

# Layered Representation for Pedestrian Detection and Tracking in Infrared Imagery

Congxia Dai, Yunfei Zheng and Xin Li

Lane Department of Computer Science and Electrical Engineering

West Virginia University

Morgantown, WV 26506-6109

congxiad,zheng,xinl@csee.wvu.edu

## Abstract

This paper introduces a layered representation for infrared imagery and studies its application into pedestrian detection and tracking. We present a generalized EM algorithm to decompose infrared images into background and foreground layers and study the phenomenon of polarity switch. We propose a hybrid (shape+appearance) algorithm for pedestrian detection, in which shape cue is first used to eliminate non-pedestrian moving objects and appearance cue is then used to pin down the location of pedestrians. We also formulate the problem of shot segmentation and present a graph matching-based pedestrian tracking algorithm. Experimental results with OSU Thermal Pedestrian Database are reported to demonstrate the excellent performance of our algorithms.

## 1 Introduction

### A. Background

Advances in sensor and computing technologies have tremendously enhanced the capabilities of machine vision. In particular, the cost of thermal sensors has reduced dramatically in the past decades and we start to witness the popularity of infrared (IR) imagery with high dynamic range and sensitivity. Since the wavelength of IR sensors is beyond visible spectrum, they often have distinguished imaging capability. In particular, long-wave IR (thermal) sensors have been widely deployed in various applications such as night-vision and all-weather surveillance.

Recently, there have been a flurry of works on pedestrian detection and tracking in IR imagery. In [6], probabilistic templates are used to capture the variations in human shape for pedestrian detection. In [13], support vector machine and Kalman filtering are adopted for detection and tracking respectively. In [14], the P-tile method is developed to detect human head first and then human torso and legs are included by local search. In [7], a particle swarm optimization

algorithm is proposed for human detection in IR imagery. In [3], a two-stage template-based method with an Adaboosted classifier was presented for pedestrian detection.

In visible spectrum, human vision system (HVS) is often used as the benchmark for the robustness and accuracy of machine vision systems. However, as we enter the IR range, several tantalizing question arises: is simulating HVS still the right approach? What do we mean by “look like” if the object is invisible to human eyes but still detectable by thermal sensors? How can we display thermal sensor data to achieve the best visualization effect (note that we are not directly seeing any object but its response to electromagnetic waves)?

All those questions indicate that conventional wisdom in computer vision research might not be applicable as we go beyond the visible spectrum due to the different imaging principle. In the early development of HVS, stimulus such as shape, contrast, and color all belong to visible range which the HVS has learned to adapt to. Therefore, even if IR sensor data can be displayed on the screen like visible grayscale images, their appearance on the screen often does not fully convey the information they carry.

For example, Fig. 1 shows two sample images from OTCBVS benchmark - the OSU thermal image database. In the left image, the contrast is particularly poor due to the rainy weather condition, which makes it difficult for HVS to discern the person at the highlighted spot. In the right image, the person is at the image boundary and partially occluded by a tree. However, if we look at the pseudo-colored version of the same sensor data, the occluded person appears to be more “visible” (refer to Fig. 2).

### B. Contributions

In this paper, we argue that the goal of computer vision beyond visible spectrum is to develop computational and statistical techniques with optimized performance for the given sensor data and tasks. Specifically, we present a layered representation for IR imagery and demonstrate its effectiveness in pedestrian detection and tracking. Layered



Figure 1: Discerning pedestrians from IR images could be tricky for human eyes due to thermal properties (left) or occlusion (right).

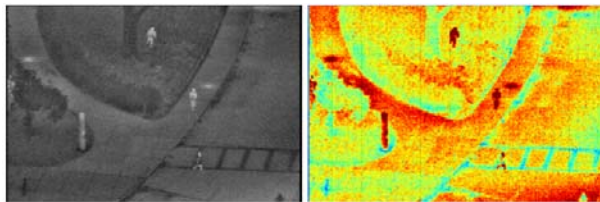


Figure 2: Visualization of sensor data affects the interpretation by HVS: the right image is arguably easier for pedestrian detection than the left one.

representations have been widely used by object tracking in visible imagery [5]. For IR imagery, layered representation is also attractive because it facilitates the statistical modeling of sensor data. We propose to decompose an IR image into two layers: background (still objects) and foreground (moving objects).

