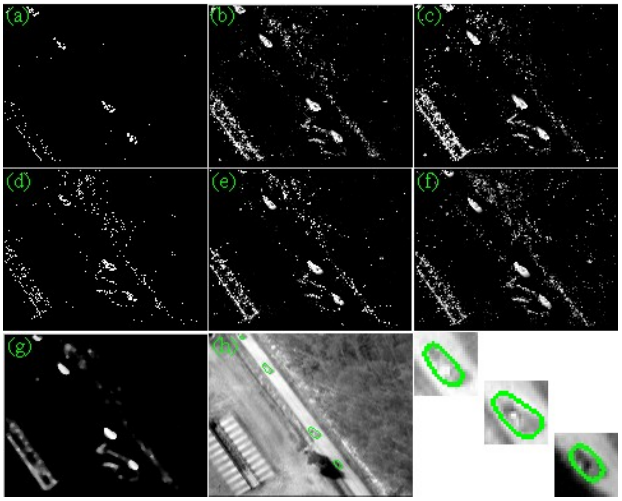

Figure 4: L=11 (a)-(c) forward MHIs at t-10, t-5, t respectively. (d)-(f) backward MHIs at t+10, t+5, t respectively. (g) combination of $H_F(t)$ and $H_B(t)$. (h) detected object contours

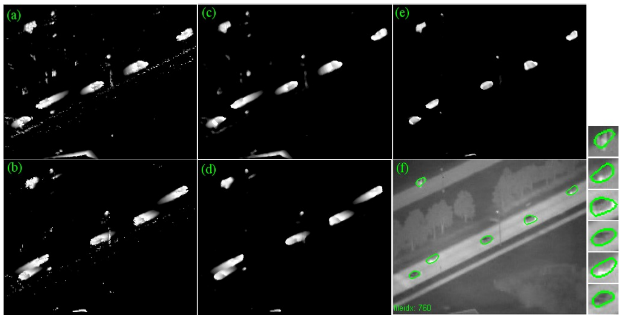

Figure 5: (a) $H_F(t)$. (b) $H_B(t)$. (c) $H_F(t)$ after median filter. (d) $H_B(t)$ after median filter. (e) combined mask (f) detected object contours, the upper-left contour in the scene is composed by four running people.

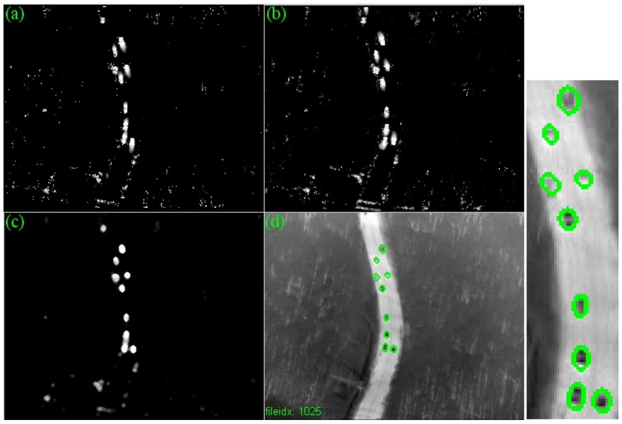

Figure 6: (a) $H_F(t)$. (b) $H_B(t)$. (c) combination without morphological operation (d) detected object contours with morphological operations

[2] As shorthand notation, we ignore the x, y in the parameter list and represent them as $H_F(\tau)$ and $H_B(\tau)$.

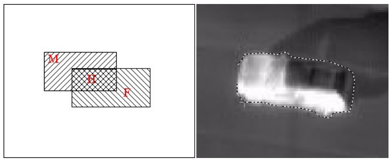

Figure 7: Evaluation process and hand labeled ground truth

in Figure 5(e). After thresholding the mask, the final object contours are shown in Figure 5(f). Furthermore, some morphological operations may be performed to improve the accuracy of the object mask as shown in Figure 6. For example, the close or dilate operations can fill any holes or gaps within the same object; while the open or erosion operations can remove thin bridges of pixels between nearby objects as well as removing small objects caused by noise.

