

# Onboard monocular pedestrian detection by combining spatio-temporal hog with structure from motion algorithm

Chunsheng Hua · Yasushi Makihara · Yasushi Yagi ·  
Shun Iwasaki · Keisuke Miyagawa · Bo Li

Received: 25 November 2013 / Revised: 22 October 2014 / Accepted: 5 November 2014 / Published online: 3 January 2015  
© Springer-Verlag Berlin Heidelberg 2014

**Abstract** In this paper, we brought out a novel pedestrian detection framework for the advanced driver assistance system of mobile platform under the normal urban street environment. Different from the conventional systems that focus on the pedestrian detection at near distance by interfusing multiple sensors (such as radar, laser and infrared camera), our system has achieved the pedestrian detection at all (near, middle and long) distance on a normally driven vehicle (1–40 km/h) with monocular camera under the street scenes. Since pedestrians typically exhibit not only their human-like shape but also the unique human movements generated by their legs and arms, we use the spatio-temporal histogram of oriented gradient (STHOG) to describe the pedestrian appearance and motion features. The shape and movement of a pedestrian will be described by a unique feature produced by concatenating the spatial and temporal his-

tograms. A STHOG detector trained by the AdaBoost algorithm will be applied to the images stabilized by the structure from motion (SfM) algorithm with geometric ground constraint. The main contributions of this work include: (1) ground constraint with monocular camera to reduce the computational cost and false alarms; (2) preprocessing by stabilizing the successive images captured from mobile camera with the SfM algorithm; (3) long-distance (maximum 100 m) pedestrian detection at various velocities (1–40 km/h). Through the extensive experiments under different city scenes, the effectiveness of our algorithm has been proved.

**Keywords** Spatio-temporal HOG · Pedestrian detection · Onboard monocular camera · Structure from motion

---

B. Li has left the Honda R&D Co., Ltd, Japan.

---

C. S. Hua (✉)  
The State Key Lab of Robotics, Shenyang Institute of Automation,  
Chinese Academy of Sciences, Shenyang 110045, Liaoning, China  
e-mail: huachunsheng@gmail.com

Y. Makihara · Y. Yagi  
Department of Intelligent Multimedia,  
ISIR of Osaka University, Osaka 560-0023, Japan  
e-mail: makihara@am.sanken.osaka-u.ac.jp

Y. Yagi  
e-mail: yagi@am.sanken.osaka-u.ac.jp

S. Iwasaki · K. Miyagawa · B. Li  
Honda R&D Co., Ltd, Wako, Japan  
e-mail: Shun\_Iwasaki@n.t.rd.honda.co.jp

K. Miyagawa  
e-mail: Keisuke\_Miyagawa@n.t.rd.honda.co.jp

B. Li  
e-mail: snakeie@gmail.com

## 1 Introduction

Due to the explosive increase of automobiles over the world in the last two decades, the requirement for protecting both the drivers and pedestrians has been kept on rising. Correspondingly, many protection systems have been proposed to improve the traffic safety, such as the seat belt, electronic stabilization program and air bag. However, most of those systems only focus on protecting the drivers. The most serious traffic accident vehicle-to-pedestrian accident still remains as a challenging task.

To avoid the vehicle-to-pedestrian accidents, more intelligent protection systems are required to anticipate accidents. Such systems are usually called as the advanced driver assistant systems (ADAS) which include the auto control, driver monitoring and pedestrian protection, etc. In this paper, we focus on the kernel task of ADAS—“pedestrian detection” by applying the visible spectrum (VS) monocular camera to

locate the pedestrians so as to avoid the vehicle-to-pedestrian accidents.

With years of research and great efforts from both the academic researcher and manufactures, many onboard pedestrian detection systems have been brought out. Most of those proposed systems mainly rely on combining the multiple sensors (such as the radar, laser range finder, far/near infrared camera and stereo camera) to locate the pedestrians. In 2008–2009, the BMW 7 Series and Mercedes-Benz E-Class vehicles were implemented with the pedestrian protection system that could work at night. By combining the monocular camera with radar, the Volvo S60 has achieved the collision warning system that can detect the people and stop the vehicle, automatically. Although, all these systems could work efficiently under some conditions, detecting pedestrians at long distance still remains as a difficult task. This is because although the depth sensors could find an obstacle at long distance, the classification of detected obstacle still relies on the pedestrian detection algorithms with VS camera. As mentioned in [1], most of the current detectors are designed for near distance (which means that pedestrians are higher than 80 pixels in the image) and perform poor even at medium distance (pedestrian height varies from 30–80 pixels).

The difficulties of onboard pedestrian detection with VS monocular camera include:

- The detection must be performed under the highly dynamic urban city scene. Here, the image quality from VS camera is usually poorer than the surveillance scenes. This is because both the illumination and weather condition are unpredictable, where the strong shadow or poor contrast in the visible spectrum may dramatically degrade the image quality.
- The appearance of both the pedestrian and background are highly variable due to the various viewpoints. Since both the pedestrian and camera are in motion, the pedestrian detection at different view angles (front, lateral, rear) and distance becomes extremely difficult, because the pedestrian appearance feature will become unclear or blurred at the long distance. Since most current pedestrian detectors require the clear pedestrian features, people at long distance remains as a challenging task for those algorithms (detailed discussion could be found in [1]).

Since the conventional onboard pedestrian detection algorithms mainly apply the appearance feature to describe the human and require the clear target feature for discriminating a pedestrian from background, they tend to suffer from the huge human appearance variation (generated by the different clothes, camera viewpoints, articulated pose, baggage, etc.) and the blurred human image at long distance.

In this paper, to overcome such problems, the temporal (or motion) features are introduced to improve the detec-

tion performance based on the fact that not only human shape but also human motion makes a pedestrian different from the surrounding environment. The pedestrian feature is described by a spatio-temporal histogram of oriented gradients which corresponds to his/her appearance and motion features. To achieve the pedestrian detection at long distance (which means the pedestrian height will be smaller than 30 pixels in the image), the images captured from the monocular VS camera are stabilized according to the SfM algorithm, through which the appearance and motion features of pedestrians could be smoothly extracted from two adjacent frames. The main contributions of our work include:

- Describing the pedestrian feature with a spatio-temporal histogram of oriented gradients (STHOG). Since pedestrians will exhibit not only their human-like shapes but also the unique human-like movements caused by their arms/legs, a pedestrian's appearance is described by the HOG feature, and his/her motion is represented by another histogram of temporal gradients computed from successive frames. Compared with the appearance features, the STHOG feature is more invariant to the changes of clothes and particularly powerful in discriminating pedestrians from cluttered background, where some background objects happen to contain the human-like shape but their motion is different from that of human.
- Geometric ground constraint according to the SfM algorithm. Since it is reasonable to assume that a pedestrian will only walking on the ground but not in the sky, the geometric constraint of a planar ground will reduce not only the computational cost but also the false alarms as the region higher than the ground (like road mark or the house roof). Here, the depth of each image pixel could be computed from the SfM by estimating the camera motion with RANSAC method. A constraint planar ground is constructed by the pixels whose depth varies 10–100 m (explanation for selecting such values could be found in Sec.3.2).
- Preprocessing for stabilizing the images captured from the monocular camera. Since the camera on a vehicle may move fast, the location of a pedestrian in two adjacent images may be discontinuous. We brought a stabilization system to estimate the pixel correspondence between two images according to the vehicle velocity and camera parameters. With the estimated pixel correspondence, we could not only smoothly compute the temporal motion from two frames but also predict the pedestrian position in the next frame which can help us to reduce the computational cost and false alarms.

This paper is organized as follows: in Sect. 2, we provide a brief review for related work about the onboard pedestrian detection; in Sect. 3, there is a detailed introduction for the

proposed system; Sect. 4 shows the comparative experimental results among the proposed system and other pedestrian detectors.

## 2 Related work

### 2.1 Pedestrian detection by the VS camera

Geronimo et al. [2] split the architectures of onboard pedestrian detection into 6 modules as: preprocessing, foreground segmentation, object classification, verification (or refinement), tracking and application. In this paper, we roughly simplify them into four main parts as: (1) preprocessing; (2) geometric constraint; (3) pedestrian detection; and (4) refinement. Although, the processing turn of pedestrian detection and geometric filter can be reversed, applying geometric filter after the pedestrian detection will be computational expensive. Therefore, in our work, we refer to applying the geometric constraint before pedestrian detection.

Since so much efforts and so many powerful algorithms have been brought out for the task of onboard pedestrian detection, covering all the related works has been beyond this paper. Good survey on this topic could be found in [1–3], where they compared the recent onboard pedestrian detection algorithms in detail and described how to design the onboard pedestrian detection system.

#### 2.1.1 Preprocessing

The task of preprocessing may include the camera calibration (stereo camera) or adjusting the exposing time of camera (both the stereo and monocular camera systems). The motivation to applying preprocessing lies in the fact that: when a vehicle moves in the city scene, some typical environments may degrade the image quality dramatically, such as the bridge tunnels, rapid motion of the vehicle itself or background (other on-coming vehicles) and abrupt reflection from the glasses. Without the preprocessing, the following modules of pedestrian detection will suffer from the poor images, which may lead to the collapse of the whole system.

The high dynamic range (HDR) method [4] can provide the high contrast images from the aforementioned urban scenes, it is reported that this method can also be applied to the near infrared (NIR) cameras to work at night. As for the stereo camera system, the camera calibration [5,6] is the main task in preprocessing. They usually firstly compute the intrinsic camera parameters and assume them to be fixed, then the extrinsic parameters will be continuously updated.