Layer separation or background extraction is easy if the camera is kept still (e.g., using the standard EM algorithm). To accommodate camera panning, we propose a generalized EM-based background mosaicing algorithm that takes the global motion into account. One tricky question arises when pedestrians remain still throughout the sequence - should we assign them to foreground or background? To overcome such difficulty, we propose an engineering solution - i.e., IR sensor can be programmed to take shots either frequently in a day (e.g., every other hour) or at particularly chosen timing (e.g., 6AM when it is unlikely to catch pedestrians).

Layered representation facilitates the detection and tracking of pedestrians. The support of foreground layer is decomposed of connected regions that correspond to moving objects. We propose to locate pedestrians from the foreground layer in two steps: 1) shape-based classification - does this object contain any pedestrian? 2) appearance-based localization - where is the exact position of each pedestrian? In the classification step, compactness and leanness of the object are extracted and trained by support vector machine (SVM) [11]. In the localization step, a modified principal component analysis (PCA) technique [10] is

developed to statistically infer the most likely position for each pedestrian. Such hybrid (shape+appearance) approach has achieved excellent performance on OTCBVS benchmark.

We also formulate the problem of shot segmentation - i.e., segment an ordered yet nonuniformly sampled image sequence into different shots. We have developed a geometric shot segmentation technique based on Hausdorff distance. Within each shot, unlike existing recursive tracking techniques based on Kalman-filter [13], we formulate tracking as a matching problem on weighted bipartite graphs. Each pedestrian is treated as a node and for every matching (edge) between two nodes from adjacent frames, we adaptively allocate the weight reflecting the tradeoff between shape-appearance similarity and geometric proximity. We also pay special attention to frames with overlapped pedestrians that are difficult for tracking.

The rest of this paper is organized as follows. Section 2 describes a generalized EM algorithm for background extraction and discusses the detection of polarity switch. Section 3 includes shape-based classification and appearance-based localization techniques for pedestrian detection. Section 4 covers shot segmentation and matching-based pedestrian tracking. Experimental results are reported in Section 5 and concluding remarks are included in Section 6.

## 2 Layered Representation

### A. Modeling of IR Imagery

Let us introduce some notations first. A sequence of IR images are denoted by  $I_k(m, n)$  where  $k = 1, 2, \dots, K$  and  $(m, n) \in \Omega = [1, M] \times [1, N]$  are temporal and spatial variables respectively. Each sequence is assumed to be taken at a large camera distance and within a short period of time such that environmental factors such as precipitation and temperature remain unchanged. Any IR image is decomposed of background and foreground layers and we make the following key observations:

- Background (still) layer

A still background is defined to be the sensed environment without any moving object (e.g., walking person, moving vehicle). Three kinds of uncertainty factors are considered in background modeling: 1) sensor-related including motion and noise characteristics of the sensor; 2) weather-related such as rainy, cloudy or sunny conditions; 3) landscape-related, namely objects that cover a wide range of thermal spectrum but physically remain still (e.g., trees, benches and post lights).

- Foreground (moving) layer

The thermal signature of moving objects is varying in both spatial and temporal domain. Taking pedestrians as an example, human head usually disperses more heat than other parts of human body because it is directly exposed to

the outer environment; arms and legs often experience more variation temporally than head and torso due to walking-related motion. Additionally, weather factors could give rise to rare events in IR images (e.g., umbrella or raincoat that affects the thermal signature of a person).

The above observations indicate that temporal constraints of motion, shape and appearance for visible imagery [1], [5] do not hold for IR imagery. Instead, we propose a two-layer representation

$$I_k = (1 - M_k)B_k + M_kF_k + W_k, \quad (1)$$

where  $B_k, F_k, M_k$  stand for background, foreground, mask layer respectively and  $W_k$  models sensor noise. When compared with more complicated layered representations developed for visible imagery, such two-layer representation is adopted for its simplicity and suitability for detection and tracking applications. Since mask layer carries critical information about moving objects, we target at a maximum-likelihood (ML) extraction of mask layer along with the background.

