Beyond Static Detectors: A Bayesian Approach to Fusing Long-term Motion with Appearance for Robust People Detection in Highly Cluttered Scenes

Jianguo Zhang and Shaogang Gong
Department of Computer Science
Queen Mary, University of London, London E1 4NS, UK
{*jgzhang, sgg*}*@dcs.qmul.ac.uk*

Abstract

In this work we present a framework for robust people detection in highly cluttered scenes with low resolution image sequences. Our model utilises both human appearance and their long-term motion information through a fusion formulated in a Bayesian framework. In particular, people appearance is modeled by histograms of oriented gradients. Motion information is computed via an improved background modeling by spatial motion constrains. Experiments demonstrate that our method reduces significantly the false positive rate compared to that of a state of the art human detector under very challenging conditions.

1 Introduction

Pedestrian detection in a busy public scene is an important yet challenging task in visual surveillance. The difficulties lie in modelling both object and background clutter contributed by a host of factors including changing object appearance, diversity of pose and scale, moving background, occlusion, imaging noise, and lighting change. Usually pedestrians in public space are captured by two dominant visual features: *appearance* and *motion*. There is a large body of work in human detection, see [2] and [3] for a survey. They can be categorized into two groups: static and dynamic people detectors. Static people detectors rely mainly on finding robust appearance features that allow human form to be discriminated against a cluttered background using a classifier such as SVM or AdaBoost searching through a set of sub-images by a sliding window. Typical features include rectified Haar wavelets [4], rectangular features [5], and SIFT (Scale Invariant Feature Transform) like features such as histogram of oriented gradients [1]. Papageorgiou et al. [4] described a pedestrian detector based on SVM using Harr wavelet features. Gavrila and Philomin [6] presented a real-time pedestrian detection system by utilizing silhouettes information extracted from edge images. The candidate of the silhouettes is selected as the one with the smallest chamfer distance to a set of learned human shape examples. On the other hand, there is little progress on dynamic detectors, although the idea of using pure motion information for human pattern recognition is not new [7, 8, 9]). Most existing work utilises optic flow. Viola et al. [5] proposed a very efficient detector using Adaboost that can achieve real-time performance. The rather simple rectangular features and the cascade structure account for the efficiency of this approach. Motion information was also taken into account through a coarse estimation of optic flow between two consecutive frames. To achieve satisfactory performance, this approach assumes that the human motion information in the test sequences is similar to those in the training set. Other related work using motion information includes human behavior recognition by distribution of 3D spatial-temporal interest points [10, 11], 3D volumetric features [12], or through 3D correlation [13]. Overall, existing methods for computing motion assume mostly that the motion is locally smooth. However this is untrue especially in busy public scenes when measuring optic flow is sensitive to noise and unreliable due to lighting change, reflection, moving background such as tree leaves (see Figure 2).

To date, work on utilising both motion and appearance information remains in its infancy. To our best knowledge, there is no work performing direct people detection using both appearance and long-term motion information. In this work, we present a robust framework for people detection in highly cluttered public scenes by utilizing both human appearance and their long-term motion information whilst reliable optic flow cannot be estimated. Our method does not require the estimation of continuous motion such as optic flow in training thus reduces the number of features required for training a classifier. It allows for any detected appearance hypothesis to be verified using a long-term motion history analysis. We show experimental results that demonstrate the efficacy and robustness of the proposed approach against that of a state of the art static people detector.

2 Methodology

In contrast to video sequences captured under well-controlled environment at frame rate, our task for people detection requires to work in highly cluttered public scene (underground) given low resolution data and low frame rate. The scene also suffers from 1) significant lighting changes, which makes the motion estimation unstable and noisy; 2) heavy occlusions, which requires the people detector to handle partial match; 3) extensive background clutters, which can cause high false alarms. To this end, we propose a robust people detection method for video sequences by fusing static appearance feature based detector with a long-term motion based attention confidence measure. An overview of our method is shown in Figure 1.

