Image and Vision Computing 27 (2009) 437–443

Contents lists available at ScienceDirect

Image and Vision Computing

journal homepage: www.elsevier.com/locate/imavis

People detection in low-resolution video with non-stationary background

Jianguo Zhang [a,*], Shaogang Gong [b]

[a] School of Electronics, Electrical Engineering and Computer Science, Queen's University Belfast, Belfast BT7 1NN, UK
[b] Department of Computer Science, Queen Mary University of London, London E1 4NS, UK

ARTICLE INFO

Article history:
Received 12 November 2006
Received in revised form 25 May 2008
Accepted 28 June 2008

Keywords:
Visual surveillance
People detection
Bayesian fusion
Long-term motion
AdaBoost

ABSTRACT

In this paper, we present a framework for robust people detection in low resolution image sequences of highly cluttered dynamic scenes with non-stationary background. Our model utilizes appearance features together with short- and long-term motion information. In particular, we boost Integral Gradient Orientation histograms of appearance and short-term motion. Outputs from the detector are maintained by a tracker to correct any misdetections. A Bayesian model is then deployed to further fuse long-term motion information based on correlation. Experiments show that our model is more robust with better detection rate compared to the model of Viola et al. [Michael J. Jones Paul Viola, Daniel Snow, Detecting pedestrians using patterns of motion and appearance, International Journal of Computer Vision 63(2) (2005) 153–161].

1. Introduction

Pedestrian detection in a busy public scene is a challenging task. 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. The problem is made harder still if the camera is placed at a distance from the scene resulting in lack of pixel details on the objects of interest. There exists a large body of work in people detection which can be broadly categorized into two groups: static and dynamic detectors [10,5]. Static people detectors [6] rely mainly on finding robust appearance features that allow human form to be discriminated against a cluttered background. This is combined with a classifier, such as SVM or AdaBoost, to search through a set of sub-images using a sliding window, or alternatively using probabilistic geometrical voting based on the local appearance features detected on the object [16].

Popular static appearance features include rectified Haar wavelets [18], rectangular features [25], and a family of SIFT (Scale Invariant Feature Transform) [15] like features such as histogram of oriented gradients [6,14]. Papageorgiou et al. [18] described a pedestrian detector using SVM classification of Haar wavelet features. Gavrila and Philomin [9] presented a pedestrian detection system by utilizing silhouettes information extracted from edge images. Viola et al. [20] proposed a detector using cascaded AdaBoost of rather simple but effective rectangular local template features. SIFT features have also been shown to be promising for pedestrian detection in static images [6].

Comparatively there is much less work on dynamic detectors, although the basic concept of using motion information for human pattern recognition is not new [13,11]. Current dynamic detectors rely typically upon short-term motion by estimating optic flow [7]. However, computing optic flow is both expensive and highly sensitive to environmental noise. Alternatively, Viola et al. [25] extended their static detector by directly boosting short-term local motion features. The model also assumes that human motion patterns in a test sequences are similar to those in the training set. However, human motion is more diverse than its appearance and it is very difficult to collect a training set covering exhaustively all possible motion styles in different temporal scales. Moreover, existing methods for computing short-term motion assume mostly that the motion is locally smooth. However this is untrue especially in busy public scenes when apparent motion is subject to lighting change, reflection, moving background such as tree leaves. Consequently such short-term motion based models are prone to false positives and misdetections due to non-stationary background clutter and diversity in human motion style.

In this work, we present a framework for robust people detection in highly cluttered public scenes with non-stationary background by utilizing both human integral gradient appearance and their long-term motion information. In particular, we utilize SIFT-like features in an integral framework that selects automatically histogram features at different scales. We compare the experimental results of this representation with that of the features selected by the models of Viola et al. [20,25]. In addition, our model does not require the estimation of continuous motion

* Corresponding author. Tel.: +44 028 9097 4783.
E-mail addresses: j.zhang@ecit.qub.ac.uk, jgzhang@dcs.qmul.ac.uk (J. Zhang), sgg@dcs.qmul.ac.uk (S. Gong).

0262-8856/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.imavis.2008.06.013

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 Viola et al.

2. Integral orientation histogram and short-term motion

Our aim is to select robust appearance and motion features suitable for highly cluttered scenes with non-stationary background. For appearance, we consider SIFT-like linear appearance features due to its success in object detection and categorization [26].