Another important difference of IR imagery from visible imagery is the heavy noise. The strength of thermal sensor noise is strong enough to be highly visible. Empirical studies have shown that the additive noise term  $W_k$  approximately observes the Gaussian distribution with zero mean and variance of  $\sigma_w \in [40, 60]$ . Such heavy noise poses a challenge to both background extraction and pedestrian detection. In background extraction, we will suppress noise components by adaptive temporal filtering; in pedestrian detection, we will intelligently choose the number of principal components to minimize the noise interference.

### B. Background Extraction

In the absence of camera motion (i.e.,  $B_k = B$ ), EM algorithm can be used to extract the still background. With camera motion, we are facing a more general background mosaicking problem [2] - each  $B_k$  can be viewed as a subset of the mosaicked image  $B$ . In this section, we present a generalized EM algorithm to mosaick the background of IR imagery under the assumption of camera panning motion.

In the basic EM algorithm without camera motion, mask layer  $M_k$  and background layer  $B_k = B$  are iteratively refined (refer to [4, page 310]). In our generalized EM algorithm, an additional image alignment step is adopted to handle the global camera motion. At the initialization, phase correlation method is used to register  $K$  IR images and produce an initial estimation of  $B$ . At each iteration, we update  $M_k$  by thresholding  $|I_k - B_k|$ , refine the alignment results by excluding the foreground pixels and then update  $B_k$  by adaptively averaging the  $K$  registered images.

The stopping criterion is set to be  $\|B^{(t+1)} - B^{(t)}\|^2 < \delta$ , where  $\delta$  is a small positive number (e.g., 0.01). Empirical studies have shown that such algorithm converges rapidly (typically 3 iterations). To improve the robustness, we use

morphological filtering to process the mask layer to eliminate small objects (connected components) and fill in the holes of moving objects. After background extraction, the set  $\Omega_{mov} = \{(m, n) | M_k(m, n) = 1\}$  consists of connected components  $R_1, \dots, R_{E_k}$  where  $E_k$  is the total number of moving objects.

One interesting phenomenon with IR imaging is the so-called ‘‘polarity switch’’. When it occurs, hot and cold ranges of thermal sensor get reversed - for instance, pedestrians that normally give rise to bright pixels could become dark pixels. To the best of our knowledge, the mechanism of polarity switch has not been well documented in the literature. The test data available for experimental studies are also limited at this point. Therefore, based on a simplified assumption that a significant portion of pedestrians would turn dark when polarity switch occurs, we propose to compute the average of  $e_k = F_k - B_k$  over the set  $\Omega_{mov}$  and use the polarity of  $e_{avg}$  as the indicator (refer to Fig. 7).

Another tricky issue with background extraction is that some pedestrians could remain still throughout the whole sequence and therefore get assigned to background instead of foreground due to the limited recording time. If the temporal interval between images is sufficiently long (e.g., an hour), such difficulty easily goes away because it is unlikely someone would remain still for hours. However, as the recording time increases, environmental factors might vary as well. Therefore, we propose a compromised solution in practice - sensor can be programmed to either take shots more frequently in a day or take some images at the sunrise everyday - those images are used as the reference only.

## 3 Static Pedestrian Detection

Unlike joint exploitation of motion and appearance for pedestrian detection in visible spectrum [12], motion cue might not be available in IR imagery. Therefore, we constrain ourselves to static pedestrian detection here and consider the shape and appearance cues only. Shape alone is difficult because human silhouette is view-dependent and becomes more complex when more than one person cluster around. Appearance alone is insufficient either because unlike visible imagery, color and texture do not contribute to the appearance variation in IR imagery. In this paper, we propose to first use shape information to reject non-pedestrian moving objects with high confidence and then resort to appearance to locate the exact position for each pedestrian. Such sequential exploitation of shape and appearance cues distinguishes our technique from previous works [3],[6].

### A. Shape-based Classification

To effectively eliminate non-pedestrian objects (vehicles and animals) from  $\Omega_{mov}$ , we consider the following shape

descriptors for a moving object: 1) compactness  $r_1 = \frac{p^2}{A}$ , where  $p$  and  $A$  measure the perimeter and area of the moving object respectively; 2) leanness  $r_2 = \frac{l_v}{l_h}$ , where  $l_h, l_v$  denotes the total length of projected segments of skeleton along the horizontal and vertical direction respectively. If the skeleton contains any bifurcation, we propose to decompose it into multiple curve segments and do the projection for each segment separately (refer to Fig. 3).