2.1 Generating static appearance hypothesis

We adopt the static people detector proposed by Dalal and Triggs [1] to generate static human presence hypothesis in each frame. To achieve scale invariance, this detector utilizes a multi-scale sliding window approach, i.e., scanning each frame at each scale level through a pyramid decomposition. Each sub-window image patch centered at location i (denoted by v_i, where $i = 1 : n$ and n is the number of patches) is transformed into a feature vector before classified into either human foreground or scene background by a classifier. The feature vector used here is a SIFT [14] like feature based on histogram of gradient orientation. The basic idea is that local object appearance and shape can often be characterized rather well by the distribution of local intensity gradients or edge directions, even without any precise knowledge of corresponding gradient or edge positions (similar work can be found in [15] using histograms of scale normalized, oriented derivatives to detect and recognize arbitrary object classes). The size of the detection window is 32×64 including 8 pixels of margin beyond the window size. A linear SVM is used as the classifier and the output of the classifier serves as the confidence measure for our static human appearance hypothesis. This approach has achieved very good detection rate in static images of outdoor scene [1], e.g. image samples from the MIT pedestrian dataset [4]. However, the lack of motion information makes this detector less robust to background clutters. This problem becomes severe in cluttered scenes with poor lighting, such as in public underground, when such a static human detector gives unacceptable false alarm rate, as shown by examples in Figures 5 and 6. Simply increasing the threshold of the score for generating the hypothesis does not result in reducing the false alarms because in such cluttered scenes, regions in background have very similar appearance to that of people, e.g. as shown in Figure 5 (e). Here the false alarms on the wall have very high scores produced by the classifier and do indeed look like standing people. Similar observations can be found in the example shown in Figure 6 (b). Given that in any public scene, people exhibit inevitably long-term-moving patterns instead of just a static pattern, we consider a detection model based on fusing detected static human presence hypothesis with their long-term motion history information as follows.

2.2 Motion confidence map

One way to utilize motion information is to compute optic flow [16, 17, 18]. However, optic flow estimation makes a strong assumption that motions are only caused by either relative movement between the camera and the object of interest or ego-motion. The accuracy of flow estimation is based on well sampled data, i.e. local smoothness. However, both assumptions are not usually satisfied. First, large lighting changes usually result in noisy flow field. Second, relatively fast action w.r.t the camera, i.e. low frame rate, also results in highly discontinuous motion which is far from smooth. Examples of estimating optic flow in a underground scene are shown in Figure 2. The optic flow was computed using a robust method proposed by Gautama et al. [18]. However, it is evident that the resulting flow field is very noisy and unstable. To address this problem, in this work we adopt an alternative long-term motion estimation approach using background extraction and subtraction, given that most surveillance CCVT systems are based on fixed views. More precisely, we utilise a Gaussian mixture background model of [19]:

$$b(x, y) = \sum_i \alpha_i g(f(x, y), \theta_{i,x,y}, \sigma_{i,x,y}), \tag{1}$$

where x, y is the location of each pixel, $(\theta_{i,x,y}, \sigma_{i,x,y})$ are the model parameters of each individual Gaussian components g, and $f(x, y)$ is the local pixel intensity. The variation of one frame $f(x, y)$ with respect to the background model is estimated as the probability distance given by

$$v(x, y) = \sum_i \alpha_i exp\left(-1/2\left(f(x, y) - \theta_{i,x,y}\right)^2/\sigma_{i,x,y}{}^2\right) \tag{2}$$

This type of motion information is very effective at highlighting changes in motion in the scene. However, this is also an undesirable property since the noisy motion caused by lighting changes is inevitably augmented. See Figure 5 (b) as an example. To suppress the noisy motion caused by lighting changes, we further take the spatial motion contrast into consideration in the Gaussian mixture model as follows:

$$v(x, y) = exp\left(-\frac{1}{2}\frac{(f(x, y) - b(x, y))^2}{\sigma_s^2}\right) \tag{3}$$

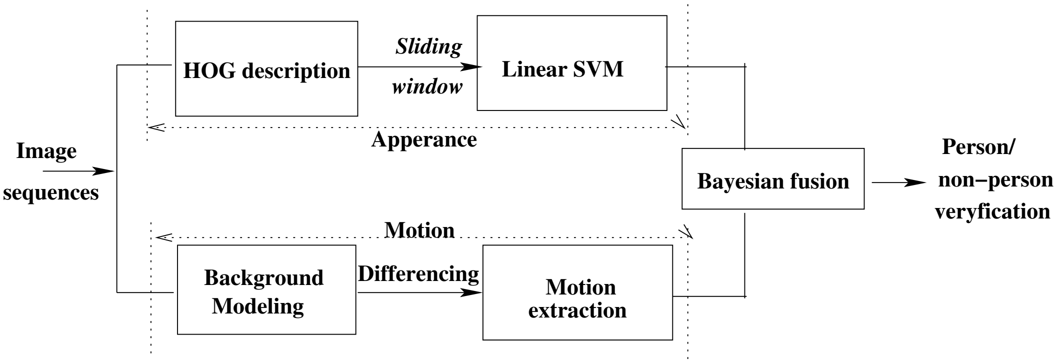


Figure 1: Flow chart of our method for pedestrian detection. An appearance based detector is used to create the initial hypothesis and long-term motion is computed by background modeling. The above cues are combined in a Bayesian framework. The final candidates are selected by thresholding.

