IEICE TRANS. INF. & SYST., VOL.E96–D, NO.6 JUNE 2013

PAPER

Pedestrian Detection by Using a Spatio-Temporal Histogram of Oriented Gradients

Chunsheng HUA[†*a)], **Yasushi MAKIHARA**[†b)], *Nonmembers*, *and* **Yasushi YAGI**[†c)], *Member*

SUMMARY In this paper, we propose a pedestrian detection algorithm based on both appearance and motion features to achieve high detection accuracy when applied to complex scenes. Here, a pedestrian's appearance is described by a histogram of oriented spatial gradients, and his/her motion is represented by another histogram of temporal gradients computed from successive frames. Since pedestrians typically exhibit not only their human shapes but also unique human movements generated by their arms and legs, the proposed algorithm is particularly powerful in discriminating a pedestrian from a cluttered situation, where some background regions may appear to have human shapes, but their motion differs from human movement. Unlike the algorithm based on a co-occurrence feature descriptor where significant generalization errors may arise owing to the lack of extensive training samples to cover feature variations, the proposed algorithm describes the shape and motion as unique features. These features enable us to train a pedestrian detector in the form of a spatio-temporal histogram of oriented gradients using the AdaBoost algorithm with a relatively small training dataset, while still achieving excellent detection performance. We have confirmed the effectiveness of the proposed algorithm through experiments on several public datasets.

key words: *spatio-temporal gradients, AdaBoost, pedestrian detection*

1. Introduction

Over the last two decades, pedestrian detection has been an important part of many computer vision tasks and has attracted the attention of numerous researchers. An effective pedestrian detection system plays an important role in smart vehicles to help avoid collisions with people. It is also one of the kernel functions of intelligent rescue robots, surveillance systems, and human activity understanding. In this paper, we identify people walking in successive frames as "*pedestrians*" and exclude people in static positions.

Owing to the lack of an explicit model to describe the non-rigid human body, we usually establish a pedestrian detector by using a machine learning algorithm to extract an implicit pedestrian model from the numerous training samples. Based on these considerations, many algorithms [5]–[7], [12], [15], [27] have been proposed, in which the detectors are trained to recognize different appearance features, such as the histogram of oriented gradients (HOG), Haar-like features, or edgelet features, using the support vector machine (SVM) or AdaBoost algorithms. The main problem with these algorithms is that false rejection occurs when pedestrian appearance variations (e.g., articulated poses, different viewpoints, and clothes, such as those shown in the left panel of Fig. 1 are large). False alarms also arise under cluttered background conditions (shown in the right panel of Fig. 1), where, for example, certain background areas happen to contain human-like shapes or the appearance of a pedestrian is strongly textured. Therefore, it is difficult for an appearance-based pedestrian detection algorithm to achieve high performance on its own under such difficult conditions.

Various pedestrian appearances

False detection caused by the cluttered background or strong texture in people

Fig. 1 Unpredictable pedestrian and background variations are the main reason for the failure of conventional appearance-based pedestrian detection algorithms. (Right-hand side shows the results of the HOG in [7].)

To overcome such problems, temporal or motion features have been introduced to improve the detection performance based on the fact that not only human shape, but also human motion can distinguish a pedestrian from the surrounding environment. This idea has proved to be useful in action recognition [16], [17], [20], [21], [24], [26] and patch-based image segmentation [18], [23].

In recent work [21], a co-occurrence spatio-temporal 3D gradient is used to describe the appearance and motion features of an object. However, when using a combination of spatial and temporal gradients, any noise in either gradient may result in its 3D gradient being allocated to the wrong bin of the histogram. Moreover, when the co-occurrence gradient is applied [21], extensive training samples are required to cover all possible combinations of appearance and motion variations. In addition, this approach tends to suffer from an over-training problem that leads to significant generalization errors and low detection performance.

Previous approaches [8], [20], [33] have combined the HOG and oriented optical flow vectors to improve the detec-

Manuscript received August 10, 2012.
Manuscript revised December 19, 2012.

[†]The authors are with the Institute of Scientific and Industrial Research, Osaka University, Ibaraki-shi, 567–0047 Japan.
[*]Presently, with the State Key Laboratory of Robotics, Shenyang Institute of Automation, Chinese Academy of Sciences, Shenyang, Liaoning, 110179, China.
a) E-mail: huachunsheng@gmail.com
b) E-mail: makihara@am.sanken.osaka-u.ac.jp
c) E-mail: yagi@am.sanken.osaka-u.ac.jp
DOI: 10.1587/transinf.E96.D.1376

tion rate. Since the optical flow is sensitive to the viewpoint of the camera and walking directions, the performance of this technique may easily be degraded if the camera viewpoint and/or the walking directions of pedestrians differ in the training and test samples. To avoid this problem, the training process for this method also needs extensive training samples to cover all possible motion directions. A detailed discussion of [8], [20], [21] can be found in Sect. 2.2.

To solve these problems, we propose a spatio-temporal HOG (***STHOG***) feature for pedestrian detection, where the appearance and motion of the pedestrian are described by two separate spatial and temporal histograms. The spatial and temporal gradients computed from successive frames are binned into different histograms according to their orientation, while the magnitude of the bins describes the strength of the spatial/temporal gradients. The advantages of the STHOG feature include: 1) robustness with respect to noise, since the gradients are not co-occurrence gradients, and spatial or temporal noise is also separated into either the spatial or temporal histogram; 2) insensitivity to changes in walking direction, since the temporal gradients of the STHOG feature describe the strength of movement, but not the walking direction, and therefore it is relatively insensitive to changes in the viewpoint within a certain range; and 3) robustness with respect to the problem of small training sample sizes, because the appearance and motion features are described separately in the STHOG feature. The STHOG feature seldom suffers from the over-training problem when trained with fewer training samples, compared with co-occurrent features (e.g., the 3DHOG feature [21]).

2. Related Work

2.1 Spatial Feature-Based Pedestrian Detection

Despite its expensive computational cost, the sliding window detection (SWD) algorithm is still regarded as one of the most efficient detection methods because of its impressive detection rate when scanning over an input image with variable scales. Some effort has been made to increase the processing speed of the SWD by applying a cascade-structure detector [6], [9], a camera geometry constraint to restrict the search area [11], [12], [19], or prior knowledge of the target shape to refine the detection result.

The Haar-wavelet cascade detector [6] is an efficient SWD method that introduces a cascade rejecter to locate people based on Haar-like features and spatial differences. The success of the cascade-structure detector is based on the assumption that the overwhelming majority of input images are background ones.

The *HOG* [7] uses a normalized histogram to describe the local shape and appearance of a target object (similar to SIFT). Local gradients are binned according to their orientation and weighted by their magnitude within a spatial grid of cells. The most discriminating HOG features for pedestrian detection are selected using a linear SVM. Zhu *et al.* [9] achieved almost real-time single-scale pedestrian detection by training a cascade detector with an integral HOG feature, while achieving almost the same detection performance as in [7]. Further improvements by including the spatial co-occurrence within a single window for each region of the HOG [18] have been reported.