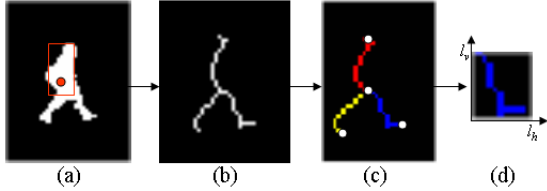


Figure 3: a) moving object; b) skeleton; c) decomposed curve segments; d) projection onto horizontal and vertical axes.

The two scale-invariant descriptors are adopted for distinguishing animals and vehicles from humans. Animals (e.g., cats and dogs) typically have smaller compactness ratio than humans due to their dominant torso; vehicles often have smaller leanness ratio due to their fat shape. The classification of any moving object into pedestrian and non-pedestrian is based on  $(r_1, r_2)$  and we adopt the SVM-based classifier [11]. Since our shape classifier does not need to separate connected pedestrians apart, it achieves excellent Receiver Operating Characteristic (ROC) performance (refer to Fig. 4).

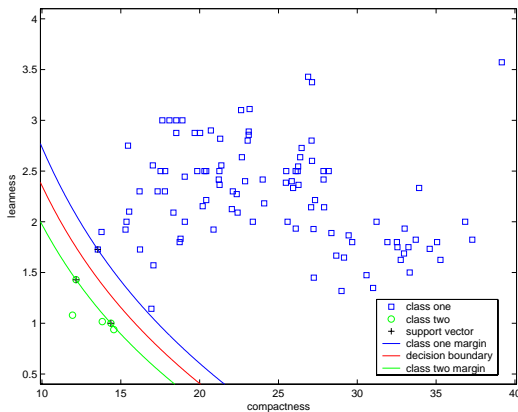


Figure 4: Shape classifier trained by SVM.

## B. Appearance-based Localization

Based on the shape classification results, we continue to exploit the appearance cue to pin down the exact location of each pedestrian within each moving object. We adopt

a modified principal component analysis (PCA) technique similar to eigenface [10]. Note that here the role of PCA is not to detect but to *localize* the pedestrians. PCA is attractive for IR imagery because thermal imaging is less sensitive to illumination and pose. The linear subspace structure of PCA lends itself to the task of separating connected or locating partially occluded pedestrians.

In our current implementation, we choose a fixed window sized  $20 \times 30$  for each pedestrian, though it is straightforward to generalize it into multi-resolution at the price of higher complexity. The PCA-based localization technique is summarized as follows.

### Stage 1: Normalization

We normalize  $e_k = F_k - B_k$  by the maximum absolute value in each frame, i.e.,

$$\bar{e}_k = \left| \frac{e_k}{e_{k,max}} \right|. \quad (2)$$

where  $e_{k,max} = \max_{m,n} |e_k(m,n)|$  and taking the absolute value is for the purpose of accommodating polarity switch. Both training and detection are performed with the normalized difference image  $\bar{e}_k$ .

### Stage 2: Training and Projection

We follow the same procedure as presented in [10] to obtain an image  $p(m,n)$  indicating the likelihood of containing a pedestrian at each pixel. The major consideration here is the choice of how many largest eigenvalues be considered as principal. The optimal value is determined by the power of noise as well as the subspace structure of signal. For OSU thermal pedestrian database, our studies show that keeping the largest twelve eigen-vectors achieves nearly optimal signal-noise separation.

### Stage 3: Location Aggregation

We first identify all local minimum in  $p(m,n)$  whose value is below a pre-selected threshold  $th$  as the candidates. Then we locally aggregate the multiple candidates into one if they are too close to each other. Specifically, if the overlapped area of two candidates is more than 30% of the window area ( $20 \times 30$ ), we aggregate them into one; otherwise they are treated as two adjacent yet different pedestrians.