Orientation of local image gradients can be sampled by a set of bins. Similar to SIFT [15], we consider 8 bins. Furthermore, gradient orientation at each pixel location can be weighted by its gradient magnitude. A weighted gradient orientation histogram for each pixel location of kth bin $(k \in 1,2,\ldots 8)$ at scale s is then defined as

$$h(x,y,s,k) = \frac{1}{Z}\oint_R \|\nabla I\|\omega(x,y,k) \tag{1}$$

where ∇I is the gradient vector, i.e., (I_x, I_y), R is the size of the local support region as a function of scale at pixel point (x,y), Z is a normalization factor, and $\omega(x,y,k)$ is an orientation counter for the kth bin denoted as

$$\omega(x,y,k) = \begin{cases} 1 & \text{if } \theta(x,y) \in k\text{th } bin \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

where $\theta(x,y)$ is the gradient orientation computed by $arctan(I_y/I_x)$. $h(x,y,s,k)$ can be computed very efficiently by constructing an integral orientation image similar to that of an *Integral Image* introduced by Viola et al. [25]. Thus the orientation histograms across different scales can be computed by four array references operated on integral orientation images, which corresponds to perform convolution of the simple rectangular filters on the original orientation images. In contrast to using SIFT-like appearance features at a fixed grid scale by Dalal and Triggs [6], this integral operation facilitates a single operation for computing local SIFT-like features at different scales. Examples of gradient orientation histograms computed from pedestrian images along with the filter used to compute the histograms in each orientation image are shown in Fig. 1.

Short-term motion information can be computed using either optic flow or frame differencing. For simplicity and cost-effectiveness, we adopt frame differencing for estimating short-term motion similar to that of [25]. To that end, five different frame-differenced images: Δ, U, L, R, D were generated using different filters during frame-differencing.

$$\begin{aligned} \Delta &= \text{abs}(f_t - f_{t+1}) \\ U &= \text{abs}(f_t - f_{t+1} \uparrow) \\ D &= \text{abs}(f_t - f_{t+1} \downarrow) \\ R &= \text{abs}(f_t - f_{t+1} \rightarrow) \\ L &= \text{abs}(f_t - f_{t+1} \leftarrow) \end{aligned} \tag{3}$$

where f_t, f_{t+1} are feature images (could be gradient images) in time, $\{\uparrow, \downarrow, \rightarrow, \leftarrow\}$ are image shift operators. This is aimed to capture motion in different directions. Examples of such 5 motion images computed for two different pairs of input frames are shown in Fig. 2.

In general, orientation histogram can be computed on *any* type of images. Thus, a motion orientation histogram can also be considered as an extension to the appearance gradient orientation histogram representation. To that end, we created and summarized several types of pedestrian features on these images. Let $I(\cdot)$ denotes the feature extraction by the *integral* operation on images. Thus, those features can be computed by the following equations, respectively:

$$AI \stackrel{\text{def}}{=} I(f(x,y,t)) \tag{4}$$

$$AG \stackrel{\text{def}}{=} I(\nabla f(x,y,t)) \tag{5}$$

$$MI \stackrel{\text{def}}{=} I(m), m \in \{\Delta, U, L, R, D\} \tag{6}$$

$$MG \stackrel{\text{def}}{=} I(\nabla(m)), m \in \{\Delta, U, L, R, D\} \tag{7}$$

More specifically, *AI* denotes the Haar-like linear appearance features as used in [20], which are computed from intensity images. *AG* represents histograms of appearance gradient orientations calculated by Eq. (1). *MI* is short-term motion information obtained by direct image differencing as used in [25]. *MG* is motion gradient orientation features and computed on differenced images as in Eq. (3). We further denote a combination of AI and MI as $AI + MI$ (the + operator represents a union of features rather than a summing of the histograms), a combination of histograms of gradient orientations plus short-term motion as $AG + MI$. In the following we use $AI, AG, MI, AI + AG, AG + MI, AG + MG$ to represent these types of features, respectively.

The importance of each feature type can be judged by the feature selection technique, e.g. AdaBoost in the context of pedestrian detection. Thus to obtain the best feature representation as well as some insights into the effectiveness of most of the features available today for pedestrian detection, we jointly boosted different types pedestrian features mentioned above, e.g., $AI + AG$. To achieve this, both the appearance and short-term motion features are then used as weak features for an AdaBoost detector. A detailed comparison as well as some discussions of each feature type is shown in Section 4.3.

The success of SIFT [15] suggests that features by histogram of orientations usually outperform appearance features. Our experiments confirm this, however, further show that this is not necessary true for the motion features. *MG* is comparable to *MI* at least in the context of using motion images produced by direct image

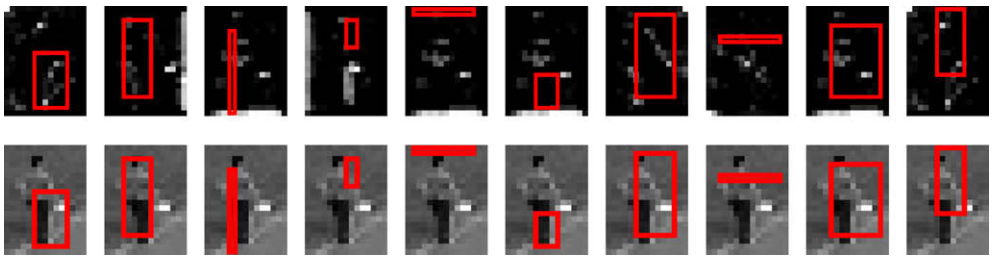

Fig. 1. Top 10 gradient orientation histogram (AG) features selected by AdaBoost when jointly boosted with Haar-like linear (AI) features. They are indicated by the selected rectangular filters (red boxes) superimposed onto the gradient orientation images (top row) and the intensity images (bottom row).

