Detection of People in Images

A.N. Rajagopalan, Philippe Burlina and Rama Chellappa

Center for Automation Research
University of Maryland
College Park, MD - 20742

Abstract

The paper describes a scheme for detecting and tracking people in images. The method effectively combines statistical information about the class of people with motion information for classification and tracking. In this scheme, the unknown distribution of the images of people is approximately modeled by learning higher order statistics (HOS) information of the 'people class' from sample images. Given a test image, statistical information about the background is learnt dynamically. A motion detector identifies regions of activity in the image sequence. A classifier based on an HOS-based closeness measure then determines which of the moving objects actually correspond to people in motion. The tracking module uses position information and an HOS-based difference measurement vector to establish correspondence. When tested on real video data with a cluttered background, the performance of the method is found to be quite good. The method can also detect people in static imagery.

1 Introduction

Detection and tracking of people in images is a very important problem and has been receiving attention for a considerable amount of time. The task is a challenging one due to various reasons. People are non-rigid objects and it is very difficult to describe them analytically. Typically, the images of people present a significant variability in color and texture patterns within the boundaries of the body. Changes in orientation of the body due to motion must also be accommodated. Moreover, imposition of any constraints on the background must be usually avoided. The increasing availability of video sensors and high performance video processing hardware opens up exciting possibilities for visual surveillance, human motion analysis, and scene understanding.

Representative works on early detecting and tracking systems may be found in [1] to [4]. These approaches rely heavily on hand-crafted models and object motion. In [1], Akita uses stick figures and generalized cone approximation of body parts. Tsukiyama and Shirai [2] find candidate persons based on the mean and the variance of brightness of an image area and detect them by using models of the toe. Leung and Yang [3] use a voting process to determine candidate edges of moving body parts. In [4], Rohr uses a volume model consisting of cylinders to represent the human body.

Recently, real-time systems have been developed that are comparatively more robust. These include the Pfinder [5] and W^4 [6] which use background scene modeling along with a combination of shape analysis and motion models to detect and track people. While Pfinder relies on color cues, W^4 is designed to work with monochromatic imagery. In [7], a graph-theoretic approach is suggested which uses change detection along with first-order prediction for moving object analysis. Papageorgiou *et al.* [8] describe a pedestrian detection system that uses the Haar wavelet for image representation and the support vector machine for learning and classification. This scheme can also detect people in a single image. In [9], Kanade *et al.* use background subtraction for moving object detection and a neural network for classification. In [10], Lipton *et al.* use temporal differencing along with object motion for detection. Classification is done based on physical parameters of the object such as area and perimeter. Tracking is performed by correlation matching of image templates.

The detection and tracking scheme described in this paper uses higher order statistics (HOS) of images of people to get a better approximation to their unknown distribution. Training data samples of people are first clustered and the statistical parameters corresponding to each cluster are estimated. Clustering is based on an HOS-based decision measure which is obtained by deriving a series expansion for the multivariate probability density function in terms of the Gaussian function and the Hermite polynomial. Background information is learnt 'on the fly'. Simple frame differencing followed by thresholding and morphological operations are used to segment the moving objects from the background. An object discrimination module uses the statistical parameters of the 'people class' and the background, in conjunction with an HOS-based difference measure, to decide which of the moving objects correspond to people in motion. Position constraint along with an intensity constraint based on the HOS-based difference measurement is used for tracking the people over an image sequence.

The HOS-based closeness measure has very good discriminating capability. We now discuss detection of people in static imagery as well as video, using this measure.

0-7803-5529-6/99/$10.00 ©1999 IEEE

2 Two-Class Pattern Classification

Let $\underline{x}$ be an N-dimensional pattern vector in the N-dimensional Euclidean pattern space $\Omega_{\underline{x}}$. Consider the two-class problem with the hypothesis

$$H_1: \ \underline{x} \text{ belonging to class } \omega_1.$$
$$H_2: \ \underline{x} \text{ belonging to class } \omega_2.$$

The hypothesis testing problem may be interpreted as dividing the pattern space $\Omega_{\underline{x}}$ into two disjoint regions Ω_1 and Ω_2. If the observed sample $\underline{x}$ is in Ω_1, we accept the hypothesis H_1 and decide that $\underline{x}$ belongs to ω_1. If $\underline{x}$ is in Ω_2, we accept H_2 and decide that $\underline{x}$ belongs to ω_2.