We note the three significant departures from conventional PCA practice. First, normalization is important to accommodate a variety of environmental conditions across the database. Second, PCA is only applied to a small fraction of pixels in the image - i.e. the moving objects in the foreground layer that have survived shape classification. In fact, the complexity can be further reduced by the use of distance transform because the local minimum in  $p(m,n)$  is likely to be close to the skeleton of each object. Third, local aggregation strategy helps more accurately align the template window with the pedestrian. Such strategy also reflects our idea of using PCA as a localization tool to handle the challenging overlapped-pedestrian cases.

## 4 Pedestrian Tracking

Tracking can be viewed as the dynamic extension of detection where motion smoothness constraint is exploited to establish the correspondence of moving objects across multiple frames. However, such motion-related constraint can only be exploited for video frames whose sample rate is sufficiently high (otherwise they are no different from static images). For an ordered yet nonuniformly-sampled collection of IR imagery, we need to do shot segmentation before tracking. A shot is defined as a collection of consecutive frames whose adjacent time interval is sufficiently small (e.g., a fraction of second).

### A. Shot Segmentation

Based on the above definition, it is reasonable to assume that frames within the same shot look more alike than those outside. In visible imagery, histogram-based techniques are often suitable for shot segmentation. However, histogram becomes less effective for IR imagery because pixels in the still background would dominate those in the moving foreground. In this paper, we propose a Hausdorff-distance based shot segmentation algorithm with low complexity.

Recall the collection of objects in foreground layer is labeled  $R_1, \dots, R_{E_k}$ . For an object  $R_k$ , we collect the endings and intersections along its skeleton and form a feature point set  $C_k$  (solid circles in Fig. 3c). To measure the distance between two feature point sets  $C_k$  and  $C_{k+1}$ , the Hausdorff distance has been widely used in the literature of computer vision [8]. We adopt the following definition of Hausdorff distance

$$H(C_k, C_{k+1}) = \frac{h(C_k, C_{k+1}) + h(C_{k+1}, C_k)}{|C_k| + |C_{k+1}|},$$

$$h(X, Y) = \max\{\min\{d(a, b)\}\}, a \in X, b \in Y. \quad (3)$$

where  $d(a, b)$  denotes the Euclidean distance between two points  $a, b$ . Note that the above definition has enforced the symmetry, i.e.,  $H(C_k, C_{k+1}) = H(C_{k+1}, C_k)$ .

One practical constraint that has not been considered in the definition of Hausdorff distance is that images have finite size. Therefore, if some pedestrian happens to enter or leave the field of view, Hausdorff distance between two frames could be large even if they are temporally close. To overcome such difficulty, we opt to exclude the pedestrians around image boundary in the calculation of Hausdorff distance. Two frames are grouped together if and only if their Hausdorff distance is below a pre-selected threshold.

### B. Graph Theoretic Tracking

Within the same shot, pedestrian tracking in IR imagery is often more difficult than that in visible imagery. Unlike visible imagery containing color and texture cues, shape is arguably the only cue that can be exploited by tracking in IR imagery. When the camera distance is large, shape discrepancy between two different person but with similar weight

and height is small. The thermal signature of a person is constantly varying due to the walking motion. Moreover, when two pedestrians walk closely or pass by each other, the overlapped shape of pedestrians experiences severe deformation, which makes tracking even more difficult.

To overcome the above difficulties, we propose the following two strategies. First, instead of considering the whole silhouette of a person, we only take the head and torso into account in tracking for the purpose of suppressing the interference of limb motion. Specifically, we calculate the centroid of any pedestrian  $u$  and take a  $10 \times 15$  rectangular box  $\Omega_{ht}$  surrounding the centroid (refer to the red box in Fig. 3a). Then the intensity profile of moving object within that box can be characterized by a 3D surface  $S_u = \{(m, n, I(m, n)) | (m, n) \in \Omega_{mov} \cap \Omega_{ht}\}$ . To measure the similarity between two pedestrians  $u, v$ , we can align them at the centroid and calculate the Hausdorff distance between  $S_u$  and  $S_v$ . Such 3D geometric approach emphasizes both the shape and appearance of pedestrians and we denote the calculated distance by  $d_{sa}$ .