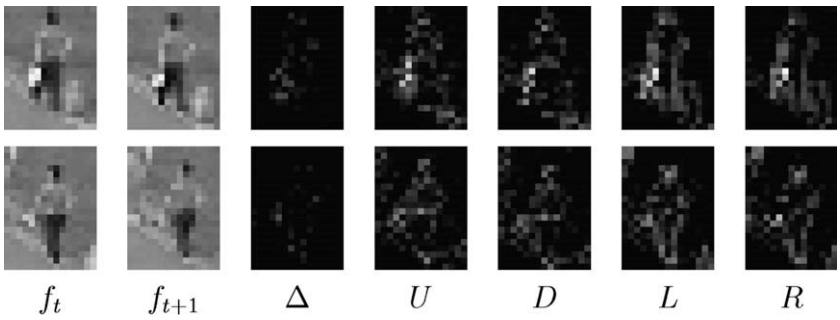

Fig. 2. Two examples of short-term motion image pairs and their corresponding 5 motion images via frame differencing after different orientation filtering.

differencing [25]. Thus a motion orientation histogram does not bring about additional advantage over simple motion filters.

2.1. Seeding detection

Popular classifiers for object detection include SVMs [19], neural networks [21], naive Bayes classifiers [22], AdaBoost [20], etc. Comparison of these different classifiers is out of scope of this paper. Thus we use the boosted stumps as Viola et al. [25], since it has been shown to work well and has real time performance. They perform feature selection and could give some insights into analyzing the importance of each feature type, such as the linear features and gradient orientation features as shown in Fig. 7, thus resulting a fairly small and interpretable classifier. The weak detector is a thresholded single features decision function, consisting of a feature (f), a threshold (θ) and a polarity (p) indicating the direction of the inequality:

$$w(x,f,p,\theta) = \begin{cases} 1 & \text{if } pf(x) < p\theta \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

The strong classifier is a linear combination of the ranked weak classifiers. In our paper, we perform 1000 round of boosting among the sets of different features types. The selected weak classifiers are used to construct the final detector. To achieve scale invariance, we run the detector on a test image or image pair at multiple scales multiplied by a scale factor 0.8 and report the location as well as the scale with strongest output. In order to combine other information evidence usually interpreted in probability form, e.g. the long-term motion information, we convert the output of each boosted classifier into a probability by using the sigmoid transform.

3. Removing false alarms using long-term motion

3.1. Detection adaptation

Directly boosting appearance and short-term motion information for pedestrian detection does not cope well with non-stationary background clutter as shown in Fig. 4Fig. 5. Because the appearance and local motion do not take into account the propagating dependencies between different frames. To address this problem, a typical approach is to deploy Bayesian temporal filtering to propagate the conditional dependencies. However Most of the existing work considers the detection and its adaptation as a separate task, where the two tasks should be addressed simultaneously in practical applications in visual surveillance. Strens and Gregory described an object detection and tracking approach using hidden Markov model (HMM) [24]. They used HMM to model the dynamics of a region of interest (ROI) rather than a single pixel. This approach can be considered as a special variant of Bayesian filtering. The approach is evaluated based on synthetic data. Hence the performance of this approach on real data in not known. The method presented in [17,4] is also along this line of research, but does not include the long-term motion information. Thus it would inevitably propagate the false positives created by the AdaBoost. At the first step, we use a similar approach as in [17] and take the output from the AdaBoost detector only as hypotheses, then adopt the Bayesian sequential estimation, particle filtering technique [12], to cope with the non-linear, non-Gaussian models. In our application, due to the occlusion and diversity of the motion of the hypothesis, the particle filtering is an idea model of our system. The basic Bayesian filtering is a recursive process in which each iteration consists of a prediction step and a filtering step described as follows:

Fig. 3. The graph models of the Bayesian verification process.

Fig. 4. Top 5 Haar-like linear (AI) features selected by AdaBoost, when jointly boosted with AG (gradient orientation histogram) features. Note that the top 10 features selected are shown in Fig. 1.

$$\text{Prediction step}: \quad (p(x_t|y_{0:t-1}) = \int p(x_t|x_{t-1})p(x_{t-1}|y_{0:t-1})\mathrm{d}x_{t-1}$$

$$\text{Filtering step}: \quad p(x_t|y_{0:t}) = \frac{p(y_t|x_t)p(x_t|y_{0:t-1})}{\int p(y_t|x_t)p(x_t|y_{0:t-1})\mathrm{d}x_t}$$

where the $p(y_t|x_t)$ is the likelihood model, and $p(x_t|x_{t-1})$ is the target dynamics model, which we set as a zero-order temporal tracking model for verification. The temporal model is defined as

$$x(t) = x(t-1) + N(0,\Sigma) \tag{9}$$

where $x(t)$ is the hidden state of location and scale of the object at time t, i.e., (l_x, l_y, s).$N(0,\Sigma)$ is a temporal prior measurement Gaussian distribution model. Here, we empirically set the parameter Σ to be diagonal as (2,3,0.3). The number of particles we used in our experiments is 30 per instance.