In the background model of Eq.(1), $\sigma_{i,x,y}$ is the strength of the motion of each pixel at (x, y), we calculate σ_s here as the mean or median of $\sigma_{i,x,y}$. Examples of motion extraction using this model are shown in Figure 5 (b) and (c), where in (b) motion was estimated using the Gaussian mixture background model without considering spatial motion contrast whilst in (c), it was taken into account. This demonstrates clearly the effectiveness of utilising the spatial motion contrast measure given by Eq.(3) for removing motion noise as compared to existing Gaussian mixture models.

2.3 Bayesian verification

For each hypothesis created by the static people detector, we verify its truthfulness using the obtained motion information. More precisely, for each detected bounding box hypothesis, we first calculate a Gaussian weighted accumulated motion score. Suppose that we have a hypothesis located as (x_0, y_0), the motion score of this hypothesis is calculated as:

$$m(x_0, y_0) = \int\int_{x,y\in\Omega(x_0,y_0)} v(x, y)g(x, y)dxdy \quad (4)$$

where $\Omega(x_0, y_0)$ is the neighborhood of the hypothesis location. Here we set it as the window size of the hypothesis created by the static human detector. $g(x, y)$ is an anisotropic Gaussian envelope given as:

$$g(x, y) = \\ exp\left(\frac{1}{2}([x, y] - [x_0, y_0])\Sigma^{-1}([x, y]^T - [x_0 - y_0])^T\right) \quad (5)$$

where Σ is the spatial correlation matrix, we set as $\begin{bmatrix} 1/2w & 0 \\ 0 & 1/2h \end{bmatrix}$, where w, h are the width and height of the hypothesis window. Thus we give higher weight to the contribution of motion near the centre of the hypothesis window, whilst less emphasis to further away from the centre. This is aimed to improve the robustness of the calculation of the motion score for each hypothesis against any error in the location of the hypothesis. Examples of measuring the Gaussian weighting mask against the corresponding bounding boxes are shown in Figure 7.

We further define the total score for detecting the presence of a person as the product of the motion score and static hypothesis confidence measure

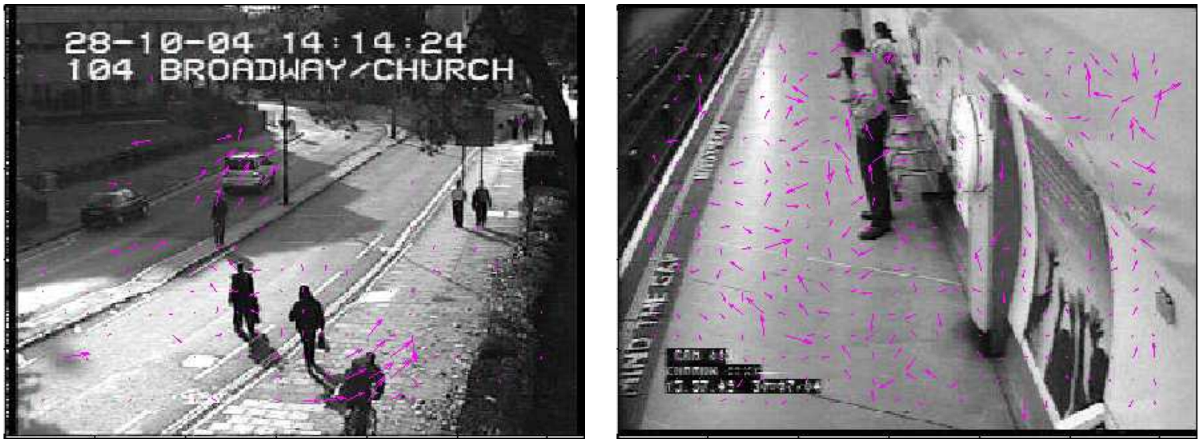
$$s(x_0, y_0) = m(x_0, y_0)\, c(x_0, y_0) \quad (6)$$

where $c(x_0, y_0)$ is the confidence measure of the static human detector at (x_0, y_0).