Second, we propose to adaptively exploit the cue of similarity and proximity in the spatio-temporal domain. If pedestrians in the scene are far away from each other, geometric proximity is often sufficient for establishing the correspondence (e.g., nearest neighbor rule). In difficult scenarios where multiple pedestrians walk closely or pass by each other, shape and appearance are often more reliable for tracking purpose. Temporally, we propose to handle frames that do not contain overlapped pedestrians first and attack difficult frames based on the tracking results obtained for easy ones. Such temporal adaptation is important to resolve the correspondence ambiguity by exploiting the pattern of walking motion.

We opt to implement the above ideas for two-frame tracking under a graph matching framework [9]. Let  $G$  be a weighted graph, in which  $2Q$  nodes denote the detected pedestrians:  $U = \{u_1, \dots, u_Q\}$  from  $I_k$  and  $V = \{v_1, \dots, v_Q\}$  from  $I_{k+1}$ . For any  $u \in U$  and  $v \in V$ , there is an edge between them whose weight is

$$w(u, v) = \alpha d_{sa} + (1 - \alpha) d_{eu}. \quad (4)$$

where  $d_{sa}$  is defined earlier,  $d_{eu}$  is the Euclidean distance between the centroid of  $u, v$  and the weight  $\alpha$  is the overlapping ratio of  $u, v$  (i.e., the ratio of overlapped area to pedestrian window size). Note that when  $\alpha > 0$ ,  $d_{sa}$  is often much larger than  $d_{eu}$  and easily dominates the weight assignment. With the above-defined weighted graph  $G$ , two-frame pedestrian tracking can be formulated as a bipartite matching problem with set  $U$  and  $V$ .

For multiple frames, we assume that frames without overlapped pedestrians are processed first. Conditioned on the tracking results for those frames, we modify the weight assignment strategy for overlapped pedestrians to reflect our

a priori knowledge about human walking motion. For example, if we assume that the motion trajectory of human walking is smooth, we might replace the Euclidean distance  $d_{eu}$  by the derivative of turning angles, i.e.,

$$w(u, v) = \alpha d_{sa} + (1 - \alpha) |\phi(\vec{u}'u, \vec{v}v) - \phi(\vec{u}v, \vec{v}v')|. \quad (5)$$

where  $u', v'$  are the matched pedestrians of  $u, v$  in  $I_{k-1}, I_{k+2}$  respectively and  $\phi(\vec{a}, \vec{b})$  denotes the turning angle between two vectors. Note that the adoption of derivative is due to the observation that for any pedestrian, the events of making a sharp turn and passing over another person are unlikely to occur at the same time.

## 5 Experimental Results

In this section, we report our experimental results with OTCBVS benchmark - OSU thermal pedestrian database [3]. Currently, there are ten test sequences in the database. Each sequence contains 18-73 frames that are taken within one minute but not temporally uniformly sampled (they are the subset of 30Hz video coming out of IR camera). The database reasonably covers a variety of environmental conditions such as rainy, cloudy and sunny days; however, only one sequence (#3) contains polarity switch. The camera is kept still all the time and non-pedestrian moving objects in the database are rare (only one moving vehicle and one walking dog).

### A. Background Extraction

We first demonstrate the performance of the proposed EM-like background extraction algorithm. To simulate camera motion, we randomly shift each frame by  $(d_x, d_y)$  ( $d_x, d_y$  are random integers between 1 and 10). Fig. 5 shows the background mosaicking result for original sequence #4 without camera motion (standard EM algorithm) and shifted sequence with simulated camera motion (generalized EM algorithm). It can be observed that generalized EM algorithm effectively separate moving objects from the background regardless of camera motion.

To illustrate the problem with motionless pedestrians, we take sequence #8 as an example. If we only use its 24 frames in background extraction, the two pedestrians will be assigned to  $B$ , which seriously affects the detection performance as observed in [3]. Since we do not have any reference image taken at the same day, we opt to add another 24 frames of sequence #10 with similar thermal characteristics. It can be observed from Fig. 6 that incorporating more frames into background extraction alleviates the problem, though some ghost shadow of two pedestrians remains (it has been experimentally confirmed that the ghost shadow does not affect the detection). We conjecture that motionless pedestrians will not pose any problem if appropriate reference frames are available (refer to the engineering solutions sketched at the end of Sec. 2).



Figure 5: Extracted background for sequence #4 without (left) and with (right) simulated camera motion.