4 Experiment analysis

4.1 Evaluation metrics

Our experiment evaluation design is shown in Figure 7. The ground truth object shape is labeled manually. Let H denote the hit area, i.e. the area belonging to the object and correctly detected, M denote the miss area, i.e. the area belonging to the object but incorrectly missed, and F denote the false alarm area, i.e. the area not belonging to the object but incorrectly detected. The hit rate is defined as

$$\text{HR} = \frac{H}{H+M} \tag{9}$$

The false alarm rate is

$$\text{FAR} = \frac{F}{H+F} \tag{10}$$

Note that miss rate is redundant with hit rate:

$$\text{MR} = \frac{M}{H+M} = 1 - \text{HR} \tag{11}$$

so we will evaluate the detection performance based on the hit rate and false alarm rate only. A perfect detection result would have HR equal to one and FAR equal to zero.

4.2 Effect of L and Δ

To achieve a good detection performance, we need to choose a suitable motion history length L. If the object has uniform intensity, s is the average moving object speed (pixels/second) during the period L, f is the frame rate (frame/second), and l is the object length in the image(pixels), then the smallest motion history length L is constrained by:

$$s\frac{L}{f} \geq l \tag{12}$$

Assuming that the speed and size of the moving object in the scene can be estimated when we set up the tracking system, we can calculate the minimum motion history length as above. Otherwise we can choose a big L conservatively to guarantee that the object shape can be detected well, although big L will lengthen the lag of the tracking system. Figure 8 shows an example in which the hit rate increases with the enlarging L while the false alarm rate decreases.

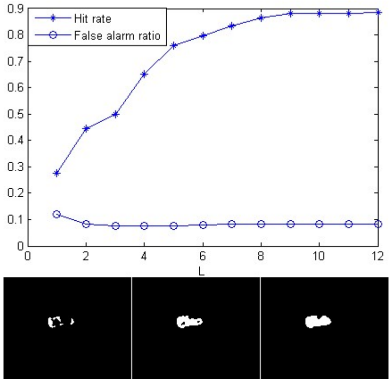

Figure 8: The effect of different history length. The bottom row shows three masks related to $L = 1, 5, 9$ respectively

Another factor that affects the frame difference is the step size Δ. Normally we choose Δ between one and four, and avoid choosing larger values since that will cause more interframe stabilization error and make the MHI noisy. If the object is moving slowly and Δ is small, then only a sliver of the object can be detected at each frame, however all the slivers will be accumulated into the final MHI with a suitable motion history length.

4.3 Experiment result

Figure 9 shows the performance evaluation for four different thermal sequences. We randomly select 20 images from each sequence and label ground truth object shapes by hand. Note that only moving objects are labeled and that the shadow is not considered to be part of the object. If all the objects in the image are static or all the objects are totally occluded, this image is replaced by another randomly chosen image. The four sequences contain trucks and small sedans driving along a road network. The image resolution of objects in the same video sequence may be large or small due to the moving and zooming camera. Different images from the same sequence may contain single or multiple objects. Some objects may be partly occluded, and some images are blurred. Despite these challenges, among all the sampled images, the hit rate is around or above 0.8 and the false alarm rate is around or below 0.4. The detected object shape tends to match the object boundary well. Some exception cases include: (1) the object slows down when going around a corner so that the motion is not obvious (the 1162nd frame of sequence 1, the 3037th and 4013th frames of sequence 2), which degrades the performance. (2) part of the object is just coming into the image, which can not be detected well (the 7537th frame of sequence 3 and the 79th frame of sequence 4). (3) the object has uniform appearance that is similar to the background. For example, the trunk of the truck in sequence 4 is dark and similar to the pavement; it is not segmented from the background perfectly.