Tracking can recover the misdetections by AdaBoost, however, prediction errors in tracking learned from previous frames could be accumulated incorrectly without some form of verification, especially, when it is initialized by a new false alarm. An example can be seen in Fig. 5. In the following, we further seek to remove these false alarms using motion evidence across a relatively large temporal scale.

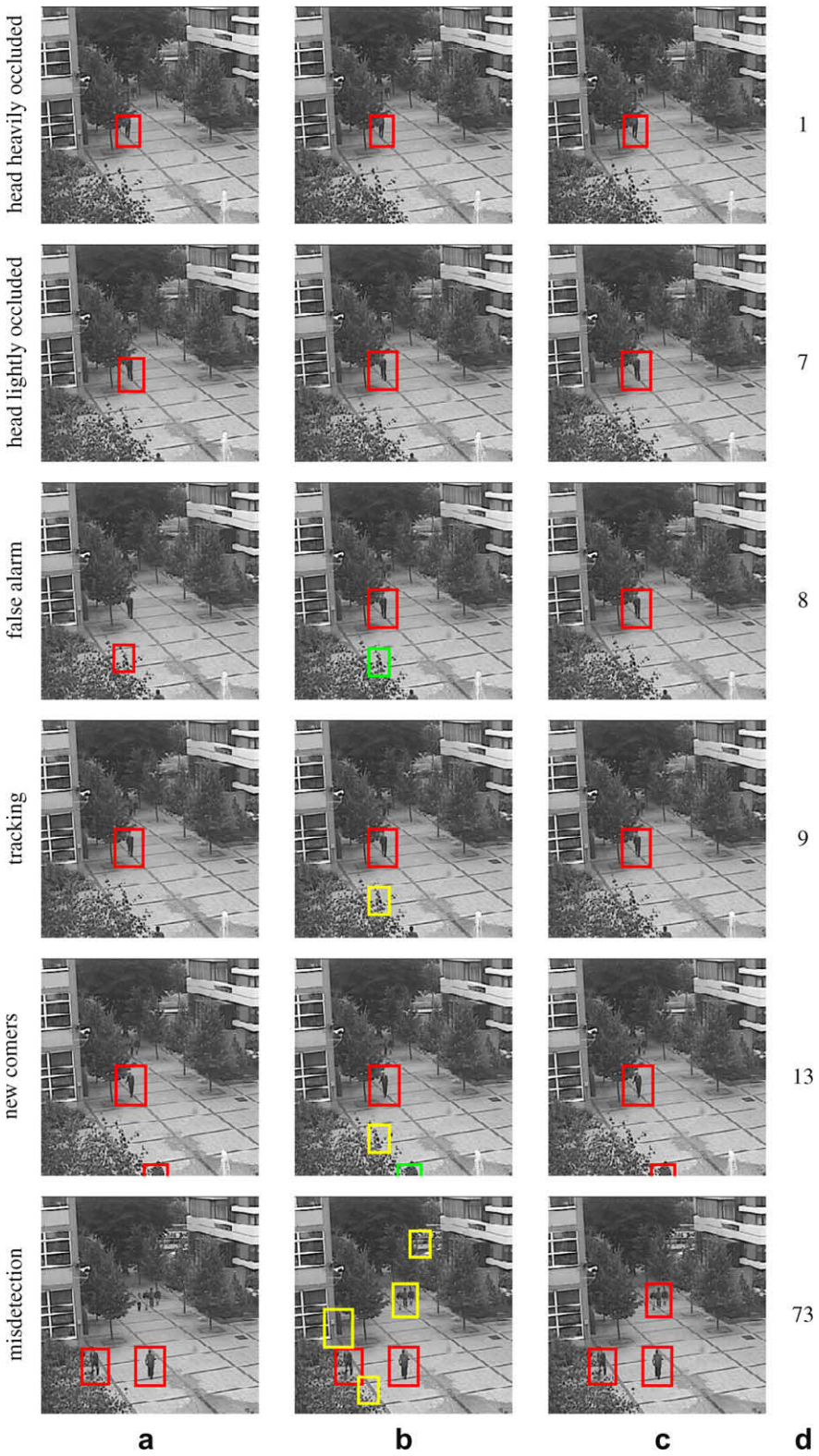

Fig. 5. Scene examples of the detection results in each step of our method. (a) The raw output of AdaBoost detector: red bounding box; (b) shows the results maintained by the tracker. Red box means the detection agreed by both the tracker and AdaBoost detector in the current frame except any new detections by AdaBoost in the current frame (shown as green boxes). Yellow box denotes the hypothesis maintained by the tracker while missed by detector on the current frame; (c) results after fusing the long-term motion information; (d) the corresponding frame in the sequence.