Wu *et al.* [12] proposed the Edgelet feature for pedestrian detection when partial occlusion occurs. The affinity function of the Edgelet is a variation of Chamfer matching that can capture both the intensity and shape of the edge. With prior knowledge of the camera position and ground-plane, the partial occlusion problem is solved by maximizing the joint image likelihood with and without occlusion.

The interest-point-based pedestrian detector [11], [19] handles partial occlusion by integrating multiple algorithms. The pedestrian hypotheses obtained from the implicit shape model detector are further verified by the 3D information from stereo cameras, camera ego-motion flow, and ground-plane geometry constraints.

To reduce the effect of a noisy background and partial occlusion, the HOG-LBP [27] detector introduces a local binary pattern (*LBP*) histogram into the conventional HOG. The histogram produced by the LBP is used to suppress random background noise, while partial occlusion is detected by checking the inner product of a cell in the SVM classification. When partial occlusion occurs, the part-based HOG-LBP detector is applied to find the occluded person.

By combining the ground plane assumption with multi-resolution models, Park *et al.* [38] improved the HOG detector when dealing with a wide range of pedestrian appearance changes in scales. Barinova *et al.* [40] presented a framework for detecting objects by describing their feature using Hough transformation. Object detection is achieved by maximizing the energy function of the log-posterior of the Hough transformation results. Roth *et al.* [41] trained a separate grid classifier for the target and background models. Object detection was achieved by comparing the similarity obtained by these two classifiers. Online updating was also carried out for the background classifier. Further improvement of this work was reported in [42].

More recent work [29], [30] showed that a combination of deformable part and holistic models could greatly improve the performance of traditional algorithms (HOG or Haar-like features). By combining richer features and more sophisticated learning techniques, they greatly improved the detection rate for complex datasets. An interesting outcome from [34] was that the virtual images could also be used for training a pedestrian detector. Wang *et al.* [37] developed an automatic algorithm to select new confident positive and negative samples for re-training the appearance-based pedestrian detector under a new unknown traffic scene.

Good surveys of the literature, new evaluation criteria, and other contributions to pedestrian detection can be found in [10], [13]–[15], [28], [32], [35], [36].

2.2 Spatio-Temporal Feature Descriptors

The motivation for using spatio-temporal features to de-

scribe a target object stems from the fact that not only its appearance, but also its motion distinguishes a target object from others. In the case of pedestrian detection, a pedestrian is assumed to have both a human shape and unique human movement such as the movement of arms and legs. Therefore, spatio-temporal features are considered to be more powerful than appearance features in pedestrian detection. The following related spatio-temporal descriptors have been proposed for action recognition and pedestrian detection.

2.2.1 3DHOG Descriptor

Kläser *et al.* [21] achieved action recognition under complex background conditions by using a 3DHOG feature descriptor that can describe both shape and motion features with a co-occurrence spatio-temporal vector. For a given 3D patch that is divided into $n_x \times n_y \times n_t$ cells, the gradients calculated in the x, y, and t directions are combined to produce a spatio-temporal vector (as illustrated in Fig. 2. The orientation of this co-occurrence spatio-temporal vector is quantized by projecting it onto an icosahedron (20 sides, which means the histogram has 20 bins) and identifying its nearest orientations. The corresponding 3DHOG descriptor concatenates gradient histograms of all cells and is normalized.

Since the 3DHOG descriptor uses a co-occurrence vector to describe human shape and motion features while training a 3DHOG descriptor, it always requires extensive training samples to cover all possible combinations of each gradient. Otherwise, the 3DHOG features tend to suffer from over-training. Moreover, since the orientation of this co-occurrence 3D vector is determined by projecting it onto each side of the icosahedron, any noise from x, y, or t will cause it to be projected onto the wrong side. To reduce the effect of noise, in [21], a 3DHOG descriptor was applied around some robustly extracted feature points, instead of directly to the full image. Our experimental results in Sect. 5 also show that applying the 3DHOG descriptor directly to the pedestrian detection task is unsuitable.

2.2.2 HOGHOF Descriptor

The HOGHOF descriptor describes both the shape and motion features of a target object with a concatenated histogram. Such a descriptor can be applied to pedestrian detection [8] and action recognition [16], [20]. To characterize the local motion and appearance, histograms of the oriented gradient and optical flow accumulated in space-time are concatenated and normalized. Walk [33] *et al.* tried to combine the HOG with optical flow (HOF) and color self-similarity to set up a new feature descriptor, where the gradient appearance, motion, and pairwise color distribution of human parts are processed simultaneously.

The problem with the HOGHOF is that it is difficult to extract the optical flow stably in a complex scene. The optical flow is easily changed as a result of different camera viewpoints, walking directions, or variations in illumination. Hence, for example, if the direction of the observation viewpoint to the pedestrian in the test video is different from that in the training samples, the HOGHOF may become very unstable. In practice, the HOGHOF descriptor also requires extensive training samples to cover the possible variations in target features (viewpoints and walking directions).

2.2.3 STGGP Descriptor

Liu *et al.* [39] proposed the spatial-temporal granularity-tunable gradient partition (**STGGP**) descriptor for pedestrian detection. The orientation of spatial-temporal gradients attributed to humans was described in a 3D Hough space, while a generalized plane was produced to partition the gradients by a back-projection from the cubic region of 3D Hough space to 3D gradient space. The generalized plane consisting of nine parameters (such as gradient strength, position, and shape of the plane) was trained by a linear SVM for pedestrian detection. Since the authors tolerated image noise by enlarging the cubic region in 3D Hough space, there was a trade-off between tolerance of noise and partition performance in the 3D spatial-temporal space. This means that enlarging the cubic region in 3D Hough space would lead to a less generalized plane in 3D gradient space, which would reduce the discriminability of this detector.

3. Spatio-Temporal HOG Feature

3.1 Spatio-Temporal Gradient

Usually, a video is considered to be a 3D space consisting of x, y, and t, as illustrated in Fig. 3. When a walking pedestrian is captured on video, his/her movement produces a unique spatio-temporal gradient vector ∇I whose orientation can be represented by the spatial (ϕ) and temporal (θ) orien-

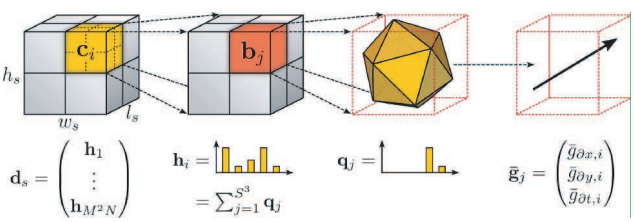

Fig. 2 Illustration of the 3DHOG descriptor in [21].