#### 2.1.2 Geometric constraint

The geometric filter/constraint could be performed by both the monocular and stereo camera system to filter out the

unreasonable pedestrian candidates. In [7], the parameters of monocular camera pose are obtained by the training results from a set of labeled images, and such camera pose helps to build the 3D geometry constraint. In [8], a V-Disparity space is produced by accumulating the disparity from stereo camera in the vertical direction to estimate the road slope and depth of a point. Other works [9] use the Euclidean space for the 3D point fitting. Ess et al. [10–12] estimate the ground plane from the Structure-from-Motion. The continuously estimated ground plane can be used for updating the camera calibration of stereo camera and filtering out the pedestrian candidates which are out of the ground plane. Keller et al. [13] brought out a novel onboard safety driving assistant system. From the 6D-Vision system, the optical flow computed from the stereo camera can identify the moving pedestrian from static vehicles at real time. Further object recognition is achieved by the HOG/Linear SVM which will be discussed in detail later. In [3], Gandhi et al. discuss about the way and the problems in designing an onboard pedestrian protection system, where they prove that the Omni cameras could also be applied for producing the 3D disparity map.

#### 2.1.3 Pedestrian detection

Since predicting the pedestrian candidates from the images is the kernel of an onboard pedestrian detection system, many efforts [14–20] have been applied to solve this problem. By taking the silhouette shape as the pedestrian feature, Gavrilu [21,22] proposed a hierarchical shape classification system based on the Chamfer matching algorithm. More recent researches focus on using the appearance descriptors to represent the pedestrian feature. Papageorgiou and Poggio [23] propose a Haar-like integral image to represent the local contrast between the target object and its surrounding background, and a Haar-feature-based detector is obtained by training this feature with support vector machine (SVM). Viola and Jones [24] further improve this feature with the AdaBoost cascade structure, the motion information and Haar-like feature are combined for the pedestrian detection with surveillance camera. The success of this cascade-structure detector is based on the assumption that the overwhelming majority of input images are background.

The HOG [25] 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. [26] achieved almost real-time single-scale pedestrian detection by training a cascade detector with an integral HOG feature, while keeping almost the same detection performance as [25]. Further improvements by including the spatial co-

occurrence within a single window for each region of the HOG [27] have been reported.

Wu et al. [28] 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 [29,30] handles partial occlusion by integrating multiple hypotheses. 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 [31] 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. [32] improved the HOG detector when dealing with a wide range of pedestrian appearance changes in scales. Barinova et al. [33] 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. [34] 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 [35].

More recent work [36,37] shows that the combination of deformable part and holistic models will greatly improve the performance of traditional algorithms (HOG or Haar-like features). By such combination of richer models and more sophisticated learning techniques, they greatly improve the detection rate under the complex datasets. In [1,38], it has been proved that under the same framework, the richer feature descriptor could outperform those simple descriptors. In their work, the combination of HOG, motion (optical flow) and self-similarity color histogram greatly increases the detection rate. An interesting outcome from [39] was that the virtual images can also be used for training a pedestrian detector. Wang et al. [40] 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. Rodriguez et al. [41] pay more attention to counting the pedestrians under the crowd scene.

They applied state-of-the-art detector of [36] to roughly locate all the possible pedestrians and estimate the number of pedestrians by minimizing the cost energy function that evaluates the detection score, overlapping information and penalizing the crowd density.

Other algorithms such as temporal tracking through detection results [29,42] have been reported. Based on the assumption that, in the temporal space, the consistency of a correct pedestrian is more robust than that of an abrupt false positive on the background, tracking the pedestrian trajectory can further improve the detection rate and reduce the false alarm rate.

As aforementioned, since most of the present pedestrian detection algorithms use the appearance information to describe the pedestrian feature, they usually suffer from the huge variation of pedestrian appearance and cluttered background where some background objects may happen to contain the human-like appearance. The normal solution is to prepare the extensive positive/negative training datasets to cover the pedestrian/background appearance variation as much as possible. However, such extensive training dataset usually leads to the over-training problem.

To overcome such problems, the temporal or motion features [43] are introduced to improve the detection performance based on the fact that is not only its appearance but also its motion that makes a target object different from others. In the case of pedestrian detection, a pedestrian is assumed to have both human shape and a unique human movement such as the periodic movement of arms and legs. Therefore, the spatio-temporal features are considered to be more powerful than appearance features in pedestrian detection. Such idea has been proved to be useful in action recognition [44–51] and patch-based image segmentation [27,52].

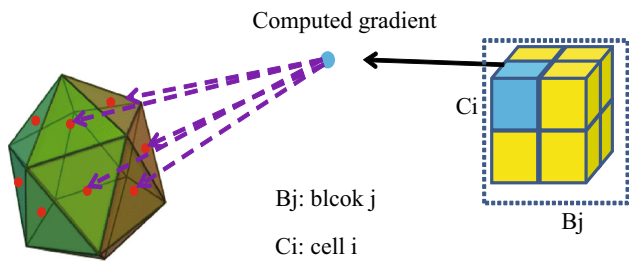
In this paper, we proposed a novel spatio-temporal HOG pedestrian detector for discriminating a pedestrian from the background (details in Sect. 3.4.1). The following related spatio-temporal feature descriptors are proposed for action recognition.

#### 2.1.4 Related spatio-temporal feature descriptors

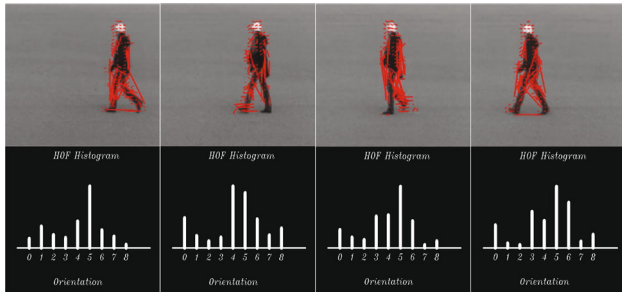
##### 3DHOG feature descriptor

Kläser et al. [47,53] achieved action recognition under complex background conditions 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. 1). 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 orien-





**Fig. 1** Illustration of the 3DHOG descriptor from [47]. The orientation of a computed spatio-temporal gradient will be calculated by projecting it onto an icosahedron



**Fig. 2** Illustration of the histograms for optical flow [44,46]. The optical flow will be represented by a histogram according to their orientation

tations. 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 [47], a 3DHOG descriptor was applied around some robustly extracted feature points, instead of directly to the full image. Our experimental results in Sect. 4 also prove that applying the 3DHOG descriptor directly to the pedestrian detection task is unsuitable.

### HOGHOF feature 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 [54] and action recognition [44,46]. To characterize the local motion and appearance, histograms of the oriented gradient and optical flow accumulated in space-time are concatenated and normalized (as shown in Fig. 2). Walk et al. [38] tried to combine the HOG with optical flow (HOF) [55] 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).

### STGGP descriptor

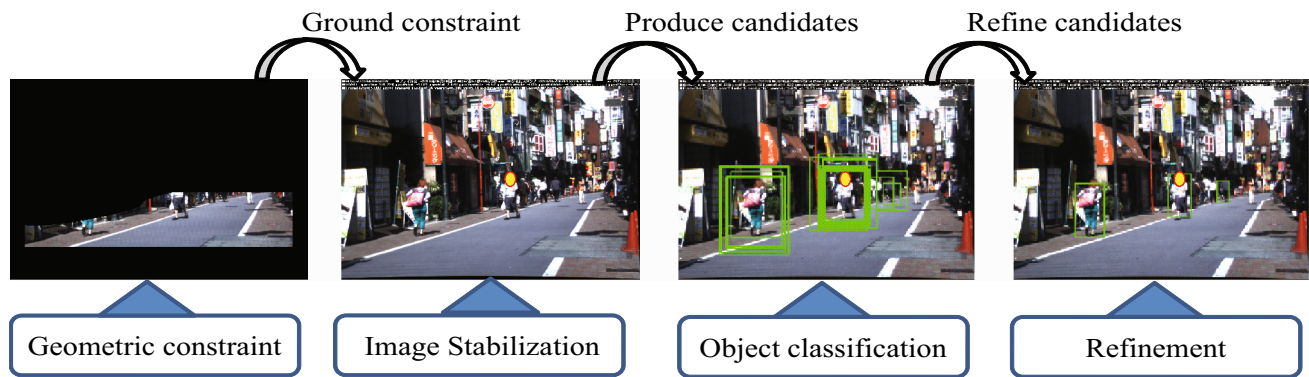
Liu et al. [56] proposed the spatial-temporal granularity-tunable gradient partition (*STGGP*) descriptor for pedestrian detection. The orientation of spatial-temporal gradients attributed to humans was described in a 3D Hough space, while a generalized plane was produced to partition the gradients by 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 tradeoff 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.

#### 2.1.5 Refinement

As for the refinement, we usually require that one target person only responds to his/her unique detection result. After the detection process, the pedestrian detector will produce a density map of correct detection results consist of their score, scale and position over the whole image. Then, the nonmaximum suppression algorithm will be applied to this density map to find the peaks that correspond to pedestrians. For such sake, mean shift [57] clustering is a good choice to find the local maximum through the whole image. Other work [33] uses the Hough transformation to describe object shape and the local nonmaximum suppression is achieved by minimizing an energy function in the Hough Forest image.