3.2. Motion confidence map

Looking at objects across a long temporal segment could give us additional cue to identify them. Fig. 6 shows examples of four hypotheses as well as their motion confidence maps. Note that though dynamic AdaBoost detector assigns a high score for each of them, we can still clearly see their difference by their long-term motion map. Here we adopt a long-term motion estimation ap-

proach using background subtraction, assuming a fixed view. More precisely, we utilize a Gaussian mixture background model [23]:

$$p(x,y,t) = \sum_i \alpha_i g(f(x,y,t), \theta_{i,x,y}, \sigma_{i,x,y}), \tag{10}$$

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 $f(x,y,t)$ is the local pixel intensity. The variation of one frame $f(x,y,t)$ with respect to the background model is estimated as the probability distance given by

$$v(x,y,t) = \sum_i \alpha_i \exp(-1/2(f(x,y,t) - \hat{\theta}_{i,x,y})^2/\hat{\sigma}^2_{i,x,y}) \tag{11}$$

$(\hat{\theta}, \hat{\sigma})$ is the estimation of (θ, σ). This type of motion information is very effective at highlighting changes in motion in the scene. However, this is also an undesirable property in our case since the noisy motion caused by lighting changes is inevitably augmented. See Fig. 8(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:

$$v(x,y,t) = \sum_i \alpha_i \exp\left(-\frac{1}{2}\frac{(f(x,y,t) - \hat{\theta}_{i,x,y})^2}{\hat{\sigma}_s^2}\right) \tag{12}$$

In the background model of Eq. (10), $\sigma_{i,x,y}$ is the strength of the motion of each pixel at (x,y), we calculate σ_s here in Eq. (12) as the mean or median of $\sigma_{i,x,y}$. Examples of motion extraction using this model are shown in Fig. 8(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 utilizing the spatial motion contrast measure given by Eq. (12) for removing motion noise as compared to existing Gaussian mixture models. To find the contributions of the motion confidence m of each hypothesis h to the object o, i.e., $p(m|o,h)$, we define it as a correlation between the motion map of a hypothesis, m_h, and pre-obtained local human motion template m_t as shown in Fig. 6 by a set of local human long-term motion maps. Note that this definition enables $p(m|o,h) \in [0,1]$, thus it can be considered as a probability distribution.

$$p(m|o,h) = \frac{< m_h \cdot m_t >}{|m_h||m_t|} \tag{13}$$

3.3. Bayesian verification

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:

$$p(o|c,m,h) = \frac{1}{Z} p(m|h,o)^{\gamma} p(c|h,o) p(h|o) \tag{14}$$

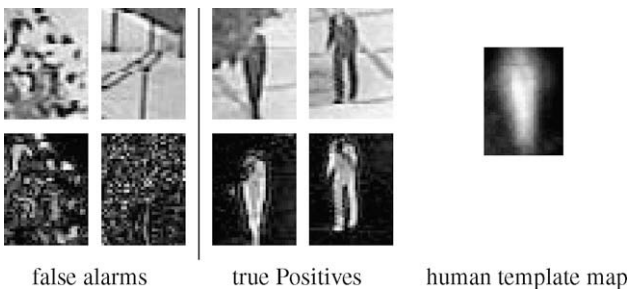

Fig. 6. The appearance images (top row) and the long-term motion maps (bottom row) of false alarms (left) and true positives (middle). Right column shows the human motion template used in our experiments.

Here, we assume that the motion m and the appearance c are conditionally independent. To understand this more clearly, the directed probability graph model of the Bayesian verification process (Eq. 14) is shown in Fig. 3, where the arrows indicate the dependencies between variables. $p(m|h,o)$ is the contribution of the motion map within the hypothesis bounding box given the object, which is computed using Eq. (13). $p(c|h,o)$ is the appearance confidence measure generated by the AdaBoost detector. $p(h|o)$ is the confidence of the hypothesis by the object detector, which in our case is a hypothesis presence indicator for a certain object class. The final candidates are selected by thresholding $p(o|c,m,h)$. We introduce the exponent γ to balance the relative confidence of these evidences and to avoid the over confidence of one of them. It can be set by cross-validation and here we use $\gamma = 0.5$.

4. Experiments

4.1. Data set

Training set: The training set is collected from the video sequences in PETS2001 database [1]. Positive examples are collected by manually tracking pedestrians through video sequences. Negative examples are uniformly randomly selected from the non-pedestrian areas across frames at different window locations and scales. Sample patches are normalized into patches of size 20×15 as the same done in [25]. Negative examples (2250) and positive examples (2250) were used during training. Examples of these images are shown in Fig. 2.