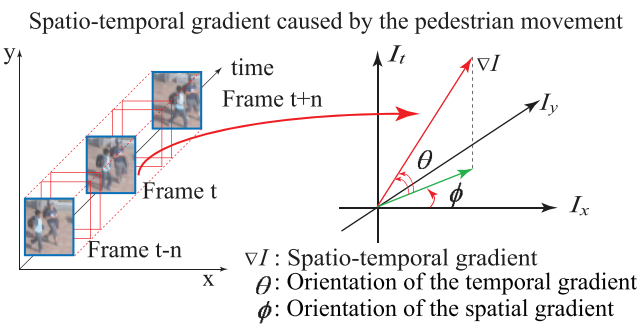

Fig. 3 Relationship between spatial and temporal gradients in STHOG.

tation in the x, y image plane and t direction. Since a pedestrian's movement, caused by the movement of the arms (or legs), and the pedestrian's shape are unique features, it is reasonable to consider that only a pedestrian can display human-like shape and motion. This is because a static background can include only the spatial shape, and not the temporal movement of a human, while a moving background seldom appears to be human-like while simultaneously exhibiting movement similar to that of a person. Therefore, spatial and temporal gradient orientations are considered to be helpful in detecting pedestrians in a video.

At frame t, since an image is represented as $I(x, y, t)$, the orientations of the spatial (ϕ) and temporal (θ) gradients at (x, y, t) are computed by:

$$\nabla I = [I_x, I_y, I_t], \tag{1}$$

$$\theta = \tan^{-1}(\|I_t\| / \sqrt{I_x^2 + I_y^2}), \tag{2}$$

$$\phi = |\tan^{-1}(I_y / I_x)|, \tag{3}$$

where gradients I_x, I_y, and I_t are simply calculated as:

$$I_x = I(x+1, y, t) - I(x-1, y, t), \tag{4}$$

$$I_y = I(x, y+1, t) - I(x, y-1, t), \tag{5}$$

$$I_t = I(x, y, t+1) - I(x, y, t-1). \tag{6}$$

3.2 Construction of STHOG

A statistical histogram structure is used to represent the STHOG features, because a histogram can describe the distribution of the STHOG features while suppressing any abrupt random noise. The computed gradient orientations ϕ and θ are identified in the spatial and temporal histograms to describe the pedestrian shape and motion, respectively. A STHOG feature is produced by concatenating the spatial and temporal histograms, which means the STHOG feature will be a long histogram.

Implementation of the STHOG is very similar to the well-known HOG (as shown in Fig. 4, where ϕ ($0 \sim 180°$) and θ ($0 \sim 90°$) are computed from three successive frames and oriented in nine directions. For normalization, each block contains $2 \times 2 \times 1$ cells, where each cell is composed of $10 \times 10 \times 3$ pixels.

The spatio and temporal histograms in each block are normalized with the $L1$ norm. A STHOG feature is set up by concatenating the two histograms, where 18 bins are established by the STHOG feature.

A detection window contains 5×7 blocks, that is 100×140 pixels by three frames, and the SWD step is 10 pixels.

We determined the size of a cell, a block and detection window experimentally so as to roughly fit to the minimum size of pedestrians in the test datasets that we used. Regarding voting the gradients into histograms, a simple distribution scheme like voting for the nearest neighbor bin could reduce the robustness to noise. To retain this robustness, an interpolation is applied linearly between the two neighbor bins nearest to a computed gradient in both the spatial and temporal histograms.

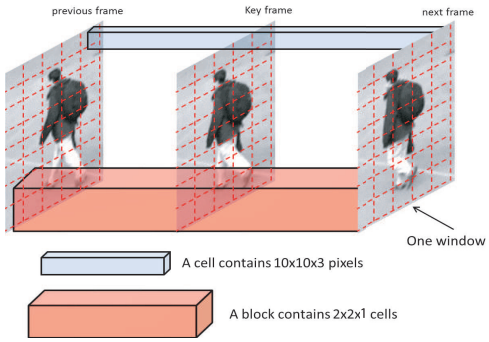

Fig. 4 Construction of the STHOG features from three successive frames.

Let ω be the magnitude of the gradient at pixel (x, y), with ϕ the orientation of its spatial gradient. Then ϕ_1 and ϕ_2 represent the orientations of the corresponding two nearest neighbor bins to ϕ in the histogram. The interpolation computation distributes the magnitude ω into two parts as:

$$\omega_1 = \omega \frac{\|\phi_2 - \phi\|}{\|\phi_2 - \phi_1\|} \tag{7}$$

$$\omega_2 = \omega \frac{\|\phi_1 - \phi\|}{\|\phi_2 - \phi_1\|} \tag{8}$$

The interpolated magnitudes ω_1 and ω_2 are accumulated for all pixels within the cell to create the spatial histogram in the STHOG feature. The temporal histogram can be set up by interpolation in the same manner.

Figure 5 illustrates the spatial and temporal histograms under different conditions. The top two rows show the STHOG features of a pedestrian with different walking poses, while the remaining three rows give the STHOG features for static or moving background regions. It is clear that even when the spatial histogram of a part of the background is similar to that of a pedestrian, such a background can still easily be distinguished from a pedestrian because it has a completely different temporal histogram. The moving legs and arms of a pedestrian produce a unique temporal histogram, while static and randomly moving backgrounds produce different temporal histograms.

Unlike in the 3DHOG descriptor, in the STHOG descriptor, since the spatial and temporal gradients are represented separately as independent histograms, orientation noise in spatial (or temporal) gradients does not affect the orientation of the temporal (or spatial) gradients, respectively. In addition, the effect of such noise can be further suppressed using statistical binning in the histogram.

Moreover, since the HOGHOF descriptor constructs a histogram of motion direction, that is, the direction of the optical flow, and the STHOG descriptor constructs a histogram of the length of the normal flow, which is defined as the optical flow projected onto the orientation of the spatial gradient (or normal vector to the contour), the STHOG is relatively invariant to changes in the viewpoints. This makes the STHOG descriptor superior to the HOGHOF descriptor

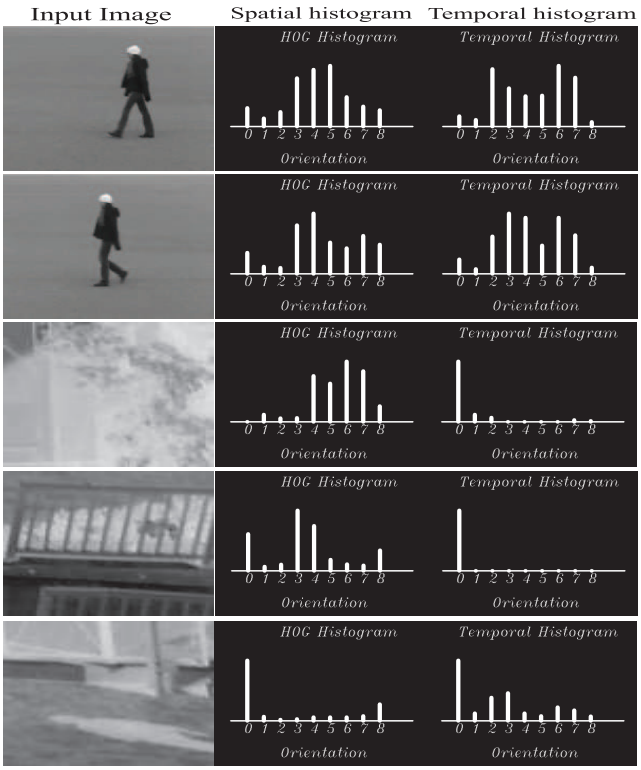

Fig. 5 STHOG features for different cases. Left column: input images. Middle column: spatial histogram. Right column: temporal histogram. The top two rows represent the STHOG features of a pedestrian with different poses. The remaining three rows show that the STHOG features for the background are completely different from those of a pedestrian.

in dealing with changes in viewpoints or walking directions.