#### 2.2 Pedestrian detection by the other onboard system

Besides the VS camera, other sensors (such as NIR, thermal infrared (TIR) camera, radar and laser scanner) are also applied for the onboard pedestrian detection task. The motivation for using those extra sensors lies in that they can



**Fig. 3** The architecture of our onboard monocular pedestrian detection system. The *arrows* refer to the function of next module. The object classification will be performed after the image stabilization process. The output of object classification will be integrated into reasonable groups by the refinement process

remove the inflexible noise that the VS camera suffers from or can provide the useful 3D depth information to filter out the unreasonable detection results (like the pedestrians in the sky or on the roof of houses). Furthermore, the image analysis with VS camera is far away from perfect, many factors such as illumination, motion blur and clustering integration will affect the final detection results.

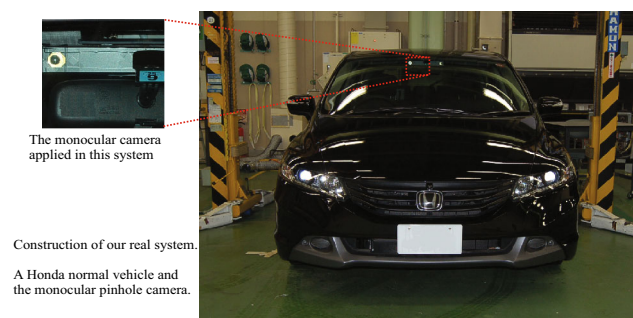
However, the TIR or NIR images could also be affected by the other hot objects (other vehicles or engines), weathers and seasons (summer or winter). It is reported [58] that pedestrians could be either warmer or colder than the surrounding background, according to different factors. The radar or laser scanner can only find an obstacle but could not tell us what it is.

Therefore, the fusion of VS camera with TIR/NIR or others sensors becomes a natural and reasonable choice. Fardi [59] applied the Kalman filter to combine the laser scanner with TIR camera for pedestrian detection. In [60], radar, velocity measurer, VS and TIR cameras are combined together for producing and verifying the pedestrian shape hypothesis.

In [61], the VS and TIR are combined for pedestrian detection, where the stereo VS camera produce the geometric constraint and the final detection results are refined in the TIR. In [62], Radar, VS and TIR cameras are implemented together for pedestrian detection. The accurate 3D ground plane is produced by the radar data and the TIR helps to filter other cold background, and the final detection results are verified by the VS camera according to the histogram of oriented edges.

### 3 Our onboard monocular pedestrian detection system

As shown in Fig. 3, the architecture of our system is composed of: geometric constraint, image stabilization, object classification (STHOG pedestrian detection), and refine-



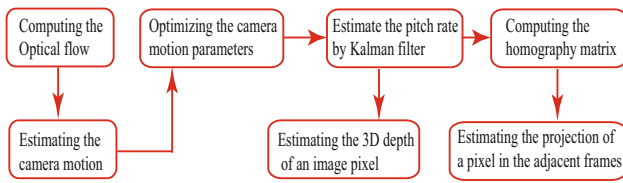
**Fig. 4** The camera setup of our system. The training and test images are captured by a monocular PointGray Camera

ment. The camera setup of our system is shown in Fig. 4, where the PointGray camera and  $47^\circ$  HFOV lens are applied. The video resolution is  $800 \times 600$  pixels which is trimmed from the  $1280 \times 960$  pixel-size images, and the overall image quality is lower than that of the images captured by the static camera with the same resolution.

#### 3.1 Structure from motion with monocular camera

As for an onboard pedestrian detection system, the real-world 3D information is always required not only for detecting the pedestrians but also for braking the vehicle to avoid the collision. Besides the aforementioned range sensors (such as radar or laser scanner), the Structure from Motion (SfM) algorithm could also provide the rather accurate 3D information from the 2D images with a monocular camera.

Figure 5 shows the flow chart of the SfM applied in our system to estimate the 3D depth information and homography matrix of image plane. Here, the Harris corner detector [63] is applied to extract the feature point from the successive frames and the optical flow of these feature points is estimated by the Lucas–Kanade algorithm [64]. The camera motion is estimated by the SfM of [65], where the epipolar constraint



**Fig. 5** The flow chart of SfM applied in our system

is set up by assuming the images captured by a monocular camera at time  $t$  and  $t + 1$  as two independent frames. In this way, a monocular camera can be considered as stereo camera along the time direction, and the essential matrix  $E$  about camera parameter could be computed by the RANSAC [66] method. Here, the matrix  $E$  will satisfy:

$$\mathbf{P}_1^T E \mathbf{P}_2 = 0, \tag{1}$$

$$\mathbf{P} = (x, y, 1), \tag{2}$$

where  $\mathbf{P}_1, \mathbf{P}_2$  represent the projection point in frame  $t$  and  $t + 1$ . The optimization for the essential matrix is achieved by the Bundle Adjustment [67] method to minimize Eq. (3), where the Gauss–Newton method is applied to Eq. (4) (which is the Taylor expansion of Eq. (3)).

$$f(x) = \sum \left[ \frac{1}{2} (\mathbf{P}_1^T E(x) \mathbf{P}_2)^2 \right], \quad x = (R, t_{\text{trans}}), \tag{3}$$

$$f(x + \delta x) = f(x) + g^T \delta x + \frac{\delta x^T H \delta x}{2}, \tag{4}$$

where  $g$  is the derivation of  $f$  as  $g = \nabla f$  and  $H$  is the Hessian matrix of  $g$  as  $H = \nabla \nabla^T f = \nabla g^T$ ,  $R$  and  $t_{\text{trans}}$  represent the rotation and translation of camera. Details about how to extract  $R$  and  $t_{\text{trans}}$  from an essential matrix could be found in [68].

The camera pose (pitch rate  $\theta$ ) is the included angle between the camera optical axis and ground plane. Here, we select the Kalman filter to estimate the camera translation  $\bar{\mathbf{t}}_{\text{trans}} = (t_{\text{trans}(x)}, t_{\text{trans}(y)}, t_{\text{trans}(z)})$ , and the pitch rate  $\theta$  could be computed as:

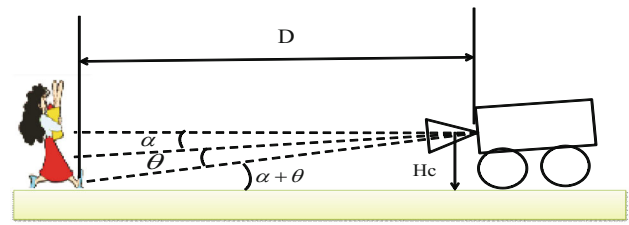
$$\theta = \tan^{-1} \left( \frac{-t_{\text{trans}(y)}}{\sqrt{t_{\text{trans}(x)}^2 + t_{\text{trans}(z)}^2}} \right) \tag{5}$$

As shown in Fig. 6, from the camera pitch rate  $\theta$ , we can estimate the depth information ( $D$ ) of a given image pixel  $\bar{\mathbf{X}}_t = (x, y, t)$  as follows:

$$D = \frac{H_c}{\tan(\alpha + \theta)}, \tag{6}$$

$$\alpha = \tan^{-1} \left( \frac{y - y_c}{f} \right) \tag{7}$$

where  $H_c$  is the height of camera,  $\alpha$  means the depression angle between the camera optical axis and image pixel  $\bar{\mathbf{X}}_t$ ,



**Fig. 6** The illustration of estimating the depth from pitch rate according to the SfM algorithm. The estimated depth will be applied for the further image stabilization and object classification

$y_c$  is the height of image center and  $f$  is the optical focus length of camera.

To estimate the corresponding position of a given image pixel  $\bar{\mathbf{X}}_t$  in its adjacent frames  $t + 1$  (or  $t - 1$ ), we need to compute the homography matrix  $H_{\text{homo}}$  of the camera movement as:

$$H_{\text{homo}} = R^T - R t_{\text{trans}} \begin{pmatrix} n \\ d \end{pmatrix}^T, \tag{8}$$

where  $n$  is the normal vector that is orthogonal to  $\alpha$ , and  $d$  is the distance from the camera center to the ground plane where a pedestrian is assumed to stand in. Then, the corresponding pixel  $\bar{\mathbf{X}}_{t+1} = (x_{(nxt)}, y_{(nxt)}, t + 1)$  to  $\bar{\mathbf{X}}_t$  could be calculated as:

$$\bar{\mathbf{X}}_{t+1} = H_{\text{homo}}^{t \rightarrow t+1} \bar{\mathbf{X}}_t. \tag{9}$$

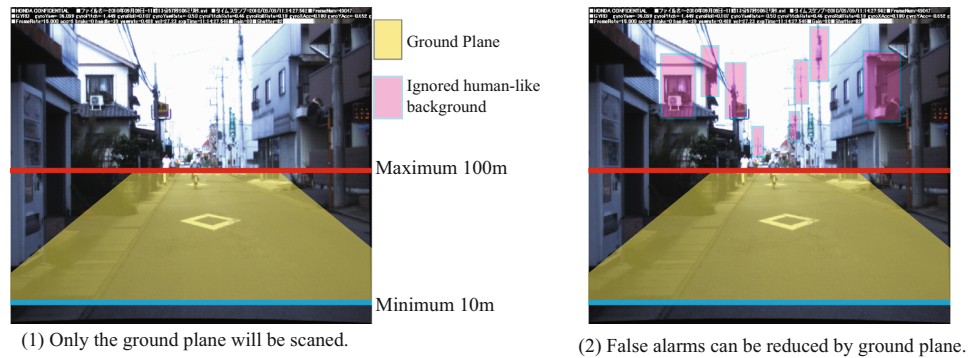
Once the rotation matrix  $R$  and translation vector  $t_{\text{trans}}$  are obtained, the pitch rate  $\theta$ , depth  $D$  to the foot bottom point of the target pedestrian standing on the ground plane and the homography matrix  $H_{\text{homo}}$  could be computed according to Eq. (5-9). The pitch rate  $\theta$  is used to compute the depth  $D$  as well as to relate the two positions between two successive frames via the homography matrix  $H_{\text{homo}}$  in Eq. 9.