Test set: Out test set contains image frames from CCTV video sequences taken from courtyard by fixed cameras, which is independent from the training set. The courtyard sequences of size 640×480 per frame contain moving human figures of low resolution. Two sequences are used in our experiments with one containing 742 frames, the other 270 frames. They present static background clutters caused by the buildings' windows and dynamic background clutters caused the movement of tree leaves. The dynamics of the background clutter can be clearly shown by the optic flow computed by a robust estimation method by using a robust method proposed by Gautama et al. [8]. See Fig. 8(a) for the optic flow. Many frames contain multiple people under severe occlusions. Fig. 5 shows some example frames from the courtyard scene.

4.2. Evaluation strategy

The performance of our method is evaluated by comparing the detected bounding box B_d of each hypothesis location to the ground truth bounding box B_g in manually annotated data. This procedure is similar to the PASCAL VOC (visual object class) competition [2], and we compute the overlapping score as

$$\lambda = \text{area}(B_d \bigcap B_g)/\text{area}(B_d \bigcup B_g) \tag{15}$$

if $\lambda > 0.5$, then B_d is considered as a true positive, otherwise it is considered as a false positive. Similar criterion has been used in [3]. We assign the detection score for each position, which indicates the confidence that the object is detected. By varying the threshold of the detection scores of each detector one at a time, we obtain the Receiver Operating Characteristics (ROC) curves of those detectors, showing false positive rate versus true positive rate.

4.3. Feature comparison

We give a comparative evaluation of different types of appearance features, short-term motion features as well as their possible

combinations as described in Section 2. AI denotes the Haar-like linear features as used in [20], which operated on intensity images. AG represents histograms of appearance gradient orientations calculated by Eq. (4). MI is short-term motion information obtained by direct image differencing as used in [25]. MG is motion gradient orientation features given by Eq. (7) and operated on differenced images. We further denote a combination of AI and MI as AI+MI, a combination of histograms of gradient orientations plus short-term motion as AG+MI.

We first examine the importance of AI features and AG features when they are jointly boosted by the AdaBoost. To save some time, a set of 16000 features (8000 AI features, 8000 AG features) are randomly selected from approximately one million features. We perform 1000 round of boosting among the feature set. Fig. 1 shows that top 10 features selected by AdaBoost. While Fig. 4 shows the top linear features selected by AdaBoost. It is important to note that all of the top ten features selected by AdaBoost are AG features, while rank of the first linear feature is 17. This clearly shows that AG features are more important the Haar-like features. This can be further verified by the statistics of AI and AG features selected by AdaBoost from a total of 100 and 1000 features. respectively. AG accounts for 86% of the top 100 features, 80% of the top 1000 features. This further confirms that the AG features dominates the selected feature sets. To give a quantitative evaluation of the discrimination power of different feature types, we show ROC curves of detection output with respect to ground-truth from manual annotations in Fig. 7. These were calculated for each feature type respectively given same feature set of 1000. From these ROC curves, we can see that motion information really improves the performance. It is very interesting to note that, AG+AI features perform better than AI features alone, even better than a combination of linear features with short-term motion information. AG features plus motion can further improve the results. AG+MG features give little improvement over AG+MI. This is probably because motion gradient information does not offer more discrimination power than motion intensity features, and direct image differencing makes the estimation of motion gradient less robust. For computational accuracy and efficiency, in the following experiments, we thus use the output of the AdaBoost with the features AG+MI.

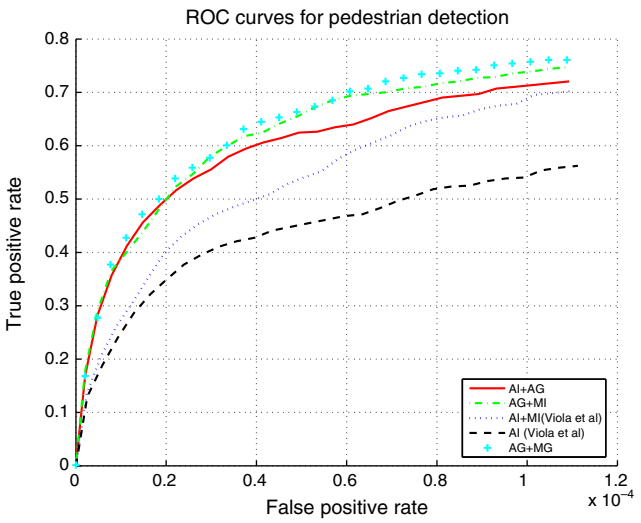

Fig. 7. Comparison of ROC curves on a courtyard sequence using different appearance and dynamic intensity features as well as Gradient Orientation features.