4. Pedestrian Detection

4.1 Training

Because of the lack of an explicit pedestrian model, machine learning is always required to complete the pedestrian detection task, where the implicit representation can be learned from extensive training samples. In this study, the simple yet efficient AdaBoost algorithm described in [4] was selected to train our STHOG pedestrian detector.

Each set of positive/negative samples comprises three successive pedestrian/background images, containing both shape and motion information (as shown in Fig. 6. The STHOG features are extracted from successive frames. The STHOG pedestrian detector obtained from AdaBoost is then applied to the video sequence to locate the pedestrian.

During the AdaBoost training process, each bin of the STHOG is considered to be a weak classifier. Because each detection window contains 10×14 cells and each cell has 18 bins, 2520 weak classifiers are prepared in one detection window. A weak classifier is defined as:

$$f_i(b_i) = \begin{cases} 1 & \text{if } P_i(b_i - \theta_i) \geq 0 \\ -1 & \text{else} \end{cases}$$

where P_i, θ_i, and b_i are the parity, bias, and bin value of the i^{th} weak classifier. A strong classifier is constructed by linearly combining $N_{wc} = 300$ weak classifiers selected via AdaBoost:

$$F(\mathbf{b}) = \sum_{i=1}^{N_{wc}} \alpha_i f_i(b_i). \quad (9)$$

Here, α_i is the weight of the i^{th} weak classifier, where the weak classifier is selected by the AdaBoost algorithm according to the values of P_i, b_i, and θ_i.

Fig. 6 Examples of positive and negative training samples for STHOG features.

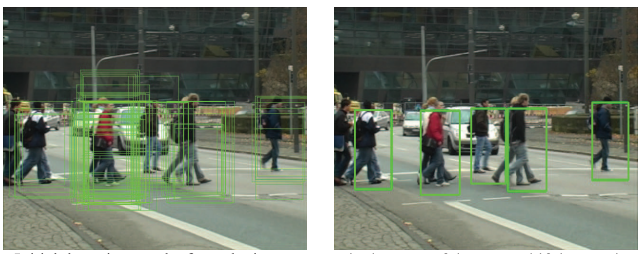

Fig. 7 Output of the STHOG detector using the image pyramid structure.

4.2 Detection

After the AdaBoost training process, the STHOG pedestrian detector scans the image from all positions. To deal with pedestrians of different heights, we selected an image pyramid structure to solve the multi-scale problem. In this image pyramid, the scale varies from 0.15 to 1.0 with the scale step of 0.05, which means that there are 18 scale levels in this structure. The STHOG detector scans over all the images within the pyramid from all positions. This process guarantees that all pedestrians with different heights are checked by the STHOG detector.

As shown in Fig. 7, integration of the output of the STHOG detector becomes the classic clustering initialization problem: we need to identify the number and position of the clusters (in this case, the pedestrian) from an unknown dataset. Therefore, we chose mean shift clustering [2] to integrate these initial detection results. Hereafter, we refer to the output of the integration results as the "detection results". Since there are no perfect clustering algorithms that are suitable for all types of clusters, one pedestrian may respond to multiple detection results, which is also the main reason for the false alarms of the STHOG detector in our recall precision curve (RPC) evaluations.

Table 1 Source of the training and test samples.

	Training Samples	Test Samples
PETS2006	S1-T1-C/1	S2-T3-C/4
	S1-T1-C/2	
	S1-T1-C/3	
	S1-T1-C/4	
PETS2009	S0-RF/Time14-03/View01	S3-MF/Time-12-43/View01
	S0-RF/Time14-03/View02	
	S0-RF/Time14-29/View02	S3-MF/Time-12-43/View02
	S2-L1/Time12-34/View08	
i-LIDS	Subway Hard	×
	CrossRoad Hard	
TUD-Crossing	×	Full Sequence
TUD-Campus	×	Full Sequence

5. Experimental Results

5.1 Datasets and Benchmarks

To evaluate the performance of our STHOG pedestrian detector, we compared the HOG [7], HOGHOF [16] (with optical flow computed by the Horn [1] method), 3DHOG [21], and proposed STHOG pedestrian detectors. Since the STHOG detector requires training and testing with successive images, the single frame-based public datasets (like INRIA and PASCAL) were unsuitable for our research. Thus, in this study, both the training and test images were selected from other public datasets with successive frames, such as PETS2006 [†], PETS2009 [††], i-LIDS [†††], TUD-Campus, and TUD-Crossing [22] (details are given in Table 1).

Regarding the training data, we manually selected 5000 positive and negative training samples from several sequences of the PETS2006, PETS2009, and i-LIDS public datasets. In our training dataset, every three samples were successive frames, as shown in Fig. 6 so as to contain the necessary shape and motion information. The negative samples were manually selected from the image regions that contain no pedestrians. It means that the negative samples could be both the static background regions and the moving object regions such as the waving leaves or shadows (like the bottom row of Fig. 5) and moving vehicles (right of Fig. 6). Most negative samples such as moving vehicles or trains came from the i-LIDS dataset, other negative moving samples like moving shadows, baggage were extracted from the PETS2006 datasets. All the pedestrian detectors were trained using the same samples with the AdaBoost algorithm.

The test sequences were selected from the following public datasets: PETS2006, PETS2009, TUD-Campus, and TUD-Crossing. The ground truth of test sequences from PETS2006 and PETS2009 was manually selected by us, such that even occluded persons were considered as pedestrians, whereas people with height $h \leq 30$ pixels were ignored.

5.2 Performance Evaluation per Window

To evaluate the performance of the STHOG, HOG, 3DHOG, and HOGHOF detectors, we trained each with 5000 positive/negative samples and compared them according to the receiver operating characteristic (ROC). Details of the test samples are given in Table 2. Figure 8 shows the ROC curves for all detectors using different test datasets, where the STHOG detector obtained the best results with almost all the data. The performance of the HOG and HOGHOF detectors seems to be similar to that of the STHOG with the PETS2006 and PETS2009-02 test data. In other words, the test data selected from these sequences are relatively simple, since few occlusions or other noise occurs.

Table 2 The number of test samples from each sequence for ROC.