### 3.2 Geometric constraint by SfM algorithm

The 3D depth distance estimated from Eq. (6) will be used to construct our geometric ground constraint which will not only reduce the computational cost but also filter out the unreasonable pedestrian candidates. Since it requires a long distance to stop a vehicle when it is moving fast (e.g., 40 km/h), we choose a range from 10 to 100 m to be the search range of our detection system. The reason for setting up the maximum depth range as 100m lies in the requirement of our system, where 100m is the braking distance for stopping a vehicle at the speed of 100km/h. While the distance from a pedestrian to vehicle is less than 10m, it will be too short to stop the vehicle (speed of 30km/h). The image pixels within such range are assumed to be on a simulated ground plane, otherwise ignored (as the unreasonable area illustrated in Fig. 7). The following modules of our system such as the image stabilization and STHOG-based pedestrian detection will be performed on this ground plane.



**Fig. 7** Geometric ground plane constraint will reduce both the computational cost and false alarms on the background regions. Like the right image, the unreasonable sky area that may look like human shape will be ignored



As illustrated in Fig. 7, the advantages of applying this ground plane constraint include:

- Reducing the computational cost (about only 40 % of the image will be searched). Objects at the infinite depth distance (such as the vehicles or background regions too far away from the camera) or in the sky (roof, windows, etc.) will be ignored. Such assumption is reasonable, because it is almost impossible for a pedestrian to walk in the sky or on the roof of a house, and the pedestrian at the infinite distance to the vehicle is not a threat to the safe driving.
- Reducing the false alarm. The background objects like the telegraph pole or windows (right part of Fig. 7) will be ignored due to their geometric position is out of the ground plane.

### 3.3 Image stabilization through three successive frames

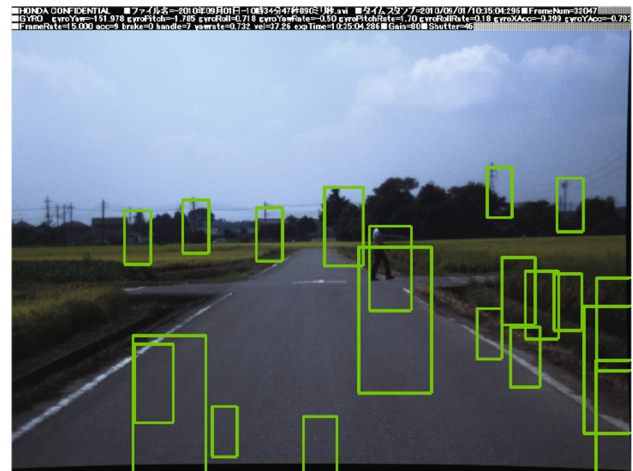
Since the onboard camera will keep on moving, not only the pedestrians but also the static background become moving due to the camera movement. In a video sequence, the motion includes: movement of walking pedestrians, scale changes of the static background objects and other moving background like vehicles and the complex motion of all the things in the image caused by the changes of camera view-points. Therefore, it becomes difficult to directly identify the human motion from the other cluttered motions as shown in Fig. 8.

To solve this problem, we set up an image stabilization method for suppressing the background motion among the successive frame and extracting the pedestrian motion smoothly.

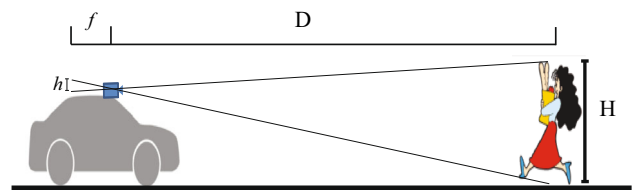
#### 3.3.1 Height estimation from monocular camera

As shown in Fig. 9, when the camera optical focus length ( $f$ ) is fixed, as for a pedestrian, his/her image height in the video sequence can be expressed as:

$$h = \frac{f * H_{ped}}{D}, \tag{10}$$



**Fig. 8** Failure sample of directly applying pedestrian detector to the onboard moving video sequence without the image stabilization. Many false alarms are produced at the unreasonable area such as the tower and trees at extreme long distance



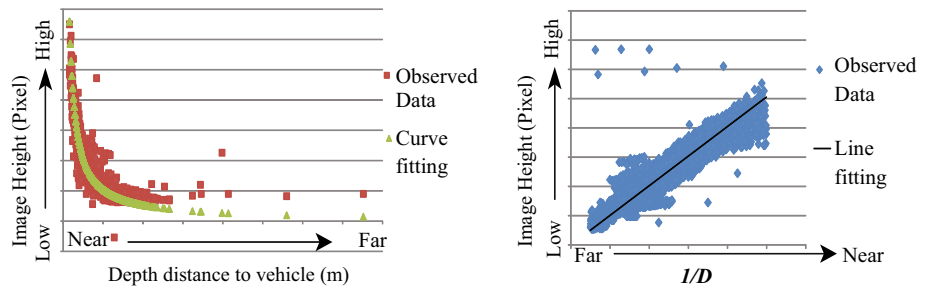
**Fig. 9** Illustration of the scene geometry

where  $H_{ped}$  is the pedestrian height in the real world,  $D$  is his/her depth distance to the vehicle. Since, as for the same person,  $f$  and  $H_{ped}$  will not change in one video sequence, his/her image height is in the inverse ratio to his depth distance to the vehicle. This equation will be correct only if the pitch rate equals to  $0^\circ$ . In the case that the camera pitch rate  $\theta \neq 0$ , the wrongly estimated pedestrian height  $h'$  would be  $h' = \frac{h}{\cos(\theta)}$ , where  $h$  is the true pedestrian height when  $\theta = 0$ .

In this way the estimation error rate  $Error(h) = \frac{h' - h}{h}$  could be translated into another format as  $Error(h) = \frac{1 - \cos(\theta)}{\cos(\theta)}$ . Therefore, the estimation error of pedestrian height will only



**Fig. 10** From our City-Scene dataset, the statistical analysis between the pedestrian height in the image and depth distance from this person to the vehicle. Left graph: curve fitting to the real data; Right graph: linear line fitting to real data



depend on the camera pitch error. Even when the pitch rate error is  $10^\circ$ , the height estimation error could still remain small as 1.54%. When the pitch rate error is as big as over  $20^\circ$ , the height estimation error will be quite problematic as being over 6.42%.

To investigate the relationship between the pedestrian image height and their depth distance to the vehicle, we prepared a City-Scene dataset which consists of more than 10,000 pedestrians. Through the statistical analysis on this dataset, we get the result of Fig. 10, where  $f * H_{ped}$  could be roughly considered as constant. As  $f$  is fixed throughout all the sequences, such a constant  $f * H_{ped}$  corresponds to the average height of all pedestrians in our dataset as 1.53 m (covering male, female, teenagers and adult). Similar result has also been reported in [1].

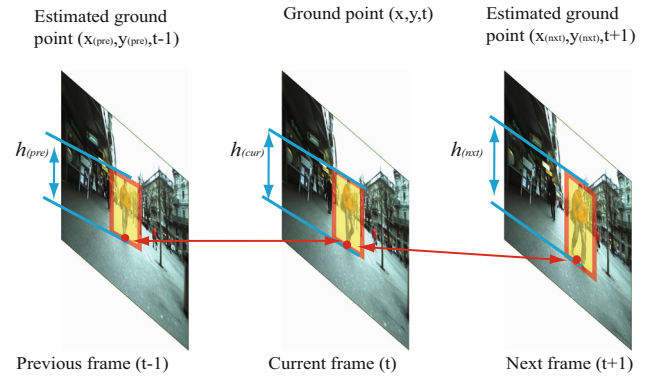
On the consideration of such constant  $f * H_{ped}$ , it will be easy for us to estimate the pedestrian height according to his/her depth distance from his/her foot point to the vehicle. In this way, even a pedestrian is far from the vehicle, we can still estimate his/her approximate height by Eq. 10. Such estimated height will be used for the following image normalization process.

### 3.3.2 Single-scale image normalization in three successive frames

Although Eq. (10) could estimate the pedestrian height in one single frame, since we use the spatio-temporal feature descriptor (details in Sect. 3.4.1) to represent the pedestrian feature, estimating the pedestrian height should be achieved in three successive frames simultaneously, but not frame by frame.

Given a ground point  $\bar{\mathbf{X}}_t = (x, y, t)$  in the current frame  $t$ , if a pedestrian is walking across this point, from Eq. (10) his height could be estimated as  $h_{(cur)}^{(t)} = \frac{const}{D_{(x,y)}^t}$ , where  $D_{(x,y)}^t$  is the 3D depth of this ground point.