4.4. Improving robustness

Fig. 5 shows detection results from each step of our system in a courtyard sequence. Fig. 5(a) shows the raw output of AdaBoost detector indicated by the red bounding boxes. Fig. 5(b) shows the results maintained by the tracker. Red box means the detection agreed by both the tracker and AdaBoost detector in the current frame except any new detections by AdaBoost in the current frame (shown as green boxes). Yellow boxes show hypotheses maintained by the tracker while missed by the detector in the current frame. Fig. 5(c) shows results after Bayesian fusing of long-term motion information. From Fig. 5(a), we can clearly see that the output of AdaBoost inevitably contains some false alarms (e.g. frame 8) and misdetections (e.g. frame 8, 73). For objects newly appeared in the scene (see Fig. 5(a) (frame 13)), AdaBoost detected successfully the new object in the scene. However, it is also worth noticing that false alarms triggered by the AdaBoost detector were then wrongly maintained by the tracker in the following frames as shown in Fig. 5(b) (from frame 9). This is an example where AdaBoost learned with short-term motion information does not work well with non-stationary background clutter. The tracker itself can not reject such an error either. However, as shown in Fig. 5(c), taking into account of long-term motion given by Eq. (14) rejected successfully these false alarms. It is also worth pointing out that our model was trained to detect people of size larger than 20×15 pixels. This explains why some of the small sized people were not detected when they were far away from the camera (e.g. in frame 1).

Fig. 9 compares the ROC curves of the method by direct detection and the one after Bayesian verification on two courtyard sequences. It is clearly shown that the proposed method outperforms the one using direct detection in terms of the ROC curves, which further suggests the effectiveness of the proposed detection method.

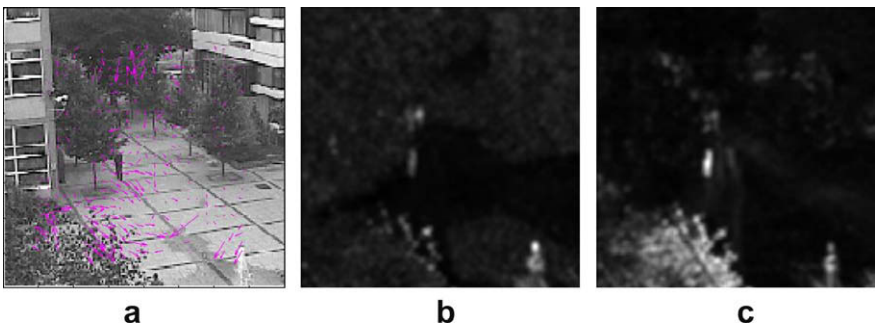

Fig. 8. Examples of the motion confidence map after background extraction. (a) The initial image with the optic flow estimated using a robust estimation method proposed by Gautama et al. [8]; (b) motion confidence map only using Gaussian Mixture by Eq. (11); (c) refined motion confidence map by Eq. (12).

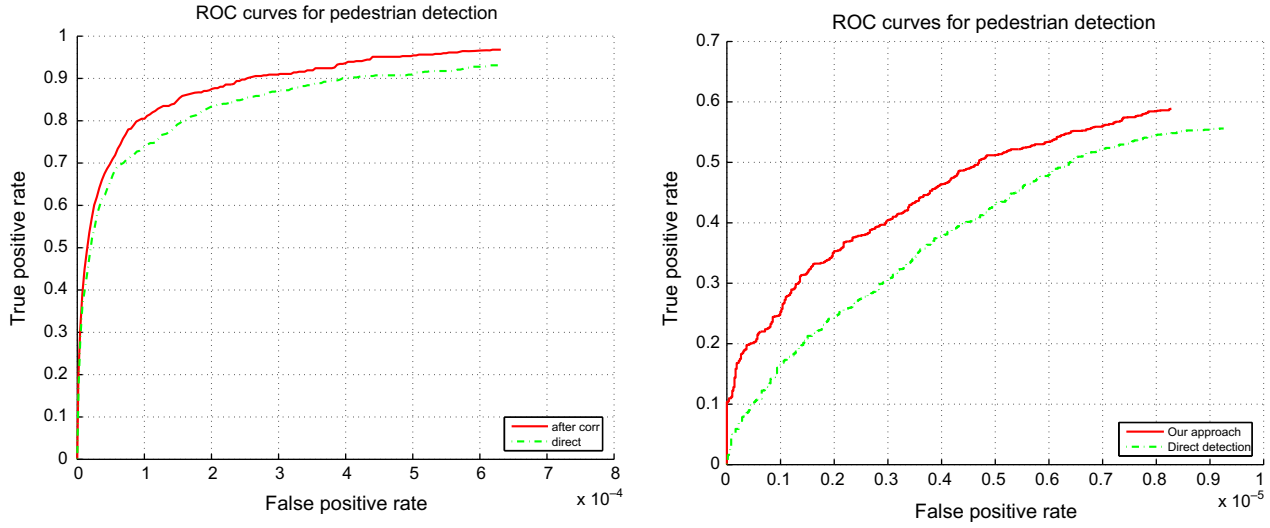

Fig. 9. Comparison of ROC curves on two courtyard sequences using direct detection and our approach.