	Positive Samples	Negative Samples
PETS2006	2840	2843
PETS2009 View01	1763	1763
PETS2009 View02	1472	1472
TUD-Crossing	2987	2987

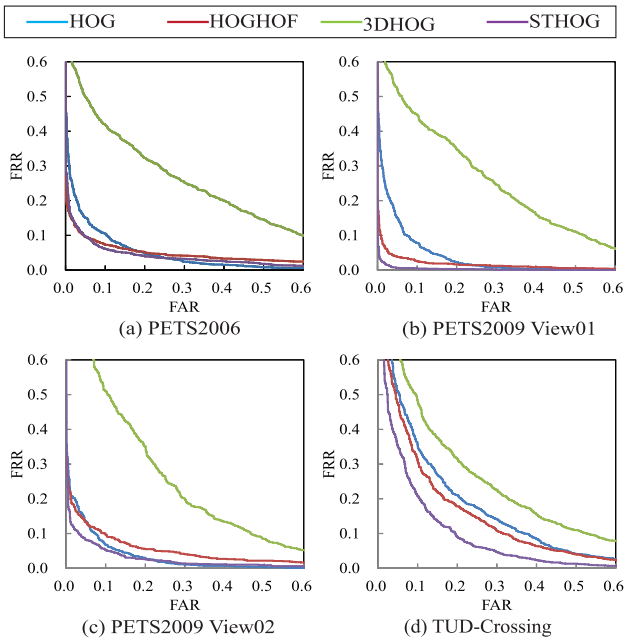

Fig. 8 ROC curves of detectors with different test datasets.

The TUD-Crossing dataset contains the most difficult test data, with complex pedestrian movements and multiple overlap-occlusions occurring frequently (as shown in Fig. 9). With regard to the overlap-occlusion problem, multiple people are located within one detection window. The person at the back can be considered as noise when trying to locate the person in front, while the person in front becomes an obstacle when detecting the person behind. Most importantly, the moving vehicles and their reflections in the glass make the background very cluttered. The main reasons for the missed detections by the STHOG detector were the complex overlapping and occlusions. For the HOGHOF and

[†] ftp://ftp.cs.rdg.ac.uk/pub/PETS2006/
[††] ftp://ftp.cs.rdg.ac.uk/pub/PETS2009/
[†††] i-Lids dataset for AVSS 2007

Fig. 9 Illustration of the overlap-occlusion in one window.

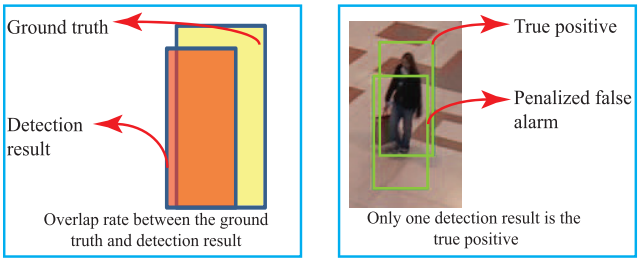

Fig. 10 Illustration of how the overlap rate and one-person-one-response principle is measured.

3DHOG detectors, besides these problems, the difficulties also included differences in the camera viewpoint from the training data, and the moving vehicles producing too many noisy optical flows and noisy temporal gradients. Because of such difficulties, they failed to detect some pedestrians and produced many false alarms in the background samples.

5.3 Performance Evaluation per Image

5.3.1 Evaluation Measurement

To evaluate the performance of each detector with a full image, we used RPCs to describe their detection abilities (details on RPCs can be found in [3]). In practice, a pedestrian detector outputs a set of detection results (the output of the integration process) with a corresponding score. We can threshold these scores to obtain the RPC throughout all the images in the test video.

In addition, although the detected windows are not always perfectly localized to the ground truth position, we consider the detection and location accuracy in a separate way and focus on the aspect of the detection. Note that the pedestrian detection has many useful applications: not only counting the number of pedestrians but also the intruder detection into a confidential area, collision warning of smart vehicles even if the localization accuracy is imperfect. More specifically, as shown in Fig. 10, a detection result is considered a *True Positive* if its overlap rate defined as the intersection ratio between the bounding boxes of a detection result and the ground truth is greater than the threshold; otherwise it is regarded as a *False Alarm*. Multiple detection results on the same person are penalized. If a pedestrian detector predicts multiple results that overlap with one ground truth, only the closest result is considered correct, while the others are penalized as false alarms. We call this the one-person-one-response principle. In this way, the false alarms

Fig. 11 Detection results for the HOG, HOGHOF, 3DHOG, and STHOG pedestrian detectors with our test sequences. From top to bottom: PETS2006, PETS2009 View01, PETS2009 View02, and TUD-Crossing.

in our RPC evaluation come from: 1) false detections in the background; 2) multiple detections rejected by the ground truth.

For each ground truth, the evaluation comparing it and the detection results is performed as follows:

1. Find the reciprocal nearest neighbor to the ground truth from the detection results.
2. Check the overlap rate between the selected detection result and the ground truth.
3. If the overlap rate is greater than the threshold, such result is a true positive result; otherwise, it is a false alarm.

$$overlap_{(rate)} = \frac{BB_{dt} \cap BB_{gt}}{BB_{dt} \cup BB_{gt}}, \tag{10}$$

Here, BB_{dt}, BB_{gt} denote the bounding boxes of the detection result and the ground truth, respectively.

The above evaluation guarantees the following important properties between the evaluated detection results and a ground truth: 1) they must be unique with respect to each other (the recurrent neural network principle); 2) they must be close enough to intersect with each other; and 3) their scale (or size) must be similar. In [36], an interesting conclusion is that the evaluation is insensitive to the overlap rate as long as it is below 0.6.

5.3.2 Qualitative Evaluation

Figure 11 shows the comparative experimental results for the HOG, HOGHOF, 3DHOG, and STHOG pedestrian detectors using our test sequences. Here, each pedestrian detector is trained using 5000 positive and negative samples

with the AdaBoost algorithm.

In the PETS2006 sequence (top row of Fig. 11, we noticed that the false alarms for the STHOG detector were due to giving multiple detection results for one person, while missed detections were caused by failing to find the occluded person. The 3DHOG and HOGHOF detectors produced many false alarms because the pedestrian walking direction in the training data was different from that in the test data. The HOG detector gave false alarms for the background, such as a chair or trash bin that was not included in the training data.

The second and third rows of Fig. 11 show the detection results for each detector using the PETS2009 View01 and View02 test sequences. Since the pedestrians frequently changed their walking direction following an "s" curve, and such conditions were not included in the training data, the HOG, HOGHOF, and 3DHOG detectors suffered from this problem. Most false alarms and missed detections for the STHOG detector were caused by the mean shift in the integration from incorrectly taking multiple persons as one or producing multiple results from one person.

In the TUD-Crossing sequence, shown in the bottom row of Fig. 11, the moving background (vehicles and their shadows), cluttered pedestrian movement, and frequent pedestrian occlusions presented challenging conditions. Most of the false alarms from the STHOG detector were caused by the imperfect clustering integration process where multiple detection results intersecting with one ground truth were penalized by our RPC evaluation criteria as false alarms. With regard to the 3DHOG, HOGHOF, and HOG detectors, these failed by producing many false alarms for vehicles.