Let the a priori probabilities for the two classes be $P(\omega_1)$ and $P(\omega_2)$. The Bayes classification rule which is based on the maximization of the a posteriori probability can then be expressed as follows [11].

$$\text{If } f_x(\underline{x}|\omega_1)P(\omega_1) > f_x(\underline{x}|\omega_2)P(\omega_2) \text{ then } \underline{x} \in \omega_1.$$
$$\text{If } f_x(\underline{x}|\omega_2)P(\omega_2) > f_x(\underline{x}|\omega_1)P(\omega_1) \text{ then } \underline{x} \in \omega_2.$$

Here, $f_x(\underline{x}|\omega_1)$ and $f_x(\underline{x}|\omega_2)$ are the conditional density functions of $\underline{x}$ given that $\underline{x}$ belongs to ω_1 and ω_2, respectively. The Bayes classifier leads to an optimal partition of the feature space of $\underline{x}$ into two disjoint regions $\mathcal{R}_1$ and $\mathcal{R}_2$ such that the average cost per decision (also called the Bayes risk), is minimized [11].

We denote

$$f_i(\underline{x}) = f_x(\underline{x}|\omega_i), \ i = 1, 2.$$

In practise, the conditional density is usually modeled as Gaussian, for mathematical convenience. However, with respect to the class of people and the background patterns, it is quite unlikely that they would be governed by a simple Gaussian distribution. Hence, we use an HOS-based expansion (derived in the next section) for modeling the conditional densities $f_1(\underline{x})$ and $f_2(\underline{x})$, corresponding to the people and the background class, respectively. The expansion uses higher order statistics of the data to get a better approximation to the underlying unknown density function.

2.1 HOS-Based Expansion

In this section, we derive a series expansion for a multivariate probability density function (p.d.f) in terms of the Gaussian function and the Hermite polynomial. An HOS-based decision measure is then derived from this expansion.

Let the random vector $\underline{X} = [X_1 \ X_2 \ \ldots X_N]^T$ and $\underline{X} \sim N(\underline{0}, I)$. If $\underline{t} = [t_1 \ t_2 \ \ldots t_N]^T$, then the moment generating function of $\underline{X}$ is given by $\Phi(\underline{t}) = E\left[\exp\left(\underline{t}^T\underline{X}\right)\right]$. Since these random variables are statistically independent, $\Phi(\underline{t}) = \exp\left(\frac{1}{2}\underline{t}^T\underline{t}\right)$. Therefore,

$$E\left[\exp\left(\underline{t}^T\underline{X} - \frac{1}{2}\underline{t}^T\underline{t}\right)\right] = 1.$$

Replacing $\underline{t}$ by $\underline{t} + \underline{s}$,

$$E\left[\exp\left(\underline{t}^T\underline{X} - \frac{1}{2}\underline{t}^T\underline{t}\right)\exp\left(\underline{s}^T\underline{X} - \frac{1}{2}\underline{s}^T\underline{s}\right)\right] = \exp\left(\underline{t}^T\underline{s}\right).$$

Expanding this equation in Taylor series, we obtain

$$E\left[\sum_{m,n=0}^{\infty}(\underline{t}^{\otimes n})^T\frac{\overline{H}_n(\underline{X})}{n!}\frac{\overline{H}_m^T(\underline{X})}{m!}(\underline{s}^{\otimes m})\right] = \sum_{m,n=0}^{\infty}\frac{1}{n!}(\underline{t}^{\otimes n})^T I_{q_n q_m}(\underline{s}^{\otimes m}),$$

where $\otimes$ represents the tensor product and $\underline{t}^{\otimes n} = \underbrace{\underline{t}\otimes\underline{t}\otimes\ldots\otimes\underline{t}}_{n \text{ times}}$. The matrix $I_{q_n q_m} = O_{q_n q_m}$ for $n \neq m$ and $I_{q_n q_m} = I_{q_n}$ for $n = m$, where $O_{q_n q_m}$ is the zero matrix with q_n rows and q_m columns while I_{q_n} is the identity matrix of dimension q_n. The vector $\overline{H}_n(\underline{x})$ is given by