5. Conclusion

In this paper, we have presented a framework for robust people detection in low resolution image sequences of highly cluttered scenes. Our model utilizes both human appearance and their long-term motion information. In particular, we adopt an integral gradient orientation histogram map to represent both appearance and short-term motion features. Tracking is also considered for correcting misdetections by appearance and short-term motion information alone. Furthermore, long-term motion is utilized to further remove false alarms in detection. We show that for pedestrian detection in video sequences, both short-term and long-term motion information play an important role.

These results can be further improved by including more examples in our AdaBoost training data for human that seen against dynamic background. At present, fusing the long-term motion information is done by direct correlation. Another improvement could be by boosting long-term motion map of the pedestrians against that of the negative samples before feeding to the Bayesian model.

References

[1] Available from: http://www.cvg.cs.rdg.ac.uk/.
[2] Available from: http://www.pascal-network.org/challenges/voc/.
[3] Shivani Agarwal, Aatif Awan, Dan Roth, Learning to detect objects in images via a sparse, part-based representation, IEEE Transactions on Pattern Analysis and Machine Intelligence 26 (11) (2004) 1475–1490.
[4] Yizheng Cai, Nando de Freitas, James Little, Robust visual tracking for multiple targets, European Conference on Computer Vision (2006) 107–118.
[5] R. Cutler, L. Davis, Robust real-time periodic motion detection: analysis and applications, IEEE Transactions on Pattern Analysis and Machine Intelligence 22 (8) (2000) 781–796.
[6] Navneet Dalal, Bill Triggs, Histograms of oriented gradients for human detection, IEEE Conference on Computer Vision and Pattern Recognition 2 (2005) 886–893.
[7] Navneet Dalal, Bill Triggs, Cordelia Schmid, Human detection using oriented histograms of flow and appearance, European Conference on Computer Vision (2006) 428–441.
[8] T. Gautama, M.M. Van Hulle, A phase-based approach to the estimation of the optical flow field using spatial filtering, IEEE Transactions on Neural Networks 13 (2002) 1127–1136.
[9] D. Gavrila, V. Philomin, Real-time object detection for "smart" vehicles, IEEE Conference on Computer Vision and Pattern Recognition (1999) 87–93.
[10] D.M. Gavrila, The visual analysis of human movement: a survey, Computer Vision and Image Understanding 73 (1) (1999) 82–98.
[11] D.D. Hoffman, B.E. Flinchbaugh, The interpretation of biological motion, Biological Cybernetics (1982) 195–204.
[12] Michael Isard, Andrew Blake, Contour tracking by stochastic propagation of conditional density, European Conference on Computer Vision 1 (1996) 343–356.
[13] G. Johansson, Visual perception of biological motion and a model for its analysis, Perception and Psychophysics 14 (1973) 201–211.
[14] I. Laptev, Improvements of object detection using boosted histograms, British Machine Vision Conference (2006).
[15] D.G. Lowe, Distinctive image features from scale-invariant keypoints, International Journal of Computer Vision 60 (2) (2004) 91–110.
[16] Krystian Mikolajczyk, Bastian Leibe, Bernt Schiele, Multiple object class detection with a generative model, IEEE Conference on Computer Vision and Pattern Recognition (2006) 26–36.
[17] K. Okuma, A. Taleghani, N. de Freitas, J. Little, D. Lowe, A boosted particle filter: Multitarget detection and tracking, European Conference on Computer Vision (2004) 28–39.
[18] C. Papageorgiou, T. Poggio, A trainable system for object detection, International Journal of Computer Vision 38 (1) (2000) 15–33.
[19] Constantine Papageorgiou, Tomaso Poggio, A trainable system for object detection, International Journal of Computer Vision 38 (1) (2000) 15–33.
[20] Michael J. Jones Paul Viola, Robust real-time face detection, International Journal of Computer Vision 57 (2) (2004) 137–154.
[21] Henry Rowley, Shumeet Baluja, Takeo Kanade, Human face detection in visual scenes, Advances in Neural Information Processing Systems (1996) 875–881.
[22] Henry Schneiderman, Takeo Kanade, A statistical model for 3d object detection applied to faces and cars, IEEE Conference on Computer Vision and Pattern Recognition (2000) 1746–1759. June.
[23] C. Stauffer, W.E.L. Grimson, Adaptive background mixture models for real time tracking, IEEE Conference on Computer Vision and Pattern Recognition (1999) 2246–2252.
[24] Malcolm J.A. Strens, Ian N. Gregory, Tracking in cluttered images, Image and Vision Computing 21 (10) (2003) 891–911.
[25] Paul Viola, Michael J. Jones, Daniel Snow, Detecting pedestrians using patterns of motion and appearance, International Journal of Computer Vision 63 (2) (2005) 153–161.
[26] Jianguo Zhang, Marcin Marszalek, Svetlana Lazebnik, Cordelia Schmid, Local features and kernels for classification of texture and object categories: a comprehensive study, International Journal of Computer Vision 73 (2) (2007) 213–238.