5.3.3 Quantitative Evaluation

Figure 12 shows the RPCs for four pedestrian detectors using the test video from all the test sequences. In Fig. 12 (a), the STHOG detector finally achieved its optimal performance with recall of about 73% and a false alarm rate of 50%. The maximum recall for the other detectors was less than 70%, with more than 70% false alarms.

Figure 12 (b) indicates that the performance of the STHOG detector greatly exceeds that of the other detectors. The STHOG detector achieved a maximum recall of 78% while maintaining a false alarm rate of 9%. The RPCs for the HOGHOF and HOG were very similar, because the optical flow in the HOGHOF was almost useless as people frequently changed their walking directions. The RPC for the 3DHOG showed its limited detection ability in the case of large variations in pedestrian movement.

In the case of the TUD-Campus test video (Fig. 12 (c)), the pedestrian appearance changed significantly in scale, while pedestrian occlusion also increased the difficulty of this video. The main reason for the missed detection by the STHOG detector was that only a holistic human model was applied and the STHOG detector failed to find heavily occluded pedestrians. The false alarms for the STHOG

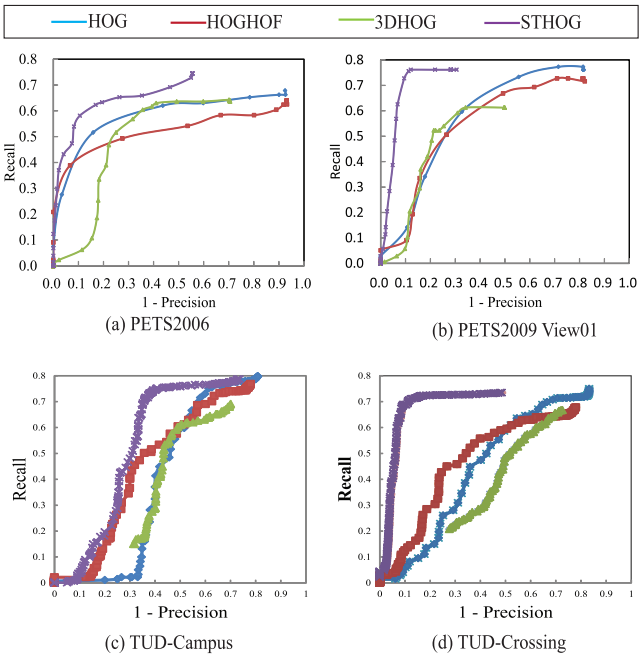

Fig. 12 RPCs for the HOG, HOGHOF, 3DHOG, and STHOG pedestrian detectors using our test sequences.

were mainly due to separating one person into multiple people. There was no clear difference between the HOG and 3DHOG detectors for this sequence, while the performance of the HOGHOF improved because it applied the optical flow separately to describe pedestrian motion.

Figure 12 (d) shows the RPC for each detector using the difficult TUD-Crossing test video. As described previously, the STHOG detector was able to separate the pedestrian from the moving cluttered background, and its maximum recall was about 73% with a false alarm rate of 8%. To achieve the same detection score, the false alarm rate of the HOGHOF and HOG detectors was over 80%, which means almost 10 times more false alarms than the STHOG. The 3DHOG detector was confused by the moving background regions that were not present in the training data, and its maximum recall was about 68% for this sequence. Since there were many occluded persons marked as the ground truth in this dataset, the STHOG detector missed these persons because only a holistic human model was applied in our work. In [40], a recall rate greater than 90% was achieved for this dataset, because the ground truth of the dataset was modified in that work.

5.3.4 Impact of the Number of Training Samples

To evaluate the impact of the number of training samples, we separately selected 500, 1000, 2000, 3000, 4000, and 5000 positive/negative samples from the training dataset and trained each detector with these different numbers of samples. If there is little change in the RPC for a detector, we assume this detector to be stable.

Figure 13 verifies that the performance of the STHOG

detector became stable quickly. That is because in the STHOG features, we used two separate features for the shape and motion without a co-occurrence feature and hence fewer samples could produce sufficient variations. The performance of the 3DHOG detector improved greatly as the training sample size increased because the co-occurrence gradient feature requires extensive training samples to avoid the generalization error caused by over-training.

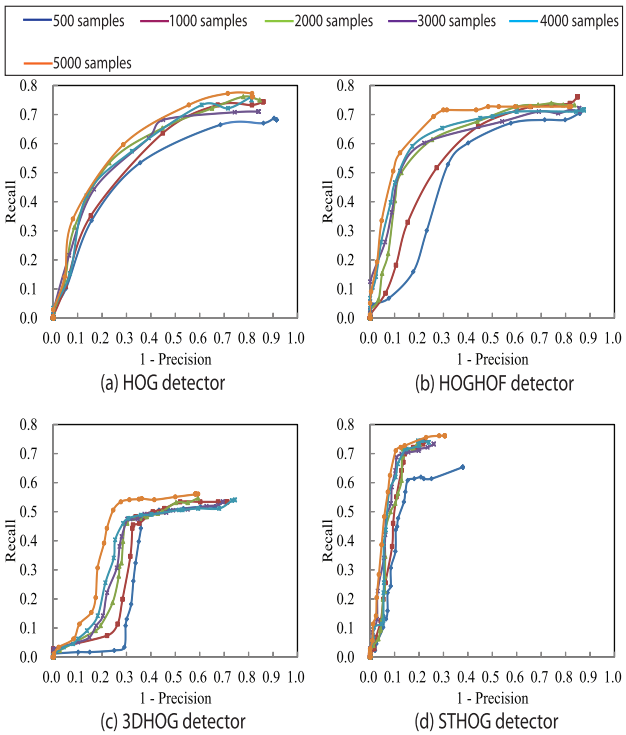

Fig. 13 RPC variations for the STHOG, HOG, HOGHOF, and 3DHOG detectors using different numbers of training samples with the PETS2009 View02 dataset.

5.3.5 STHOG Detector with Static Persons

One suspicion is the performance of our STHOG pedestrian detector with the static persons. Since there will be little motion information extracted from those static people, the temporal histogram in STHOG feature will be compressed. Therefore, in such case, a STHOG detector will only use the appearance feature to detect a people. Figure 14 shows the comparative experiments between STHOG and HOG detectors with static persons under an elemental school scene. In this scene, two girls were standing in the center of images, where the right girl kept static while the left girl moved around through the sequence. Since some similar positive samples were included into our training dataset (images were captured from this environment on other days), both detectors could find the static girl. However, since another girl moved around and changed her appearance, the HOG detector failed to find her, while the STHOG detector was successful to detect her due to her clear motion information. The STHOG detector failed to detect the moving girl in the first frame, because another person's leg was included into her detection window from the left-top and polluted her STHOG feature in the window. That was also the failure reason of HOG detector in the first frame. Here, we would like to argue that if there were no similar positive sample in the training dataset, it would be very possible that neither the STHOG nor the HOG detector could detect those unknown static people. To avoid such problem, various positive training samples should be included.