The verification process can be formulated by the Bayesian rule. For a bounding box hypothesis, we wish to find the probability of the presence of an object given motion confidence m and appearance measure c, $p(o|c, m, h)$, which is given by the Bayesian rule as follows:

$$p(o|c, m, h) = p(m|h, o)\, p(c|h, o) \quad (7)$$

Here we assume that the motion m and the appearance c are conditionally independent. $p(m|h, o)$ is the motion confidence within the hypothesis bounding box given the object, which is computed using Eq.(3). $p(c|h, o)$ is the appearance confidence measure generated by the static object detector. The final candidates are selected by thresholding $p(o|c, m, h)$.



(a) (b)

Figure 2: Optic flow estimation comparison in different sequences. (a) regular flow on well-captured outdoor sequence; (b) noisy flow on a real underground sequence. Note that in (b), the upper-right corner has some distinct noisy optic flow caused by lighting changes and object reflections.

3 Experimental results

3.1 Data set

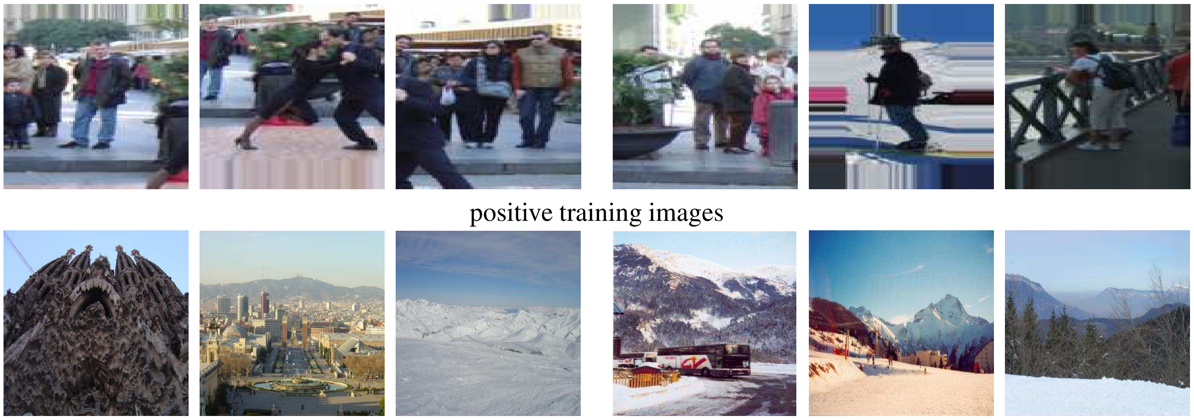
Training set: We use the challenging INRIA [1] image dataset for training, which is totally independent to our test video data. The INRIA dataset also does not contain any motion information. It contains 607 positive training images, together with their left-right reflections (1214 images in all). A fixed set of 12180 patches are randomly sampled from 1218 person-free training images. Examples of these images are shown in Figure 3.
Test set: Out test set contains image frames from CCTV video sequences taken from underground stations and platforms by fixed cameras. One set of images is from a train platform containing 3710 frames. The other set is from a ticket office area containing 160 frames. In contrast to other video sequences captured under well-controlled environment, these sequences present significant lighting changes and background clutter. Many frames contain multiple people under severe occlusions. Figure 4 shows 6 consecutive example frames from the platform scene.

3.2 Detection results

Examples of detections at the ticket office area scene from our dynamic detector are compared and shown against those from a static people detector [1] in Figure 5. The detected boxes are displayed on each frame. Figure 5 (d) shows detection results from the static detector whilst Figure 5 (e) and (f) show the detections of our dynamic detector. Figure 5 (f) is an improved version of Figure 5 (e) by increasing the robustness of motion estimation using spatial contrast measure given in Eq.(3). Thus it further removes additional false alarms. Another detection examples from the sequences of the underground ticket office are shown in Figure 6. The dataset presents a lot of background clutters where ticket machines appear to be similar to the appearance of people standing there. So it is reasonable for the static detectors to produce false detections at where some ticket machines are located, as shown in Figure 6 (b). Our dynamic detector has shown to be able to remove those false alarms.