To demonstrate that this approach generalizes to other kinds of scenes, we also tested it on five challenging non-thermal sequences. 40 images are chosen randomly from each sequence and Figure 10 provides some localization results for each sequence. The first column shows an airfield video with flat background. The vehicles turn around or pass by each other. The second column shows two vehicles in a forest. The vehicles may pass through tree shadow or become partially occluded. The third column shows an intersection with static background, in which multiple vehicles exist[3]. The fourth column displays two different weather conditions (snow and fog) at the same intersection. From the results, we can see that objects can be localized well except when the objects move close to each other in the intersection (the 1190th frame in column 3) or when an object's intensity is very similar to the background (the 1250th frame in column 3).

5 Conclusion

Motion history images accumulate change detection results with a decay term over a short period of time. The MHI contains more motion information than a single motion image generated by frame difference. Instead of only showing where the motion is, MHI answers the questions of "what went where"and "how did it go there". Each moving object has a fading trail, with the trail showing the direction of movement. By combining the forward MHI and backward MHI together, we can get a contour shape for the moving object at the current frame. The experiments show the effectiveness of the approach. In addition, the method is much faster to run and to implement than multi-scale optical flow. There are only several subtraction and comparison operations for each pixel at each iteration. However for optical flow method, if each pixel has a k*k window, the computation cost is roughly increased by a factor of k^2.

Future work will implement this localization approach within a complete tracking system. The motion, shape and appearance features of detected objects will be combined to represent and track the object. Furthermore, based on the detected object location and initial shape estimation, more accurate local segmentation methods can be performed around the object to get better layer representations of the object and background.

[3] Downloaded from the Karlsruhe University at: http://i21www.ira.uka.de/image_sequences/

References

[1] A. Bobick and J. Davis, "The Recognition of Human Movement Using Temporal Templates," IEEE Transactions Pattern Analysis and Machine Intelligence, 23(3): 257-267 March 2001.

[2] G. Bradski and J. Davis. "Motion segmentation and pose recognition with motion history gradients, " Fifth IEEE Workshop on Application of Computer Vision, 238-244, Dec. 2000

[3] G. Halevi and D. Weinshall, "Motion of Disturbances: Detection and Tracking of multi-Body non-Rigid Motion,"IEEE Conference on Computer Vision and Pattern Recognition, Puerto Rico, 1997, pp. 897-902.

[4] I. Haritaoglu, D. Harwood and L. Davis, "W4: Real-time Surveillance of People and Their Activities," IEEE Transactions Pattern Analysis and Machine Intelligence, 22(8): 809-830 August 2000.

[5] M. Irani and P. Anandan, "A Unified Approach to Moving Object Detection in 2D and 3D Scenes," IEEE Transactions Pattern Analysis and Machine Intelligence, 20(6): 577-589 June 1998.

[6] R. Kumar, H. Sawhney et.al, "Aerial Video Surveillance and Exploitation," in Proceedings of the IEEE, 89(10): 1518-1539 October 2001.

[7] N. Paragios and R. Deriche, "Geodesic Active Contours and Level Sets for the Detection and Tracking of Moving Objects," IEEE Transactions Pattern Analysis and Machine Intelligence, 22(3): 266-280 March 2000.

[8] C. Stauffer and W. Grimson, "Learning Patterns of Activity Using Real-Time Tracking," IEEE Transactions Pattern Analysis and Machine Intelligence, 22(8): 747-757 August 2000.

[9] A. Strehl and J. Aggarwal, "Detecting Moving Objects in Airborne Forward Looking Infra-Red Sequences," IEEE Workshop on Computer Vision Beyond the Visible Spectrum: Methods and Applications, 1999

[10] L. Wixson, "Detecting Salient Motion by Accumulating Directionally-Consistent Flow, " IEEE Transactions Pattern Analysis and Machine Intelligence, 22(8): 774-780 August 2000.

[11] H. Yalcin, R. Collins and M. Hebert, "Background Estimation under Rapid Gain Change in Thermal Imagery," Second IEEE Workshop on Object Tracking and Classification in and Beyond the Visible Spectrum, 20-26, June 2005.