$$\overline{H}_n(\underline{x}) = \left[(\underline{D_t}^{\otimes n})\exp\left(\underline{t}^T\underline{x} - \frac{1}{2}\underline{t}^T\underline{t}\right)\right]_{\underline{t}=0}$$

and $\underline{D_t} = \left[\frac{\partial}{\partial t_1}\ \frac{\partial}{\partial t_2}\ldots\frac{\partial}{\partial t_N}\right]^T$. The dimensions of $\overline{H_n}(\underline{x})$ and $\overline{H_m}(\underline{x})$ are given by q_n and q_m, respectively. By equating the coefficients of $\underline{t}$ and $\underline{s}$ on both sides, we obtain the following important orthogonality relation

$$E\left[\underline{H}_n(\underline{X})\underline{H}_m^T(\underline{X})\right] = I_{p_n p_m}. \qquad (1)$$

Note that the expectation is w.r.t $N(\underline{0}, I)$. In (1), $\underline{H}_n(\underline{x})$ is a vector whose elements are given by the product $\left(\prod_{i=1}^{N}\frac{H_{k_i}(x_i)}{\sqrt{k_i!}}\right)$ for all permutations of k_i, $i = 1, \ldots N$, such that $\sum_{i=1}^{N} k_i = n$. The dimensions of the vectors $\underline{H}_n(\underline{x})$ and $\underline{H}_m(\underline{x})$ are given by p_n and p_m, respectively. The term $H_{k_i}(x_i)$ is the Hermite polynomial of order k_i and is defined as

$$H_{k_i}(x_i) = \left[\frac{\partial^{k_i}}{\partial t_i^{k_i}}\exp\left(t_i x_i - \frac{1}{2}t_i^2\right)\right]_{t_i=0}.$$

Similarly, one can derive an orthogonality relation in terms of $N(\underline{\mu}, R)$ as

$$E\left[\underline{H}_n\left(R^{-\frac{1}{2}}(\underline{X} - \underline{\mu})\right)\ \underline{H}_m^T\left(R^{-\frac{1}{2}}(\underline{X} - \underline{\mu})\right)\right] = I_{p_n p_m},$$

where $R^{\frac{1}{2}} = UD^{\frac{1}{2}}U^T$, and U and D can be obtained through an eigenvalue-eigenvector decomposition of the matrix R. Let $\underline{Y} = R^{-\frac{1}{2}}(\underline{X} - \underline{\mu})$ and $\underline{y} = R^{-\frac{1}{2}}(\underline{x} - \underline{\mu})$. If $\underline{X}$ has mean $\underline{\mu}$ and covariance R, then using the above orthogonality relation, the

multivariate probability density function $f(\underline{x})$ can be written as

$$f(\underline{x}) = N(\underline{\mu}, R)\left(1 + \sum_{n=3}^{\infty} E\left[\underline{H}_n^T(\underline{Y})\right] \underline{H}_n(\underline{y})\right) . \quad (2)$$

In accordance with the Bayes classification rule, we define the HOS-based closeness measure as $-\log f(\underline{x})$ which is given by

$$-\log N(\underline{\mu}, R)\left(1 + \sum_{n=3}^{\infty} E\left[\underline{H}_n^T(\underline{Y}))\right] \underline{H}_n(\underline{y})\right) . \quad (3)$$

When $f(\underline{x})$ is Gaussian, the HOS-based decision measure neatly reduces to the Mahalanobis distance.

3 Detection in a Single Image

In this section, we propose an HOS-based people detection scheme that finds people by searching an image for different views of people at all points of the image and across different scales. The methodology involves statistical parameter estimation of the 'people class', dynamic background learning, and classification.

We use a statistical distribution-based model for people as well as the background. It is to be expected that the conditional density function for these two classes is unlikely to be well-modeled by a simple Gaussian fit $N(\underline{\mu}, R)$. The unknown p.d.fs' $f_1(\underline{x})$ and $f_2(\underline{x})$ are approximated up to their m^{th} order joint moment by (2) which uses higher order statistics to get a better approximation to the unknown p.d.fs'. As a compromise between accuracy of representation and computational complexity, we choose $n = 3$ in our experiments.