To quantify the detection performance, we also perform an quantitative evaluation of both the dynamic detector and static detector on the ticket-office scene from which manually labelled ground truth was available. By varying the threshold of the detection scores one at a time, we obtain the Receiver Operating Characteristics (ROC) curves of those detectors, showing false positive rate versus true positive rate (see Figure 8). When comparing with the ground truth annotation, we measured the overlap score between the detected bounding box and ground true bounding box. A detection with overlap score larger than 50% is labeled as a 'match'. For further explaining the role of motion information played in improving the detection rate, we also plotted the ROC curve of a pure motion detector, i.e. the bounding boxes are weighted only by their motion information as computed by Eq.(3). From this experiment, it is clear that the motion information plays a critical role in accurate detection. Simply using the motion information alone also gave good detection as shown by the ROC curve. This is because most of the motions in this particular scene were caused actually by human movement. The ROC curve shows that our dynamic detector improves significantly the

[1] The dataset may be downloaded from http://lear.inrialpes.fr/data.



positive training images

negative training images

Figure 3: Positive and negative training examples used in our experiments.

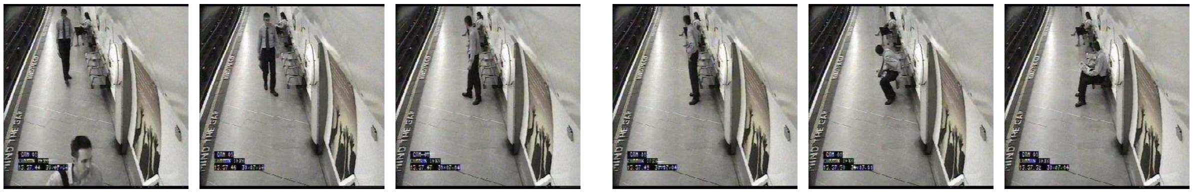


Figure 4: Six consecutive frames from the test sequences.

performance of the static detector by Dalal and Triggs [1]. The false alarms rate has been greatly reduced. For example, to achieve a detection rate of 70% on the ticket-office scene, our detector produces 130 false alarms whilst the detector by Dalal and Triggs generated 370 false alarms, over 3 times more.

4 Discussion and conclusions

In this paper, we presented a framework for robust people detection in highly cluttered scenes with low resolution image sequences. Our model utilises both human appearance and their long-term motion information through a fusion formulated in a Bayesian framework. In particular, people appearance is modeled by histograms of oriented gradients. Motion information is computed via an improved background modeling by spatial motion constrains. This is an extension of the static detector proposed by Dalal and Triggs. Experiments demonstrate that our method reduces significantly the false positive rate compared to that of the state of the art static human detector under very challenging conditions. At present, our model is based on the long-term motion information, and requires fixed camera viewpoint during detection. Building a hybrid model of both long-term and short-term motion information could possibly give more robust detections and also be adaptive to some background and viewpoint change.

References

[1] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” in *International Conference on Computer Vision & Pattern Recognition*, vol. 2, June 2005, pp. 886–893.

[2] D. M. Gavrila, “The visual analysis of human movement: A survey,” *Computer Vision and Image Understanding*, vol. 73, no. 1, pp. 82–98, 1999.

[3] R. Cutler and L. Davis, “Robust real-time periodic motion detection: Analysis and applications,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 781–796, 2000.

[4] C. Papageorgiou and T. Poggio, “A trainable system for object detection,” *International Journal of Computer Vision*, vol. 38, no. 1, pp. 15–33, 2000.

[5] P. Viola, M. J. Jones, and D. Snow, “Detecting pedestrians using patterns of motion and appearance,” *International Journal of Computer Vision*, vol. 63, no. 2, pp. 153–161, 2005.

[6] D. Gavrila and V. Philomin, “Real-time object detection for “smart“ vehicles,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 1999, pp. 87–93.

[7] G. Johansson, “Visual perception of biological motion and a model for its analysis,” *Perception and Psychophysics*, vol. 14, pp. 201–211, 1973.

[8] D. D. Hoffman and B. E. Flinchbaugh, “The interpretation of biological motion,” *Biological Cybernetics*, pp. 195–204, 1982.

[9] R. C. Aswin C Sankaranarayanan and Q. Zheng, “Tracking objects in video using motion and appearance models,” in *IEEE International Conference on Image Processing*, 2005.