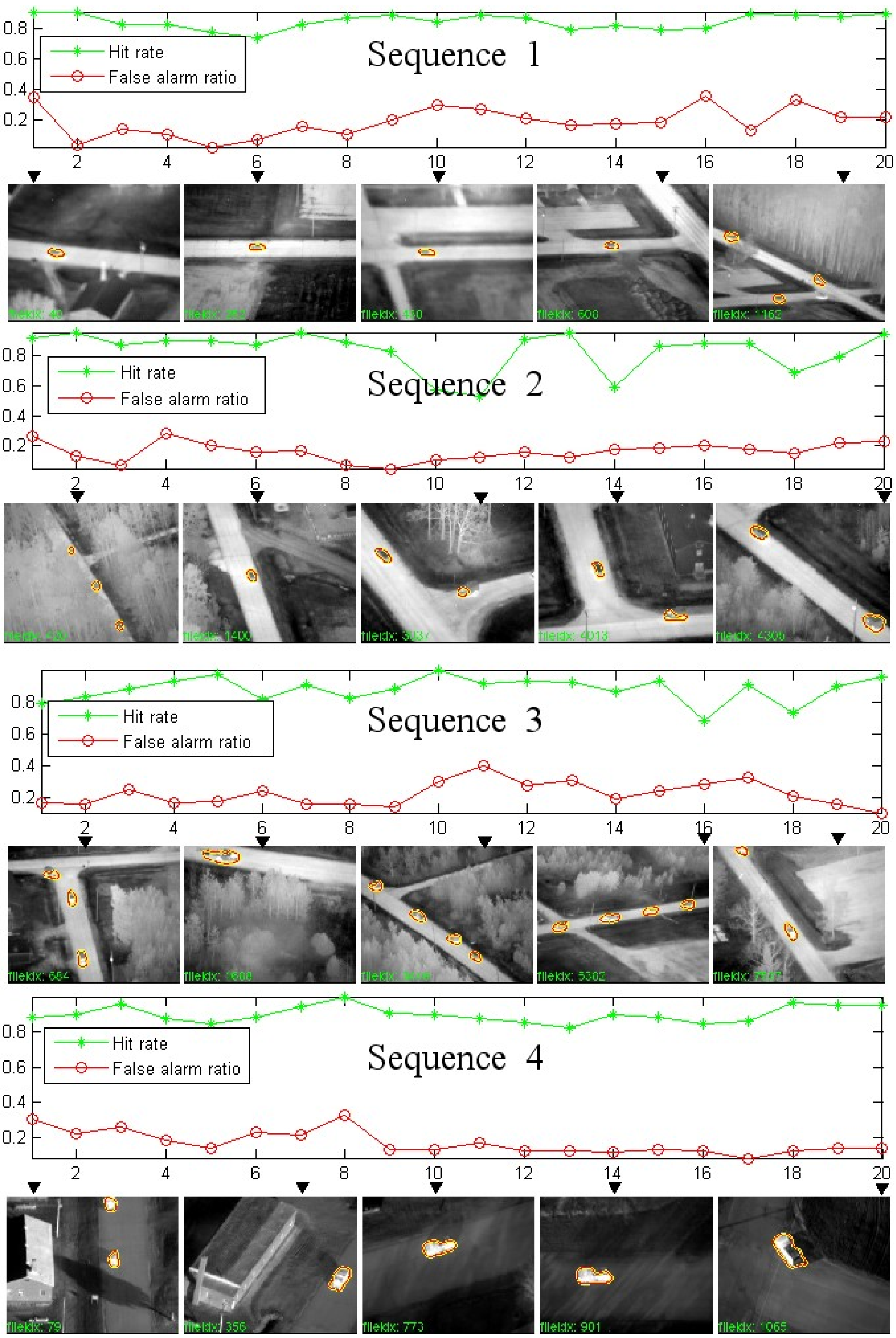

Figure 9: Evaluation on the thermal videos