3.1 Statistical Parameter Estimation

We model the distribution of people by fitting the data samples of people with multi-dimensional clusters. The idea of using multi-dimensional clusters to model the p.d.f may be traced back to the works in [12]. Traditional k-means clustering algorithms based on the Euclidean or the Mahalanobis distances [11, 13] work satisfactorily under Gaussian assumptions. However, if the actual distribution of the data is non-Gaussian, then traditional k-means may fail to yield satisfactory results. Hence, we use a modified k-means clustering algorithm that utilizes higher order statistics for improved clustering. The closeness measure that we use for clustering is given by (3). It was found, after some experimentation, that six clusters were adequate for our purpose. The data samples in the clusters are used to learn the mean, the covariance and the joint third-order statistics corresponding to each of the clusters.

3.2 Dynamic Background Learning

Given a test image, the background is learnt dynamically as follows. Initially, the test image is scanned at its highest resolution for image patterns that are not people. Since background patterns usually far outnumber people in a given test image, the trick is to use a loose threshold to separate background patterns based on the already available statistical knowledge of the people. Naturally, a loose threshold will not capture all the background patterns. However, since the background usually constitutes a major portion of the test image, it is possible to get sufficient number of samples that are not people. These patterns are next distributed into six clusters using the HOS-based closeness measure and the statistical parameters corresponding to each of the six clusters are estimated.

3.3 Locating People

The test image is searched for the presence of people at all points in the image and across different scales by using the HOS-based measure given by (3). A vector of difference measurements of the test pattern is computed with respect to each cluster (six corresponding to people and six corresponding to the background) using the HOS-based closeness measure. If the minimum difference value corresponds to that of a people cluster and is less than a specified threshold, the test pattern is declared as belonging to the 'people class', else not. The above condition helps to reduce the false alarms considerably. Knowledge of the background allows us to relax the threshold which in turn leads to an improvement in the people detection rate while simultaneously keeping down the number of false matches.

4 Detection in Video

In most practical situations, it is moving objects in a scene that are primarily of interest. When an image sequence or video data is available, motion-based information can be exploited to achieve better accuracy and speed as the search can be now restricted to areas in and around the motion regions only.

In our approach to detection and tracking of people in video, we first identify regions of the image that contain moving objects by a combination of thresholding and morphological operations. A discrimination module then ascertains whether the moving objects correspond to people in motion or not by using an HOS-based statistical classifier. Tracking a subject over subsequent frames is carried out by finding the closest match in the next frame based on certain constraints, such as continuity of position co-ordinates and the HOS-based difference measurement values of the object with respect to the clusters. Note that the HOS-based closeness measure constitutes an important component in both our detection and tracking modules. We now describe each of these modules in brief.

4.1 Segmentation

Assuming a stationary camera, foreground objects are segmented from the background in each frame of the video sequence by frame differencing followed by thresholding. However, simple thresholding can result in incomplete extraction of a moving object, erroneous extraction of non-moving pixels, and legitimate extraction of illegitimate objects. Hence, morphological operations are used to reconstruct incomplete targets and to remove extraneous noise. A judicial combination of the erosion and dilation operations removes

isolated noise detections and connects the fragments of objects produced by the pixel-based detection decisions into contiguous motion regions. The result of the morphological operations is a binary image with the areas of motion identified. The illegitimate targets must be subsequently removed by the HOS-based classifier.

4.2 Object Discrimination

Once the motion regions are identified, the task is to determine which of the moving objects actually correspond to people in motion. This is carried out by statistical learning, image search and classification.

4.2.1 Statistical Learning

The learning process consists primarily of gathering statistical information about the class of people and the background. This is achieved as follows.

- An HOS-based k-means clustering algorithm is used on a training set comprising of images of people to derive information about the mean, the covariance and the joint higher order statistics (usually up to order 3) corresponding to each of the six clusters.
- Given the image frame and the knowledge-base of the people class, the statistics of the background (mean, covariance and HOS corresponding to each of the six clusters) is then learnt using the dynamic background learning algorithm described in Section 3.2.

4.2.2 Image Search

In this step, we search for the presence of people in and around the detected motion regions. A window of suitable size is chosen about each of these regions. The area within the window is then searched for possible target at all points and across different scales. The size of the window is not very critical. For each test pattern within the window, a vector of HOS-based difference measurements of this pattern is computed with respect to each of the 12 clusters using equation (3) and the available statistical knowledge-base of the people and the background. The vector of difference measurements corresponding to each test pattern is then passed on to the classifier along with the centroidal locations of the test patterns.