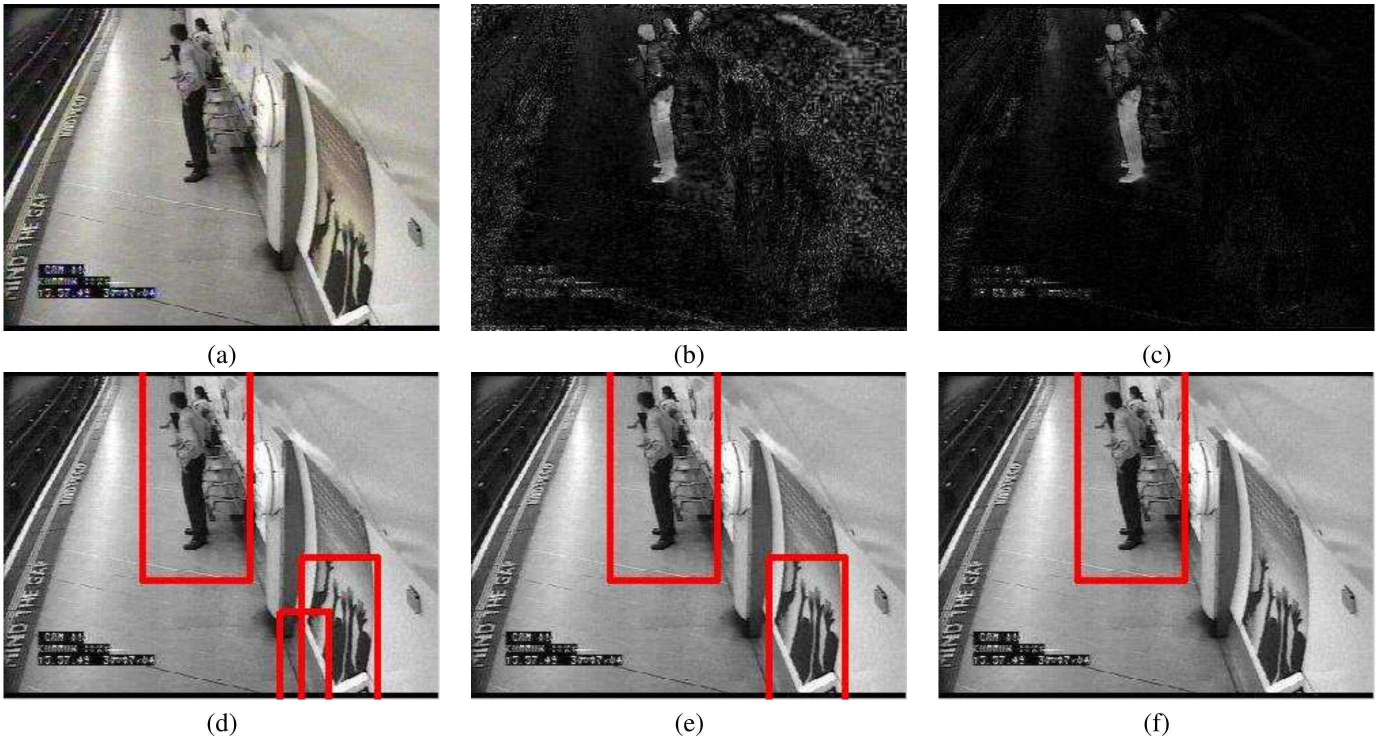


Figure 5: Examples of human detection in each step. (a) the original slice; (b) initial motion confidence map only using Gaussian mixture; (c) refined motion confidence map; (d) initial hypothesises by using pure static human detector; (e) detection results by using the motion map of (b); (f) refined detection results by using the motion map of (c).