In general, when running our STHOG detector on a desktop PC with an Intel C2D 3.16 GHz CPU and 4 GB memory, it takes 4 to 5 seconds to obtain multiple-scale de-

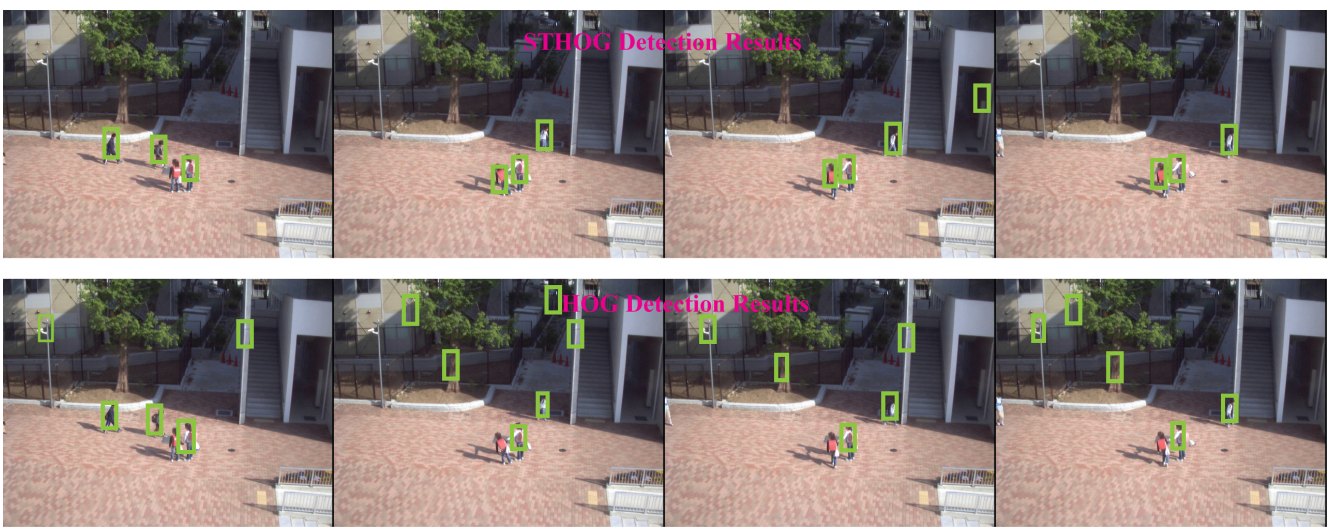

Fig. 14 Comparative experiment between STHOG and HOG detectors with static people. Top row: the detection results of STHOG detector; Bottom row; the detection results of HOG detector. By including similar samples in the training samples, either detector could find the right static girl, but STHOG detector was superior in detecting the left moving girl.

tection results for a 720×570 pixel image by scanning over the image pyramid for all positions and scales. We believe that the sparse sampling in the image pyramid [31] greatly increased the processing speed.

6. Conclusions and Future Research

In this paper, we proposed an algorithm for pedestrian detection using spatio-temporal features. By describing the appearance and motion features with separate spatial and temporal histograms, our STHOG detector is more capable of detecting a pedestrian in cluttered situations. This is because not only the appearance information but also the unique human movement generated by the movement of arms/legs distinguishes a pedestrian from the background. We also proved that even when trained with a fairly small training dataset, the STHOG pedestrian detector still achieves high detection performance for various complex datasets.

Future work on our STHOG detector will be influenced by the following: 1) since the integration of initial detection results greatly affects the final evaluation, better clustering integration is always required; 2) the occlusion and overlapping problems need to be tackled in conjunction with the part-based detector; 3) temporal tracking through the detection results may also improve the recall rate; and 4) performance of the STHOG detector on a mobile platform should be investigated.

References

[1] B.K.P. Horn and B.G. Schunk, "Determining optical flow," Artificial Intelligent, vol.17, pp.185–203, 1981.
[2] Y.Z. Cheng, "Mean shift, mode seeking, and clustering," IEEE Trans. Pattern Anal. Mach. Intell., vol.17, no.8, pp.790–799, 1995.
[3] S. Agawal and D. Roth, "Learning a sparse representation for object detection," 7th Euro. Conf. Compu. Vision, 2002.
[4] P. Viola and M. Jones, "Robust real-time face detection," Inter. J. Compu. Vision, vol.57, pp.137–154, 2004.
[5] P. Viola and M. Jones, "Robust real-time object detection," Inter. J. Compu. Vision, 2001.
[6] P. Viola, M. Jones, and D. Snow, "Detecting pedestrians using patterns of motion and appearance," Inter. J. Compu. Vision, vol.63, no.2, pp.13–161, 2005.
[7] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," IEEE Conf. Comp. Vis. Patt. Reco., pp.886–893, 2005.
[8] N. Dalal, B. Triggs, and C. Schmid, "Human detection using oriented histograms of flow and appearance," Euro. Conf. Compu. Vision, pp.428–441, 2006.
[9] Q. Zhu, S. Avidan, M. Yeh, and K. Cheng, "Fast human detection using a cascade of histograms of oriented gradients," IEEE Conf. Comp. Vis. Patt. Reco., pp.1491–1498, 2006.
[10] S. Munder and D.M. Gavrila, "An experimental study on pedestrian classification," IEEE Trans. Pattern Anal. Mach. Intell., vol.28, no.11, pp.1863–1868, 2006.
[11] B. Leibe, N. Cornelis, and L.V. Gool, "Dynamic 3D scene analysis from a moving vehicle," IEEE Conf. Comp. Vis. Patt. Reco., 2007.
[12] B. Wu and R. Nevatia, "Detection and tracking of multiple, partially occluded humans by Bayesian combination of edgelet based part detectors," Inter. J. Compu. Vision, vol.75, no.2, pp.247–266, 2007.
[13] D.M. Gavrial and S. Munder, "Multi-cue pedestrian detection and tracking from a moving vehicle," Inter. J. Compu. Vision, vol.73, no.1, pp.41–59, 2007.
[14] T. Gandhi and M.M. Trivedi, "Pedestrian detection systems: Issues, survey and challenges," IEEE Trans. Intel. Transportation Systems, vol.8, no.3, pp.413–430, 2007.
[15] O. Tuzel, F. Porikli, and P. Meer, "Human detection via classification on Riemannian manifolds," IEEE Conf. Comp. Vis. Patt. Reco., 2007.
[16] I. Laptev and P. Perez, "Retrieving actions in movie," Inter. Conf. Comp. Vis., 2007.
[17] P. Scovanner, S. Ali, and M. Shah, "A 3-dimensional SIFT descriptor and its application to action recognition," Intern. Conf. Multimedia, pp.357–360, 2007.
[18] Y. Yamauchi and H. Fujiyoshi, "People detection based on co-occurrence of appearance and spatiotemporal features," Inter. Conf. Patt. Recog., 2008.
[19] M. Andriluka, S. Roth, and B. Schiele, "People-tracking-by-detection and people-detection-by-tracking," IEEE Conf. Comp. Vis. Patt. Reco., 2008.
[20] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," IEEE Conf. Comp. Vis. Patt. Reco., 2008.
[21] A. Kläser, M. Marszalek, and C. Schmid, "A spatio-temporal descriptor based on 3D-gradient," Brit. Mach. Vis. Conf., pp.995–1004, 2008.
[22] M. Andriluka, S. Roth, and B. Schiele, "People-tracking-by-detection and people-detection-by-tracking," IEEE Conf. Comp. Vis. Patt. Reco., 2008.
[23] Y. Murai, H. Fujiyoshi, and T. Kanade, "Combined object detection and segmentation by using space-time patches," Asia. Conf. Comp. Vis., pp.915–924, 2009.
[24] J. Sun, X. Wu, S.C. Yan, L.F. Cheong, T.S. Chua, and J.T. Li, "Hierarchical spatio-temporal context modeling for action recognition," IEEE Conf. Comp. Vis. Patt. Reco., 2009.
[25] P. Dollar, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: A benchmark," IEEE Conf. Comp. Vis. Patt. Reco., 2009.
[26] H. Wang, M.M. Ullah, A. Kläser, I. Laptev, and C. Schmid, "Evaluation of local spatio-temporal features for action recognition," Brit. Mach. Vis. Conf., 2009.
[27] H.Z. Wang, X. Han, and S.C. Yan, "An HOG-LBP human detector with partial occlusion handling," Inter. Conf. Comp. Vis., 2009.
[28] M. Enzweiler and D.M. Gavrila, "Monocular pedestrian detection survey and experiments," IEEE Trans. Pattern Anal. Mach. Intell., vol.31, no.12, pp.2179–2195, 2009.
[29] P.F. Felzenszwalb, R.B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," IEEE Trans. Pattern Anal. Mach. Intell., vol.32, no.9, pp.1627–1645, 2010.
[30] P.F. Felzenszwalb, R.B. Girshick, and D. McAllester, "Cascade object detection with deformable part models," IEEE Conf. Comp. Vis. Patt. Reco., 2010.
[31] P. Dollar, S. Belongie, and P. Perona, "The fastest pedestrian detection in the west," Brit. Mach. Vis. Conf., 2010.
[32] L. Bourdev, S. Maji, T. Brox, and J. Malik, "Detecting people using multually consistent poselet activations," Euro. Conf. Compu. Vision, 2010.
[33] S. Walk, N. Majer, K. Schindler, and B. Schiele, "New features and insights for pedestrian detection," IEEE Conf. Comp. Vis. Patt. Reco., 2010.
[34] J. Marin, D. Vazquez, D. Geronimo, and A.M. Lopez, "Learning appearance in virtual scenarios for pedestrian detection," IEEE Conf. Comp. Vis. Patt. Reco., 2010.
[35] D. Geronimo, A.M. Lopez, A.D. Sappa, and T. Graf, "Survey of pedestrian detection for advanced driver assistance systems," IEEE Trans. Pattern Anal. Mach. Intell., vol.32, no.7, pp.1239–1258, 2010.
[36] P. Dollar, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: An evaluation of the state of the art," IEEE Trans. Pattern Anal. Mach. Intell., 2011.