4.2.3 Classification

The classification procedure is as follows.

Step 1. Based on the vector of difference measurements obtained from the search step, that test pattern x^* and the cluster i^* which result in a minimum difference value for a given motion region are determined as

$$x^*, i^* = \arg \min_{x, i, x \in C_i} - \log N(\underline{\mu_i}, R_i) \cdot \left(1 + \sum_{n=3}^{\infty} E\left[\underline{H}_n^T(R_i^{-\frac{1}{2}}(\underline{X} - \underline{\mu_i}))\right] \underline{H}_n(R_i^{-\frac{1}{2}}(\underline{x} - \underline{\mu_i}))\right).$$

In the above equation, C_i represents the i^{th} cluster and $1 \leq i \leq 12$. The first 6 clusters are used to model the class of people while the rest of the 6 clusters are used to model the background.

Step 2. The test pattern x^* is classified as belonging to the people class if the following two conditions are met:

a. The HOS-based difference measurement value is less than an optimally selected threshold value T_0.

b. The cluster corresponding to the minimum value of x^* belongs to the set of people clusters.

In other words, x^* belongs to the class of people class if

$$- \log N(\underline{\mu_{i^*}}, R_{i^*}) \left(1 + \sum_{n=3}^{\infty} E\left[\underline{H}_n^T(R_{i^*}^{-\frac{1}{2}}(\underline{X} - \underline{\mu_{i^*}}))\right] \underline{H}_n(R_{i^*}^{-\frac{1}{2}}(\underline{x^*} - \underline{\mu_{i^*}}))\right) < T_0,$$

and

$$1 \leq i^* \leq 6 .$$

Step 3. If the test pattern x^* belongs to the class of people, then the centroid of the test pattern along with the vector of HOS-based difference measurement values are passed onto the tracking module.

Step 4. The above steps are repeated for every motion region to check for the presence of people within each motion region.

In Step 2a above, the threshold T_0 is determined empirically from several frames.

4.3 Tracking

The tracking module must be capable of tracking multiple people against complex background. Most systems for target tracking are based on either the Kalman filter or the correlation technique. However, we describe here a scheme that uses centroidal locations in conjunction with the HOS-based difference measurement vector for tracking. The tracking procedure consists of the following steps.

Step 1. The centroid corresponding to a detected foreground region is compared with the centroids of the objects detected in the earlier frame using the simple Euclidean norm.

Step 2. If the difference in the displacements of the centroids is less than a certain threshold value, then correspondence is established.

Step 3. If there are multiple foreground regions that are likely candidates for match with an object in the previous frame, then the average value of the HOS-based difference measurements is used to establish a unique correspondence.

The tracking method described above works satisfactorily as long as the positioning of the people is not very complex. Velocity estimates could be computed for the motion regions and used together with the locations of the centroids for improved performance. The method could fail under occlusion or when people cross each other.

5 Experimental Results

We present results on the performance of the proposed HOS-based people detection system for both static and video image data. The training set consisted of about 500 grey images of people, each of dimension 16×32 pixels. The system was tested on real images with complicated background. The training set was distinct from the test set.

Figure 1 shows the output results for some test static images. Each detected object is represented by drawing a box around it. Multiple boxes represent detection at different scales. Considering the complexity of the problem which lies in being able to detect people under different lighting conditions and positional orientations, we note that the method is able to detect people quite well in all the images.

The method was next tested on real video data comprising of many people moving around in a parking lot. The image sequence was captured with a stationary camera. Each of the detected objects is represented by a box with a certain grey level. The grey value of the box is an identity assigned to a detected object so that the correctness of the tracking algorithm can be deciphered. Figure 2 shows the output results corresponding to some of the frames in the image sequence. Note that all the people have been detected and correctly tracked in each of the frames. Even those with non-frontal views are detected. The potential of the HOS-based closeness measure is quite evident from the results. The system is quite capable of detecting and tracking multiple people in motion in natural outdoor lighting conditions, while simultaneously being able to reject background clutter.