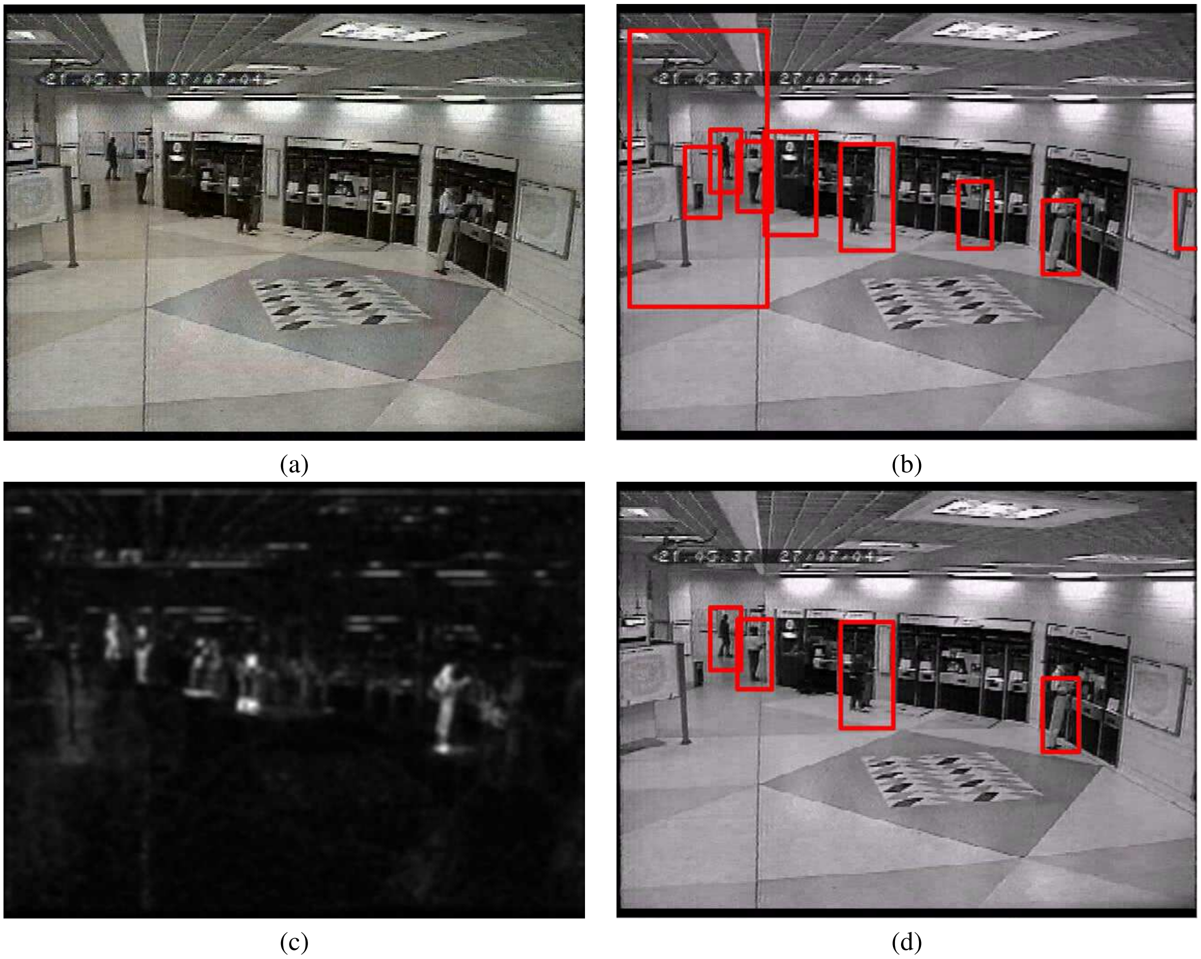


Figure 6: Examples of human detection in the ticket office with heavy background clutter. (a) the original slice; (b) initial hypothesises use pure static human detector; (c) motion confidence map; (d) detection results using the motion map of (c).