[37] M. Wang and X.G. Wang, "Automatic adaption of a generic pedestrian detector to a specific traffic scene," IEEE Conf. Comp. Vis. Patt. Reco., 2011.
[38] D. Park, D. Ramanan, and C. Fowlkes, "Multiresolution models for object detection," Euro. Conf. Compu. Vision, 2010.
[39] Y. Liu, S. Shan, X. Chen, J. Heikkila, W. Gao, and M. Pietikainen, "Spatial-Temporal Granularity-tunable Gradients Partition (STGGP) descriptor for human detection," Euro. Conf. Compu. Vision, 2010.
[40] O. Barinova, V. Lempisky, and P. Kohli, "On the detection of multiple object instances using hough transforms," IEEE Conf. Comp. Vis. Patt. Reco., 2010.
[41] P.M. Roth, S. Sternig, H. Grabner, and H. Bischof, "Classifier grids for robust adaptive object detection," IEEE Conf. Comp. Vis. Patt. Reco., 2009.
[42] S. Sternig, P.M. Roth, and H. Bishcof, "Learning of scene-specific object detectors for classifier Co-grids," Int'l Conf. on Advanced Video and Signal-Based Surveillance, 2010.

Chunsheng Hua received his BE degree in electronic engineering from Shenyang University of Technology in 2001. He received his MS degree from the Department of Mechanical and System Engineering at Kyoto Institute of Technology in 2004 and his Ph.D. degree in computer vision from the graduated school of system engineering at Wakayama University in 2007. From 2007 to 2010, he worked as postdoctoral researcher at the Institute of Scientific Industrial Research of Osaka University and was promoted to be a specially assigned assistant professor in 2010. Since October 2012, he joined the Shenyang Institute of Automation Chinese Academy of Sciences as a professor. He is a member of IPSJ. He received the Funai award of IPSJ Digital Courier for young researchers, Yamashita Memorial Award of IPSJ and SCIE award in 2006, 2008 and 2009, respectively. His research interests include machine learning, pattern recognition, clustering algorithms, object tracking, pedestrian detection and sensor fusion, etc.

Yasushi Makihara was born in Japan in 1978, and received the B.S., M.S., and Ph.D. degrees in Engineering from Osaka University in 2001, 2002, and 2005, respectively. He is currently an Assistant Professor of the Institute of Scientific and Industrial Research, Osaka University. His research interests are gait recognition, morphing, and temporal super resolution. He is a member of IPSJ, RJS, and JSME.

Yasushi Yagi is the Director of the Institute of Scientific and Industrial Research, Osaka university, Ibaraki, Japan. He received his Ph.D. degrees from Osaka University in 1991. In 1985, he joined the Product Development Laboratory, Mitsubishi Electric Corporation, where he worked on robotics and inspections. He became a Research Associate in 1990, a Lecturer in 1993, an Associate Professor in 1996, and a Professor in 2003 at Osaka University. International conferences for which he has served as Chair include: FG1998 (Financial Chair), OMINVIS2003 (Organizing chair), ROBIO2006 (Program co-chair), ACCV2007 (Program chair), PSVIT2009 (Financial chair), ICRA2009 (Technical Visit Chair), ACCV2009 (General chair), ACPR2011 (Program co-chair) and ACPR2013 (General chair). He has also served as the Editor of IEEE ICRA Conference Editorial Board (2007–2011). He is the Editorial member of IJCV and the Editor-in-Chief of IPSJ Transactions on Computer Vision & Applications. He was awarded ACM VRST2003 Honorable Mention Award, IEEE ROBIO2006 Finalist of T.J. Tan Best Paper in Robotics, IEEE ICRA2008 Finalist for Best Vision Paper, MIRU2008 Nagao Award, and PSIVT2010 Best Paper Award. His research interests are computer vision, medical engineering and robotics. He is a fellow of IPSJ and a member of RSJ and IEEE.