The corresponding ground points to  $\bar{\mathbf{X}}_t$  in the previous and next frames could also be estimated from Eq. (9) and expressed as  $\bar{\mathbf{X}}_{t-1}$  and  $\bar{\mathbf{X}}_{t+1}$ . As shown in Fig. 11, the estimated height of a pedestrian around  $\bar{\mathbf{X}}_{t-1}$ ,  $\bar{\mathbf{X}}_{t+1}$  in the two adjacent frames could be calculated as:



**Fig. 11** Illustration of single-scale image normalization in three onboard successive frames. The foot points of one pedestrian in three successive frames will be estimated according to the SfM algorithm and the corresponding height will be computed by Eqs. 11, 12

$$h_{(pre)}^{(t-1)} = h_{cur}^{(t)} * \frac{D_{(x,y)}^t}{D_{(x,y)}^t + Vel^t * \Delta t} \tag{11}$$

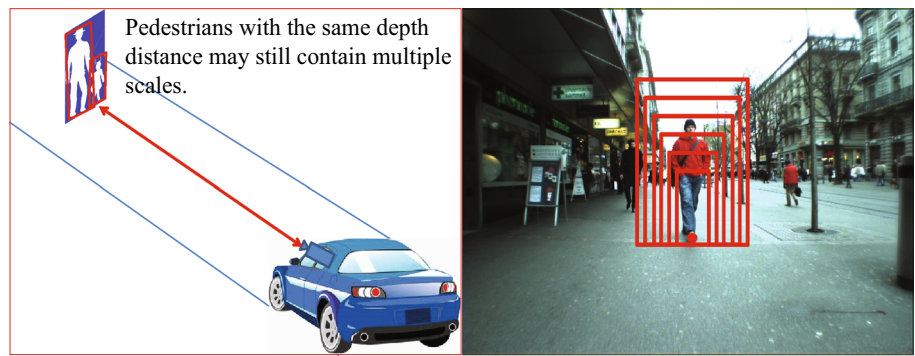
$$h_{(nxt)}^{(t+1)} = h_{cur}^{(t)} * \frac{D_{(x,y)}^t}{D_{(x,y)}^t - Vel^t * \Delta t}, \tag{12}$$

where  $Vel^t$  refers to the vehicle velocity in the current frame  $t$ ,  $\Delta t$  is the shutter speed of camera. According to the estimated height ( $h_{(pre)}^{(t-1)}$ ,  $h_{(cur)}^{(t)}$ ,  $h_{(nxt)}^{(t+1)}$ ) and  $\bar{\mathbf{X}}_{t-1}$ ,  $\bar{\mathbf{X}}_t$ ,  $\bar{\mathbf{X}}_{t+1}$  in three frames, we could select the normalization windows in each frame with a fixed aspect ratio (here, 1/1.4) around each ground point.

The advantages of this image normalization include:

- Attenuating the background movement. Since the three normalization windows are extracted from the same foot point in the successive frames and the window size also correspond to the same object in the real world, the camera motion in the normalized windows will be attenuated. From such windows, we could easily compute the pure foreground motion which will be further categorized by the STHOG pedestrian detector (as Sect. 3.4.1).
- Automatic adjustment according to the distance and velocity. According to Eq. (9), (12), since the normalization windows are determined according to the depth distance and vehicle velocity, our system could automatically stabilize the image windows through various speeds.

**Fig. 12** Illustration of the multiscale image normalization. In this way, even multiple pedestrians with various heights are at the same distance to a vehicle, the proposed algorithm can still deal with such condition



**(a)** Image normalization should be performed with multiple scales.

**(b)** Illustration of the multiple scale image normalization at one ground point.

### 3.3.3 Multiscale image normalization

As in Fig. 12, despite containing the same depth distance to the vehicle, the height of pedestrians may vary greatly due to the age (child or adult) or sex (male or female). The pedestrian height estimated by Eqs. 10 and 12 should also be flexible to follow such height variance. Therefore, the estimated heights ( $h_{(pre)/(cur)/(nxt)}$ ) is changed to  $\beta * h_{(pre)/(cur)/(nxt)}$  where  $\beta = 0.7-1.3$ . Therefore, Eq. 10 could be considered as a specific case as  $\beta = 1.0$ . As the average pedestrian height estimated in Eq. 10 is 1.53 m, such multiple scale image normalization could follow the variation of real pedestrian height from 1 to 2 m. Here, by changing  $\beta$ , this normalization could deal with wilder variation of people's height.

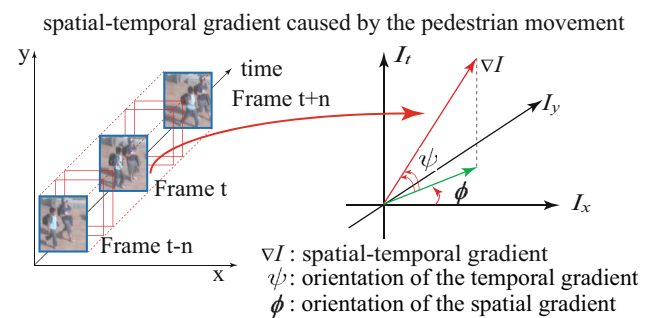
In this way, at a given ground point, the multiple normalization window will be set up according to the scale variation. The appearance and motion information computed from these multiscale windows will be used as the input for the following pedestrian detection process.

## 3.4 Object classification

### 3.4.1 Spatio-temporal HOG feature

#### Spatio-temporal gradient

Usually, a video can be considered as a 3D space consisting of  $x$ ,  $y$  and  $t$ , as illustrated in Fig. 13. When a pedestrian's walking is captured, his/her movement will produce a unique spatio-temporal gradient vector  $\nabla I$  whose orientation can be represented by the spatial ( $\phi$ ) and temporal ( $\psi$ ) orientation in the  $x$ ,  $y$  image plane and  $t$  direction. Since the pedestrian movement caused by the periodic movement of the arms (or legs) and the pedestrian shape are the unique features, it is reasonable to consider that only a pedestrian can display a human-like shape and motion. This is because the static background will only contain the spatial shape but not the temporal movement, and the moving background sel-



**Fig. 13** Relationship between spatial and temporal gradients in STHOG

dom simultaneously appears to be human with the periodic movement of people. Therefore, the spatial and temporal gradient orientations are considered to be helpful in detecting pedestrians in a video.

At frame  $t$ , as an image is represented as  $I(x, y, t)$ , the orientation of the spatial ( $\phi$ ) and temporal gradients ( $\psi$ ) at  $(x, y, t)$  will be computed as:

$$\nabla I = [I_x, I_y, I_t], \quad (13)$$

$$\psi = \tan^{-1} \left( I_t / \sqrt{I_x^2 + I_y^2} \right), \quad (14)$$

$$\phi = |\tan^{-1}(I_y/I_x)|, \quad (15)$$

where the gradients  $I_x$ ,  $I_y$ ,  $I_t$  are simply calculated as:

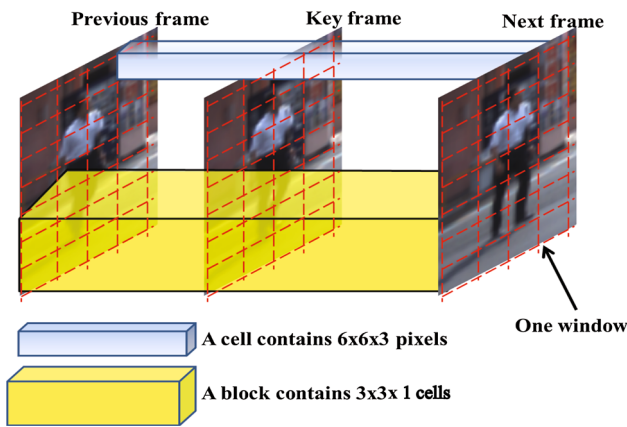
$$I_x = I(x+1, y, t) - I(x-1, y, t), \quad (16)$$

$$I_y = I(x, y+1, t) - I(x, y-1, t), \quad (17)$$

$$I_t = I(x, y, t+1) - I(x, y, t-1). \quad (18)$$

#### Construction of STHOG

The statistical histogram structure is used to represent the STHOG features, because a histogram can both describe the distribution of the STHOG features and suppress any abrupt random noise. The computed gradient orientation  $\phi$  and  $\psi$  are identified in the spatial and temporal histograms to describe



**Fig. 14** Construction of the STHOG features from three successive frames

both the pedestrian shape and motion. The STHOG feature is produced by concatenating the spatial and temporal histograms, which means the STHOG feature will be a single concatenated feature vector.

The implementation of STHOG is very similar to the well-known HOG (as shown in Fig. 14), where  $\phi$  ( $0-180^\circ$ ) and  $\psi$  ( $-90-90^\circ$ ) are computed from three successive frames and oriented in nine directions. For normalization, each block contains  $3 \times 3 \times 1$  cells, where each cell is composed of  $6 \times 6 \times 3$  pixels. The spatial and temporal histograms in each block are normalized with the  $L1$  norm. A STHOG feature is set up by concatenating the two histograms, where 9 spatial and 9 temporal bins are separately established by the STHOG feature.

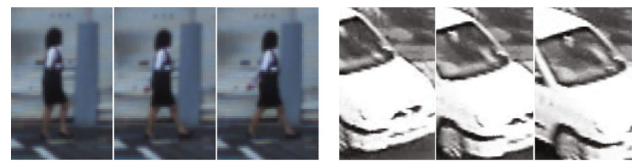
A detection window contains  $6 \times 9$  blocks, that is  $48 \times 66$  pixels by three frames, and the detection step is 6 pixels. We determined the size of a cell, a block and detection window experimentally so as to roughly fit to the minimum size of pedestrians in the test datasets that we used. 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.

Let  $\omega$  be the magnitude of the gradient at pixel  $(x, y)$ , with  $\phi$  the orientation of its spatial gradient. Then  $\phi_1$  and  $\phi_2$  represent the orientations of the corresponding two nearest neighbor bins to  $\phi$  in the histogram. The interpolation computation distributes the magnitude  $\omega$  into two parts as:

$$\omega_1 = \omega \frac{\|\phi_2 - \phi\|}{\|\phi_2 - \phi_1\|} \tag{19}$$

$$\omega_2 = \omega \frac{\|\phi_1 - \phi\|}{\|\phi_2 - \phi_1\|} \tag{20}$$

The interpolated magnitudes  $\omega_1$  and  $\omega_2$  are accumulated for all pixels within the cell to create the spatial histogram in



**Fig. 15** Illustration of the positive and negative training samples for STHOG features. Each set of training samples is composed of three successive image patches

**Fig. 15** Illustration of the positive and negative training samples for STHOG features. Each set of training samples is composed of three successive image patches

the STHOG feature. The temporal histogram can be set up by interpolation in the same manner.