We have also did a preliminary test with polarity switch detection. Fig. 7 plots the calculated  $e_{avg}$  for all ten test sequences. It clearly indicates the negative value for #3, which is the only one with polarity switch among ten.



Figure 6: Extracted background using 24 frames in #8 only (left) and 48 frames in #8, #10 together (right).

### B. Pedestrian Detection

Fig. 9 shows the detection result for all ten sequences in the database. We have adopted the terminology in [3] to facilitate the comparison. When compared with [3], our approach noticeably works better on sensitivity performance. It should be noted that the ground truth provided in the database only include those people that are at least 50% “visible”. Highly-occluded pedestrians are not counted in the ground-truth, which has negative impact on the detection performance of our scheme. Fig. 10 shows some examples of miss detection (due to overlapping) and arguably correct detection (not included in the ground-truth table due to occlusion).

### C. Pedestrian Tracking

In Fig. 7, we also show the shot segmentation result for sequence #9. The total 73 frames are segmented into four separate shots. The starting frames of four shots are shown in Fig. 8. In tracking experiments, the following heuristics is used in finding the optimal matching: if a pedestrian is isolated, the best matching can be found by direct local search (i.e.,  $\alpha = 0$ ); only for overlapped pedestrians ( $\alpha > 0$ ), we need to exhaustively try out different assignments. Fig. 11 shows the tracking result for frames No. 13-18 of sequence #5. Each pedestrian is marked by a different color. It can be observed that tracking is successful despite slight overlapping of pedestrians. Note that since

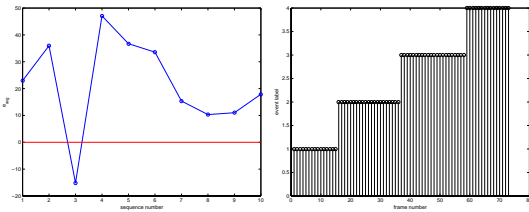


Figure 7: Polarity switch detection result (left) and shot segmentation result for sequence #9 (right).

our tracking is based on detection results tracking will not work if any pedestrian is missed at the detection stage.

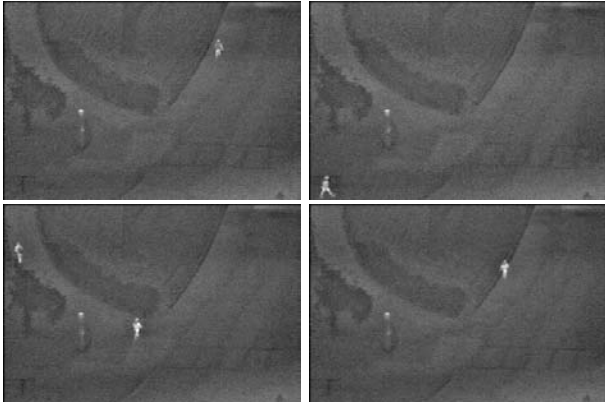


Figure 8: Frames No. 1, 16, 37, 59 in sequence #9 - they are the starting frames of new shots.

## 6 Conclusions

This paper presents a two-layer representation for detection and tracking of pedestrians from IR imagery. A generalized EM algorithm is proposed for layer decomposition. A two-stage pedestrian detection algorithm is developed to first classify moving objects of the foreground layer based on their shapes and then localize each individual pedestrian based on the appearance. Such sequential exploitation of shape and appearance cues leads to superior performance to previous works. We also study the shot segmentation problem and present a graph-matching based algorithm for tracking pedestrians within the same shot. Preliminary results on pedestrian tracking are reported.

We believe tracking can help further improve the detection performance (i.e., from static to dynamic). Within each shot, the motion cue can be exploited to resolve the ambiguity with overlapped pedestrians. A robust overlap detection mechanism is desirable. The phenomenon of polarity

switch remains poorly understood - e.g., under what condition does it occur? More extensive experiments on background mosaicing, polarity switch detection and pedestrian tracking need to be done when more test data become available (e.g., IR imagery acquired more frequently, with camera motion and/or polarity switch).