6 Conclusions

We have described a scheme for detecting and tracking multiple people against a cluttered background in image sequences. The method effectively combines statistical information about the target object with spatio-temporal information for classification and tracking. It derives the higher order statistics from data samples of people to get a better approximation to the distribution of people. Motion information is used to localize moving objects. The background is learnt dynamically while testing. The detection module uses an HOS-based classifier to determine which of the moving objects actually correspond to people in motion. Detected people are tracked over subsequent frames using position co-ordinates and the HOS-based difference measurement of the target object. The system successfully detects and tracks multiple people, even against complex backgrounds. The algorithm is also robust to orientation, changes in scale, and lighting conditions.

There are several directions that we are pursuing to improve the performance of the system and to extend its capabilities. When the motion regions become too many, the task of isolating them can be quite difficult. We are currently working on this issue. Investigations are also on to improve the tracking module to handle situations such as temporary occlusions. Derivation of the bounds on the error in classification is underway. We hope to extend this scheme to moving platforms in conjunction with image stabilization algorithms.

References

[1] K. Akita, "Image sequence analysis of real world human motion", *Pattern Recognition*, vol. 17, pp. 73-83, 1984.

[2] T. Tsukiyama and Y. Shirai, "Detection of the movements of persons from a sparse sequence of TV images", *Pattern Recognition*, vol. 18, no. 3, pp. 207-213, 1985.

[3] M.K. Leung and Y.H. Yang, "Human body motion segmentation in a complex scene, *Pattern Recognition*, vol. 20, no. 1, pp. 55-64, 1987.

[4] K. Rohr, "Incremental recognition of pedestrians from image sequences", *Proc. Computer Vision and Pattern Recognition*, 1993, pp. 8-13.

[5] C.R. Wren, A. Azarbayejani, T. Darrell and A.P. Pentland, "Pfinder: Real-time tracking of the human body", *IEEE Trans. Pattern Anal. and Machine Intell.*, vol. 19, no. 7, pp. 780-785, 1997.

[6] I. Haritaoglu, D. Harwood and L.S. Davis, "W^4: Who? When? Where? What?", *Intl. Conf. on Face and Gesture Recognition*, (Nara, Japan), 1998.

[7] T. Olson and F. Brill, "Moving object detection and event recognition algorithms for smart cameras", *Proc. DARPA Image Understanding Workshop*, May 1997, pp. 159-176.

[8] C.P. Papageorgiou, M. Oren and T. Poggio, "A trainable system for people detection", *Proc. DARPA Image Understanding Workshop*, 1997, pp. 207-214.

[9] T. Kanade, R.T. Collins, A.J. Lipton, P. Burt and L. Wixson, "Advances in co-operative multi-sensor video surveillance", *Proc. DARPA Image Understanding Workshop*, (Monterey, California), Nov. 1998, pp. 3-24.

[10] A.J. Lipton, H. Fujiyoshi and R.S. Patil, "Moving target classification and tracking from real-time video", *Proc. DARPA Image Understanding Workshop*, (Monterey, California), Nov. 1998, pp. 129-136.

[11] R. O Duda and P. E. Hart, *Pattern Classification and Scene Analysis*, John Wiley & Sons Inc., 1973.

[12] K. Sung and T. Poggio, "Example-based learning for view-based human face detection", *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 20, pp. Jan. 98.

[13] A. K. Jain and R. C. Dubes, *Algorithms for Clustering Data*, Prentice-Hall Inc., Englewood Cliffs, 1988.