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 in dealing with changes in viewpoints or walking directions.

### 3.4.2 STHOG pedestrian detection

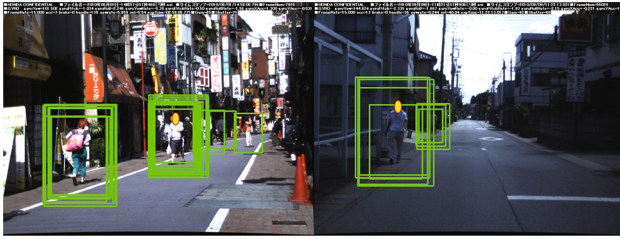
#### Training

Because of the lack of an explicit pedestrian model, machine learning is always required for finishing the pedestrian detection task, where the implicit representation can be learned from the extensive training samples. In this paper, the simple yet efficient AdaBoost algorithm described in [69] is selected to train our STHOG pedestrian detector.

Each set of positive/negative samples is composed of three successive pedestrian/background images, where both the shape and motion information are included (as in Fig. 15). The STHOG features are extracted from such successive frames. The STHOG pedestrian detector obtained from AdaBoost will be applied to the video sequence to locate the pedestrian.

During the AdaBoost training process, each bin of the STHOG is considered as a weak classifier. Because each detection window is composed of  $6 \times 9$  blocks, a block consists of  $3 \times 3$  cells and each cell has 18 bins, totally 8,748 weak classifiers are prepared in one detection window. A weak classifier is defined as:





**Fig. 16** Examples of initial detection results by the STHOG detector with the proposed detection system

$$f_i(b_i) = \begin{cases} 1 & \text{if } P_i(b_i - \eta_i) \geq 0, \\ -1 & \text{else} \end{cases} \quad (21)$$

where  $P_i$ ,  $\eta_i$  and  $b_i$  are the parity, bias and bin value of the  $i$ th weak classifier. A strong classifier is constructed by linearly combining  $N_{wc} = 600$  weak classifiers selected via AdaBoost:

$$F(\mathbf{b}) = \sum_{i=1}^{N_{wc}} \gamma_i f_i(b_i), \quad (22)$$

where  $\gamma_i$  is the weight for  $i$ th weak classifier, where the weak classifier is selected by the AdaBoost algorithm according to the value of  $P_i$ ,  $b_i$ ,  $\eta_i$ .

### Detection

After the AdaBoost training process, a STHOG pedestrian detector will scan the image from all the positions which satisfy the geometric constraint—their depth distance to the vehicle is in the range 10–100 m. To follow the height variation caused by the different pedestrians, as aforementioned in Sect. 3.3.3, at a given ground point, the images will be normalized as multiscale from three frames. The STHOG feature computed from unified multiscale images will be verified by a STHOG pedestrian detector. Running over those multiscale-normalized windows will guarantee all pedestrians with different heights to be checked by the STHOG detector.

The initial detection results like Fig. 16 will be further integrated by the following refinement process.

### 3.5 Refinement

To integrate the initial detection results, the nonmaximum suppression method is usually applied. As shown in Fig. 16, this process becomes a classic clustering initialization problem: we need to identify the number and position of clusters (here, corresponding to pedestrians.) from an unknown dataset. Compared with the fixed-bandwidth mean shift algorithm that suffers from the fixed kernel window, the adaptive mean shift clustering is a smarter method, where the bandwidth to each data is determined by the distance from it to its  $K$ -Nearest Neighbor ( $KNN$ ). Since the variation of cluster



**Fig. 17** Examples of the refinement results with the relative adaptive mean shift clustering

size is huge in the detection results of an onboard scene (like the big pedestrians who are close to the vehicle and small pedestrians being far from the camera), it becomes difficult to priory choose the proper value of  $K$ . Here, we choose a relative adaptive mean shift clustering method (similar to [40]), where the bandwidth is determined as  $\bar{s}/6$ , and  $\bar{s}$  is the mean value of the height of all initial detection results in the current image. In this way, the bandwidth of mean shift clustering in each frame will be adaptive. Figure 17 shows the refinement results with the proposed clustering method.

## 4 Experimental results

### 4.1 Benchmarks and datasets

The comparative experiments were taken on the dataset (described in Table 1) among HOG, HOGHOF, 3DHOG and STHOG pedestrian detectors which were the standard benchmarks for pedestrian detection. After the proposed geometric constraint process and multiscale image stabilization, the stabilized images were used as the input for each detector and the initial detection results of each detector will be integrated by the same refinement algorithm. Although we notice the other onboard pedestrian detection systems, we focused on the above four benchmarks in this paper due to the unavailability of open source codes for those systems.

As shown in Table 1, we prepared a City-Scene test dataset composed of 11 different test scenes with various background scenes, velocity and road conditions, and the dataset was accompanied with velocity information recorded for each frame. The resolution of test image is  $800 \times 600$  pixels. Totally, 2,390 frames and 11,272 ground truth bounding boxes are prepared in our test dataset. The pedestrian height varies from 23 to 518 pixels and the vehicle velocity changes between 1–41 km/h. The training dataset was selected from the other videos captured by the same vehicle. In our training dataset, every three successive samples (like forementioned in Sect. 3.4.2) are used as one positive/negative training set. To contain the necessary shape and motion information, the training samples were also selected from different viewpoints at various velocities. Totally, 12,024 positive and 53,130 negative samples are selected for training, and all the compared



**Table 1** Source of the test datasets

	Total frames	Total bounding boxes	Min height (pixel)	Max height (pixel)	Velocity (km/h)	Road condition
CityScene-01	149	241	33	365	35–38	Straight
CityScene-02	139	139	23	183	35–41	Straight
CityScene-03	163	163	23	231	7–41	Straight
CityScene-04	72	72	40	119	40–41	Curve
CityScene-05	98	201	40	296	37–38	Straight
CityScene-06	362	1,353	36	249	8–13	Straight/curve
CityScene-07	319	2,801	56	518	1–20	Straight/curve
CityScene-08	442	4,386	33	468	1–20	Straight/curve
CityScene-09	165	268	30	422	26–32	Straight
CityScene-10	160	384	40	404	23–27	Straight
CityScene-11	100	164	35	367	39–40	Straight

**Table 2** Pedestrian height distribution of the test datasets

	Total number of pedestrians	Pedestrians at short distance (1–30m)	Pedestrians at middle distance (30–70m)	Pedestrian at long distance (70m–)
CityScene-01	241	92 (38.2%)	142 (58.9%)	7 (2.9%)
CityScene-02	139	24 (17.2%)	70 (50.4%)	45 (32.4%)
CityScene-03	163	29 (17.8%)	57 (35.0%)	77 (47.2%)
CityScene-04	72	15 (20.8%)	57 (79.2%)	0 (0%)
CityScene-05	201	98 (48.8%)	103 (51.2%)	0 (0%)
CityScene-06	1,353	1,013 (74.9%)	339 (25.0%)	1 (0.1%)
CityScene-07	2,801	1,887 (67.4%)	914 (32.6%)	0 (0%)
CityScene-08	4,386	3,246 (74.0%)	1,139 (26.0%)	1 (0%)
CityScene-09	268	71 (26.5%)	168 (62.7%)	29 (10.8%)
CityScene-10	384	188 (49.0%)	196 (51.0%)	0 (0%)
CityScene-11	164	61 (37.2%)	100 (61.0%)	3 (1.8%)

pedestrian detectors are trained by the same samples with the AdaBoost algorithm. Although we notice other public dataset for onboard pedestrian detection such as ETH, TUD-Brussels and Caltech datasets, they were unavailable to test our algorithm due to the lack of velocity information, which is essential for our image stabilization technique with the SfM algorithm. Only the Daimler Stereo Dataset [70] is suitable for our system due to the vehicle velocity is included in this dataset, and we performed the comparative experiment with this public dataset.

In Table 1, the limitation of vehicle velocity in our dataset is set up as 40 km/h due to the consideration that our vehicle is assumed to mainly move in the urban residential street. According to the traffic laws all over the world (such as USA<sup>1</sup>, Japan<sup>2</sup> and other countries<sup>3</sup>), in the residential area, the

maximum velocity of vehicles is usually limited to 40 km/h to reduce the risk of traffic accidents according to the normal road condition. Therefore, on consideration of the global traffic laws, we limit the vehicle velocity in our test dataset as 40 km/h. However, we would like to argue that such limitation is not the performance limitation of proposed system, as shown in the latter Sect. 4.2.2 our system could detect a pedestrian at the distance of 100 m to our vehicle and such a distance could be enough to brake a vehicle moving at the speed of 100 km/h.

The detailed pedestrian height distribution of the test datasets is described in Table 2. Here, we define that if the distance from a pedestrian to the vehicle is 1–30 m, such a distance will be considered as short distance, when that distance is between 30 m and 70 m as middle distance, and if such a distance is over 70 m as long distance. Among all the test datasets, seven datasets mainly contain the pedestrians at middle distance (CityScene-01, 02, 04, 05, 09, 10 and 11), one dataset mainly contains the pedestrians at long dis-

<sup>1</sup> [http://en.wikipedia.org/wiki/Speed\\_limits\\_in\\_United\\_States](http://en.wikipedia.org/wiki/Speed_limits_in_United_States).