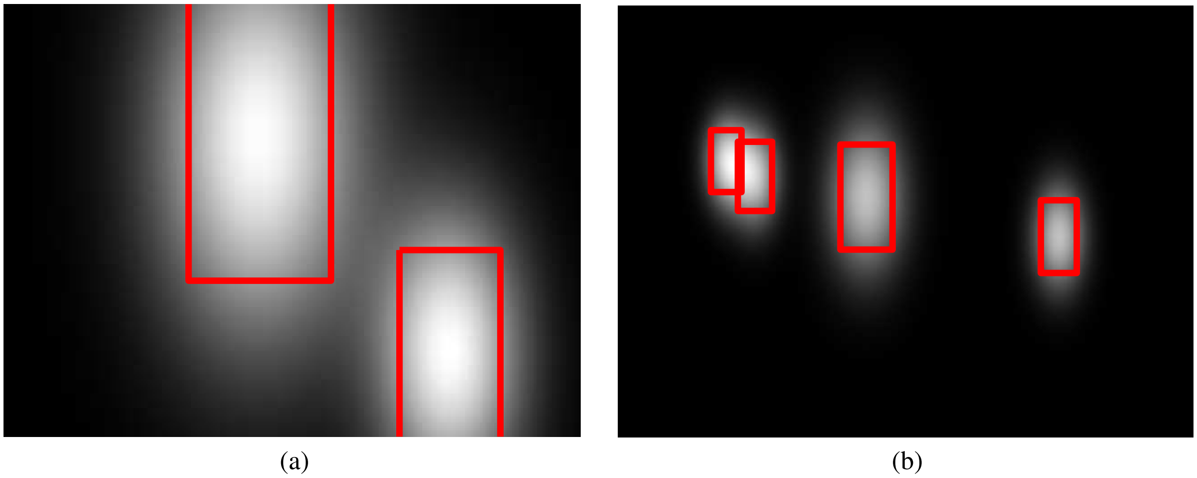


Figure 7: Gaussian weighting masks corresponding to the hypothesises bounding box. (a) corresponds to Figure 5(e); (b) corresponds to Figure 6(d).

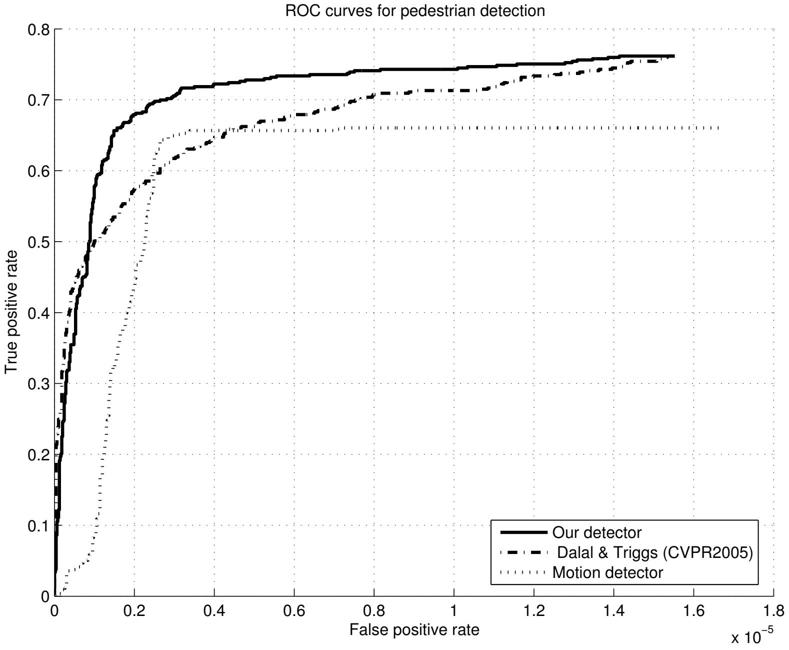


Figure 8: ROC curves on the sequences of the underground ticket office for both the dynamic and static people detectors as well as a pure motion based people detector

[10] C. Schuldt, I. Laptev, and B. Caputo, “Recognizing human actions: A local svm approach,” in *International Conference on Pattern Recognition*. Cambridge, UK, 2004, pp. 32–36.

[11] I. Laptev, “On space-time interest points,” *International Journal of Computer Vision*, vol. 64, no. 2, pp. 107–123, 2005.

[12] Y. Ke, R. Sukthankar, and M. Hebert, “Efficient visual event detection using volumetric features,” in *International Conference on Computer Vision*, 2005, pp. 166–173.

[13] O. Boiman and M. Irani, “Detecting irregularities in images and in video,” in *International Conference on Computer Vision*, 2005, pp. 462–469.

[14] D. G. Lowe, “Distinctive image features from scale-invariant keypoints,” *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[15] B. Schiele and J. Crowley, “Recognition without correspondence using multidimensional receptive field histograms,” *International Journal of Computer Vision*, vol. 36, no. 1, pp. 31–50, 2000.

[16] B. Horn and B. Schunck, “Determining optical flow,” *AI Memo 572,Massachusetts Institue of Technology*, 1980.

[17] E. P. M. Proesmans, L. Van Gool and A. Oosterlinck, “Determination of optical flow and its discontinuities using non-linear diffusion,” in *European Conference on Computer Vision*, vol. 2, 1994, pp. 295–304.

[18] T. Gautama and M. Van Hulle, “A phase-based approach to the estimation of the optical flow field using spatial filtering,” *IEEE Transactions on Neural Networks*, vol. 13, pp. 1127–1136, May 2002.

[19] T. Xiang and S. Gong, “Beyond tracking: Modelling activity and understanding behaviour,” *International Journal of Computer Vision*, pp. 195–204, 2006, to appear.