## References

- [1] S. Ayer and H. Sawhney. Layered representation of motion video using robust maximum-likelihood estimation of mixture models and mdl encoding. In *IEEE Conference on Computer Vision*, pages 777–784, 1995.
- [2] P. Burt and E. Adelson. A multiresolution spline with application to image mosaicking. *ACM Transactions on Graphics*, pages 217–236, 1983.
- [3] J. Davis and M. Keck. A two-stage approach to person detection in thermal imagery. In *Proc. Workshop on Applications of Computer Vision*, 2005. IEEE OTCBVS WS Series Bench.
- [4] D. Forsyth and J. Ponce. *Computer Vision: a Modern Approach*. Prentice-Hall, 2002.
- [5] H. S. H. Tao and R. Kumar. Object tracking with bayesian estimation of dynamic layer representations. *IEEE Transactions on Pattern Anal. Mach. Intell.*, 24:75–89, 2002.
- [6] H. Nanda and L. Davis. Probabilistic template based pedestrian detection in infrared videos. In *Proc. Intell. Vehicles Symp.*, 2002.
- [7] Y. Owechko, S. Medasani, and N. Srinivasa. Classifier Swarms for Human Detection in Infrared Imagery. In *IEEE Int. Workshop on Object Tracking and Classification Beyond the Visible Spectrum*, 2004.
- [8] W. J. Rucklidge. Efficiently locating objects using the hausdorff distance. *Int. J. of Computer Vision*, 24:251–270, 1997.
- [9] A. Shokoufandeh and S. Dickinson. Graph-theoretical methods in computer vision. pages 148–174, 2002.
- [10] M. Turk and A. Pentland. Eigenface for recognition. *Journal of Cognitive Neuro-science*, 3:70–86, 1991.
- [11] V. Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag, 1995.
- [12] P. Viola, M. Jones, and D. Snow. Detecting pedestrians using patterns of motion and appearance. In *Proc. IEEE Conf. Computer Vision*, 2003.
- [13] F. Xu and K. Fujimura. Pedestrian detection and tracking with night vision. In *Proc. Intell. Vehicles Symp.*, 2002.
- [14] M. Yasuno, N. Yasuda, and M. Aoki. Pedestrian Detection and Tracking in Far Infrared Images. In *IEEE Int. Workshop on Object Tracking and Classification Beyond the Visible Spectrum*, 2004.

Sequence No.	# Frames	# People	#TP		#FP		Sensitivity		PPV	
			[3]	Ours	[3]	Ours	[3]	Ours	[3]	Ours
1	31	91	88	91	0	0	0.97	1.00	1.00	1.00
2	28	100	94	99	0	0	0.94	0.99	1.00	1.00
3	23	101	101	100	1	2	1.00	0.99	0.99	0.98
4	18	109	107	109	1	0	0.98	1.00	0.99	1.00
5	23	101	90	101	0	0	0.89	1.00	1.00	1.00
6	18	97	93	97	0	0	0.96	1.00	1.00	1.00
7	22	94	92	94	0	0	0.98	1.00	1.00	1.00
8	24	99	75	99	1	1	0.76	1.00	0.99	0.99
9	73	95	95	95	0	0	1.00	1.00	1.00	1.00
10	24	97	95	94	3	3	0.98	0.97	0.97	0.97
1-10	284	984	930	979	6	6	0.95	0.99	0.99	0.99

Figure 9: Detection results for OSU thermal pedestrian database.

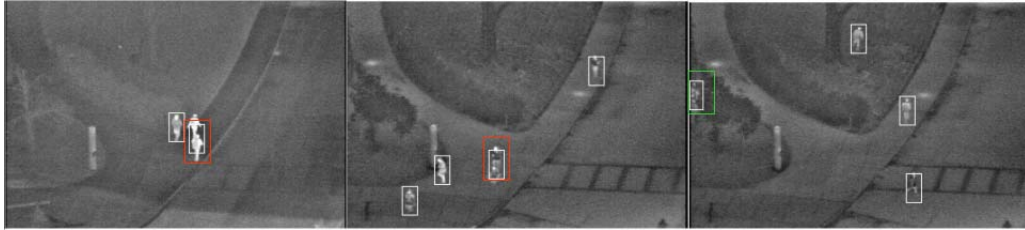


Figure 10: Examples of miss detection (red box) and arguably correct results (green box).



Figure 11: Tracking results for frames No. 13-18 in sequence #5.