<sup>2</sup> <http://www.web-pbi.com/speed.htm>.

<sup>3</sup> [http://en.wikipedia.org/wiki/Category:Speed\\_limits\\_by\\_country](http://en.wikipedia.org/wiki/Category:Speed_limits_by_country).

tance (CityScene-03), and the rest 3 ones mainly focus on the pedestrians at short distance (CityScene-06, 07, 08).

#### 4.2 False positive per image (FPPI) evaluation

Although the receiver operator characteristic (ROC) is a popular evaluation measurement for the pedestrian detection task, the ROC can only describe the performance of a classifier but not a detector. As for an onboard pedestrian detection, since the performance evaluation is always performed across the full images but not within the detection windows, we selected the false positive per image (FPPI) evaluation measurement to describe the performance of each pedestrian detector.

During the FPPI evaluation, in the current frame  $t$ , when 50% size of one ground truth intersects with a predicted bounding box, such bounding box is counted as a true positive (TP), otherwise a false positive (FP). The detection rate (DR) and FPPI of a detector through one test sequence will be computed as

$$DR = \frac{\sum_{t=1}^n N_{TP}^t}{\sum_{t=1}^n N_{GT}^t}, \quad (23)$$

$$FPPI = \frac{\sum_{t=1}^n N_{FP}^t}{n}, \quad (24)$$

where  $N_{GT}^t$ ,  $N_{TP}^t$  and  $N_{FP}^t$  are the numbers of ground truth, true positive and false positive at  $t$ th frame and  $n$  is the total number of frames in a test sequence. Figure 18 and 19 show the FPPI evaluation of HOG, HOGHOF, 3DHOG and STHOG detectors on each test dataset with/without the proposed image stabilization.

##### 4.2.1 FPPI evaluation without image stabilization

In the case of onboard pedestrian detection, image stabilization is a key-important preprocess to improve the system performance. To evaluate the performance of proposed image stabilization method, we first evaluated the FPPI of four detectors without image stabilization.

The four kinds of detectors were directly applied to the image pyramid structure, where the scales of input images will vary from 0.35 to 1.0 and the scale factor is 0.05 (it means 14 layers in this structure). In such image pyramid, all detectors will scan over all test images at all positions and scales.

Figure 18 shows the detection performance of all detectors was quite poor without the image stabilization process. That is because we could not include all the possible variations (such as illumination, vehicle velocity and road condition) into the training dataset, those pedestrian detectors could not directly distinguish the shape/movement of a pedestrian from that of the moving background (which is related to the

vehicle velocity). Meanwhile, too many unnecessary background regions (that may contain human-like regions) were also scanned, which lead to too many false alarms in the unreasonable area (like the sky or the house roof). In this experiment, among the compared four kinds of detectors, STHOG and HOGHOF detectors are superior to the rest two detectors, and HOGHOF detector seems to get much better results. That is because the STHOG feature captures a sort of motion magnitude (temporal gradient) which will be greatly affected by the vehicle movements with different velocities. Since what the HOGHOF feature captures is not the motion magnitude but the motion orientation, it may be insensitive to the velocity changes to some extent. However, note that none of the compared four features could get good results without image stabilization, in other words, the image stabilization process is essential to get good results. Meanwhile, the ranking of each pedestrian detector is also unstable, which means that their detection results were more like to depend on the test scene and such results were unreliable.

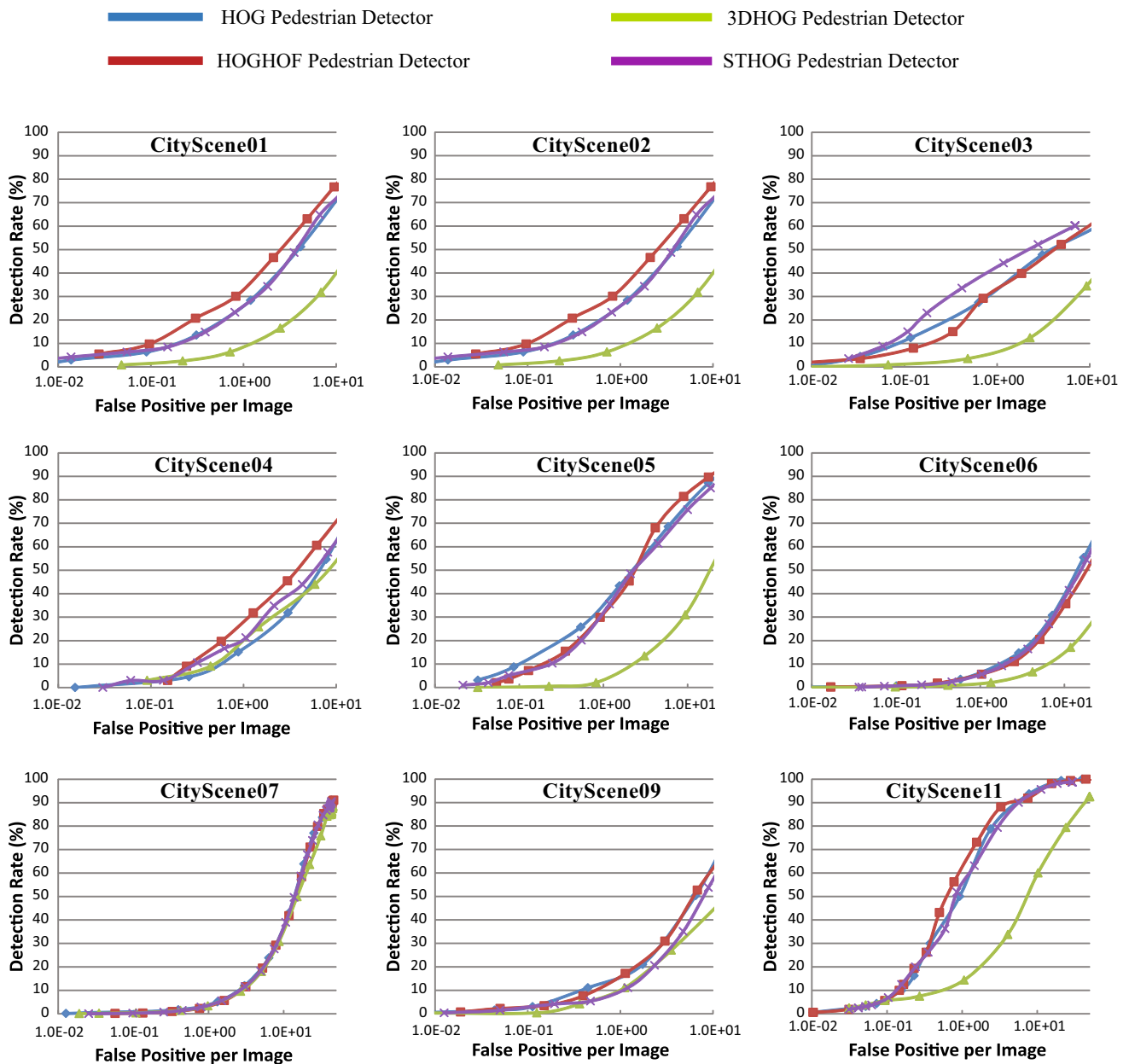
It has been reported in [71] that the STHOG feature outperforms the other features with static background, and such conclusion has also been proved in the latter Sec.4.2.2.

##### 4.2.2 FPPI evaluation with image stabilization

We did another experiment, where the detectors and test sequences were the same as described in Sect. 4.2.1. After the geometric ground constraint and multiscale image normalization, the four kinds of pedestrian detectors will be applied to the normalized image window, where at the same depth distance the window size will grow from 0.7 to 1.3 with the scale step of 0.1 (as described in Sect. 3.3.3).

Figure 19 shows the FPPI evaluation of all compared detectors with the proposed image stabilization process. Since the detection windows through three frames were normalized in multiple scales, both the scale changes and camera motion among frame were compressed. With such stabilization, the performance of all detectors was greatly improved. The STHOG pedestrian detector outperformed other detector through almost all sequences, the HOGHOF ranked the second, HOG ranked the third and 3DHOG detector lost almost all the sequences. Such stable ranking of all detectors across different sequences implies that the comparative experiments were quite fair and the ranking is independent of the test sequences.