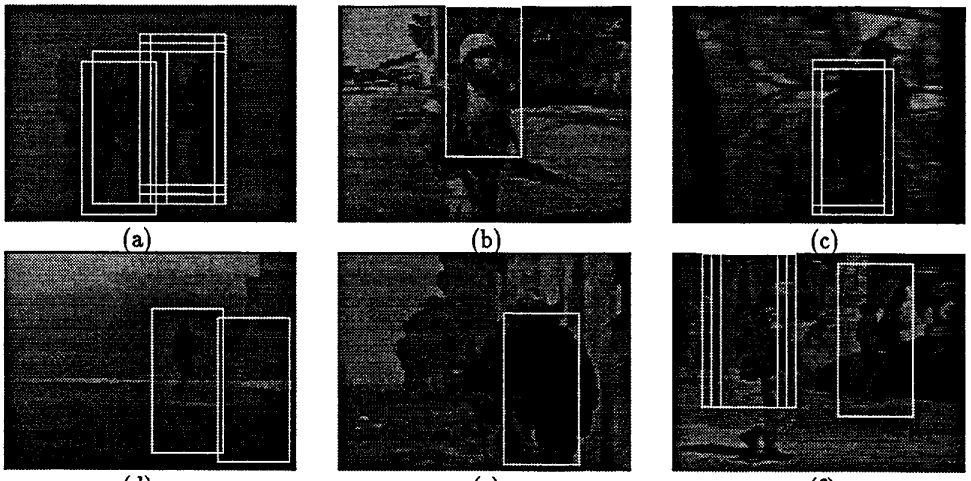

Figure 1. Representative results on static imagery using the proposed scheme.

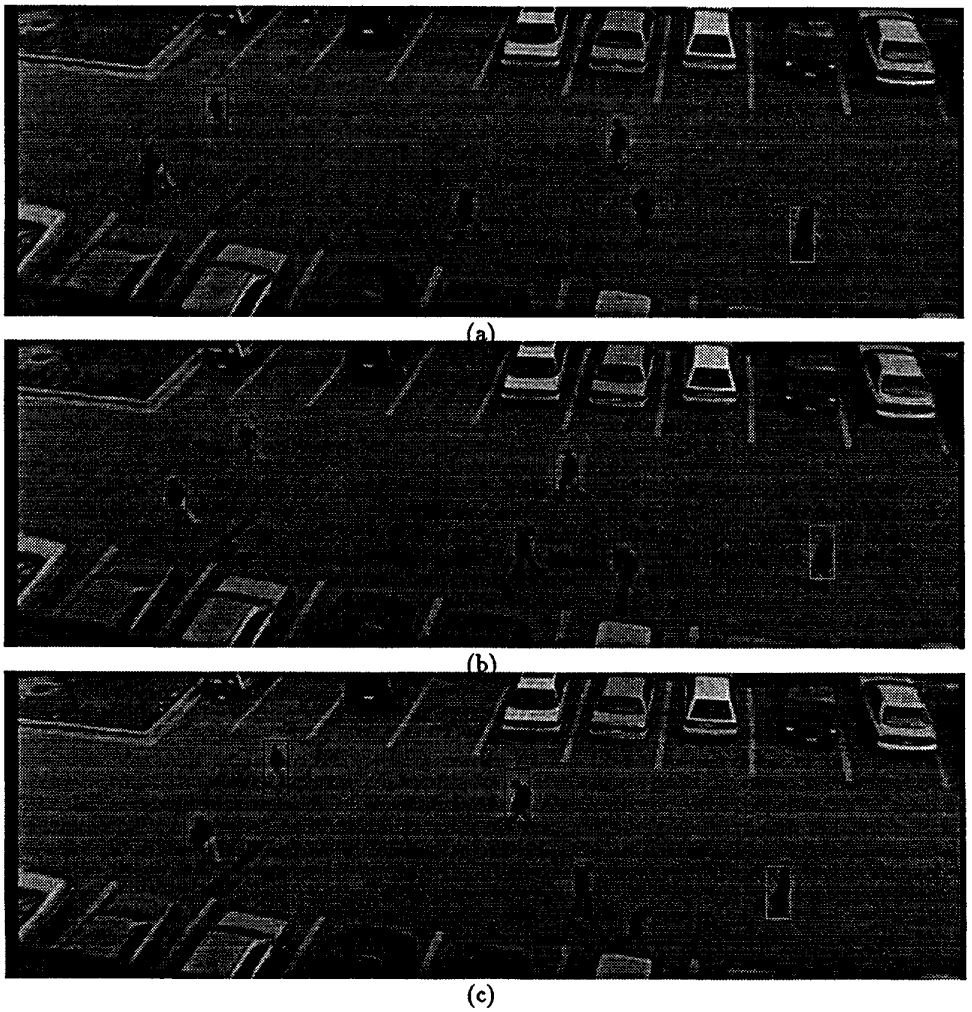

Figure 2. Detection results for video data. (a,b,c) Frames 5, 25, and 45, respectively.