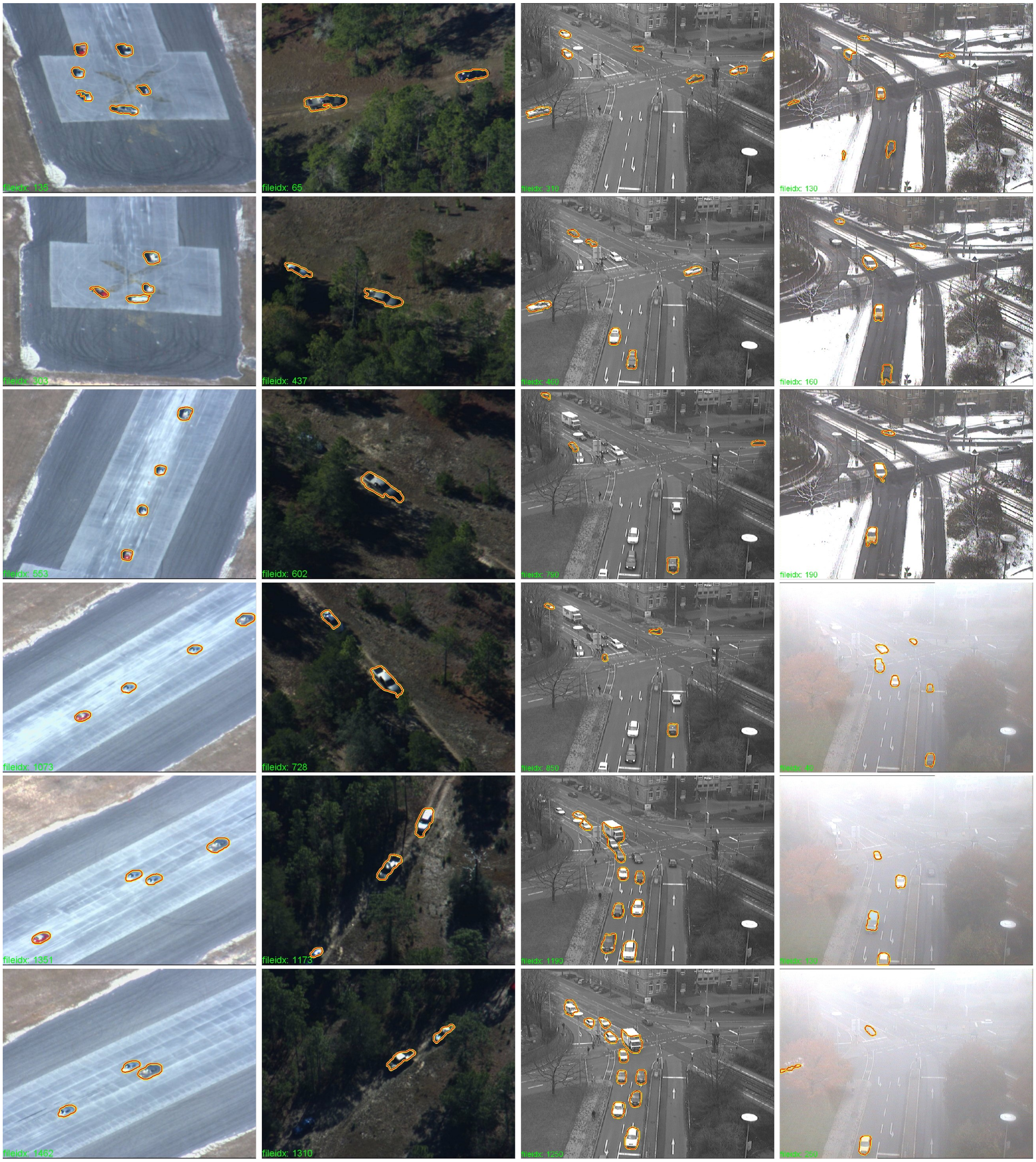

Figure 10: Test on different scenes