The main reason for the missed detection of STHOG lies in the occlusion problem. As described in [1], in a test scene, most pedestrians (over 70%) are occluded in at least one frame. Since we only use the holistic pedestrian model for detection, when occlusion happens, the detection rate of all detectors is insufficient. Detailed discussion about the false positive and missed detection of STHOG could be found in the following Sect. 4.3. In CityScene-07 scene, except



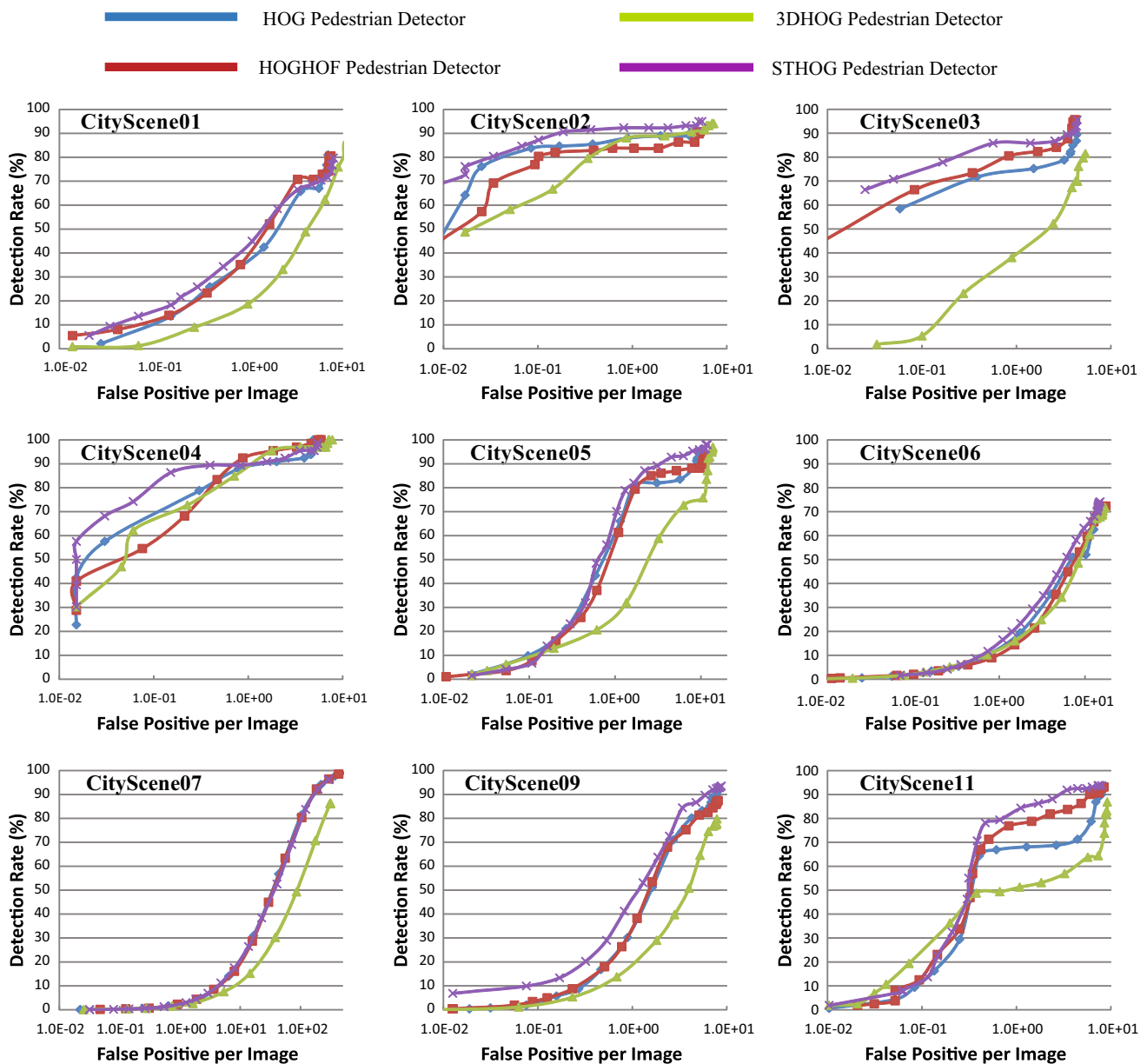
**Fig. 18** FPPI of the HOG, HOGHOF, 3DHOG and STHOG pedestrian detectors without our image stabilization method through the test dataset. Here, the performance of all detectors is too poor to be applied

into the real test scene. This result indicates that directly applying the pedestrian detectors to the test image is improper

the 3DHOG, the performance of other pedestrian detectors seems to be the same. That is because, besides the STHOG detector, the rest detectors tend to predict extensive large bounding box all over the search area. Since the FPPI evaluation will consider a prediction a true positive as long as its intersection with the ground truth is over the threshold, their detection rate will look like that of STHOG detector. However, as shown in the bottom row of Fig. 20, the performance of STHOG detector is more accurate than the other detectors.

Figure 20 shows the comparative experimental results among the four kinds of pedestrian detectors with the pro-

posed onboard system under different conditions. The top two rows (CityScene-01, 02 test sequences) show the examples of detecting the pedestrians at long distance where the pedestrian height was 65 and 28 pixels (in Row two, the distance is close to 100m). In such cases, the HOG and HOGHOF detectors became insensitive to the objects that were far from the vehicle, because the shape and optical flow motion of those objects became unclear in the images, and this would lead to the missed detection. The STHOG achieved the most robust and good performance under such condition because it directly combined the spatial and tem-



**Fig. 19** FPPI of the HOG, HOGHOF, 3DHOG and STHOG pedestrian detectors with our image stabilization method through the test dataset. Here, the STHOG detector won most sequences, HOGHOF, HOG and

3DHOG detectors rank 2, 3 and 4. The stable ranking of all detectors indicates that this evaluation is reasonable and independent on the test scene

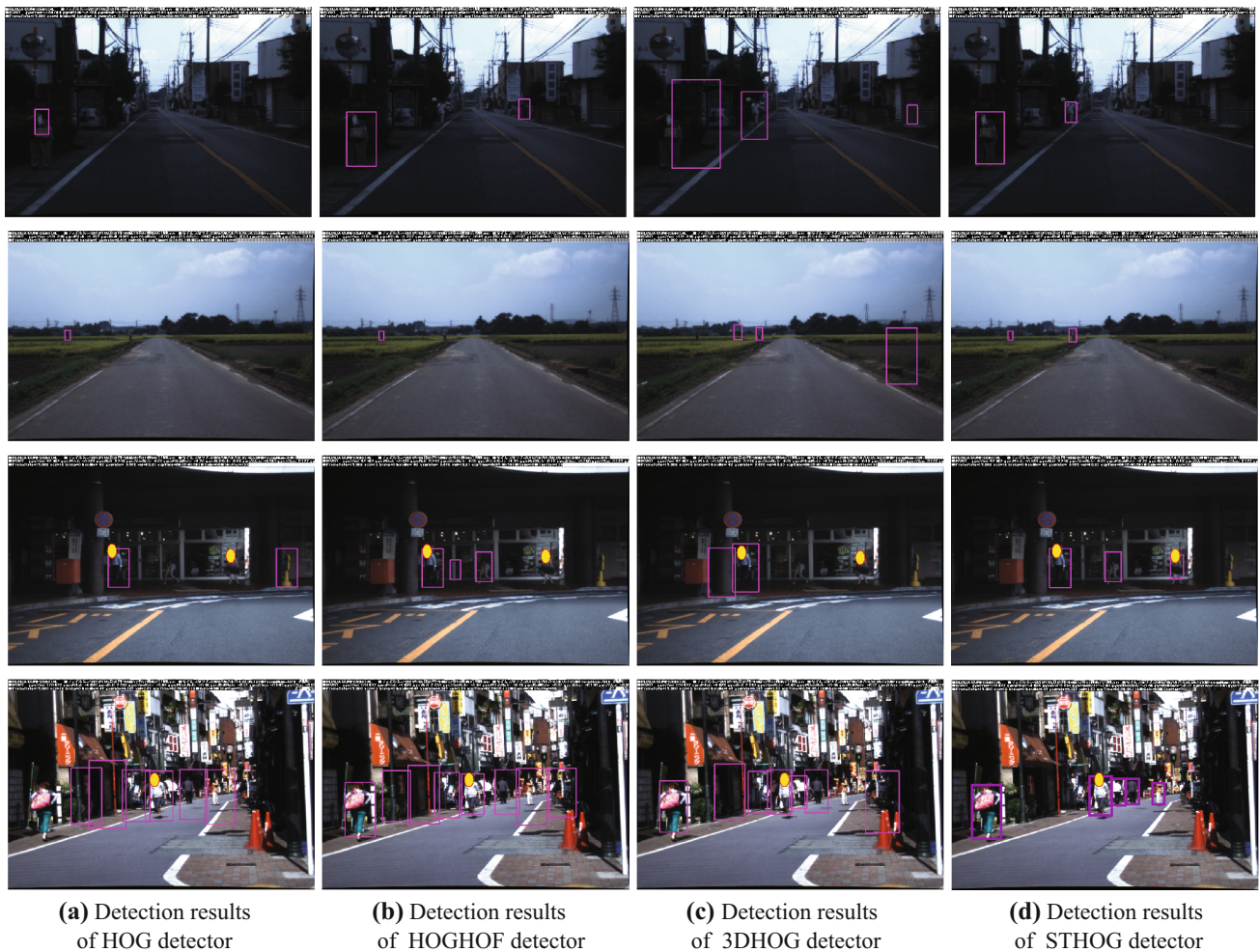
poral gradients to describe the pedestrian feature (shape and motion). The 3DHOG failed in these two sequences because the co-occurrence 3D gradient was susceptible to the unpredicted image noises. (such as the illumination changes or the unexpected vehicle movement).

The third row corresponds to CityScene-06 scenes, where the yaw rate of vehicle was high and there were a lot of human shadows in the glass. The HOG and 3DHOG detectors gave false alarms on the background objects looked like human (like the pillar or signpost) which we failed to include into the training dataset. The HOGHOF detector produced the false

alarm on background due to the same reason and missed a person where the reflection from glass behind the pedestrian made it difficult to compute the correct optical flow from the real pedestrian.

The bottom row of Fig. 20 corresponds to the complex CityScene-07, where many human-like background objects made the pedestrian detection task difficult. The HOG detector produced many false alarms on the background region and failed to locate some true pedestrians with new appearance such as carrying bag or umbrella. The HOGHOF and 3DHOG pedestrian detectors found almost all pedestrians at





**Fig. 20** Comparative experimental results of HOG, HOGHOF, 3DHOG and STHOG pedestrian detectors on the proposed onboard monocular detection system on different test sequences (corresponding to CityScene-01,02,06,07 of Table 1) under complex conditions. Column **a** HOG detector; **b** HOGHOF detector; **c** 3DHOG detector;

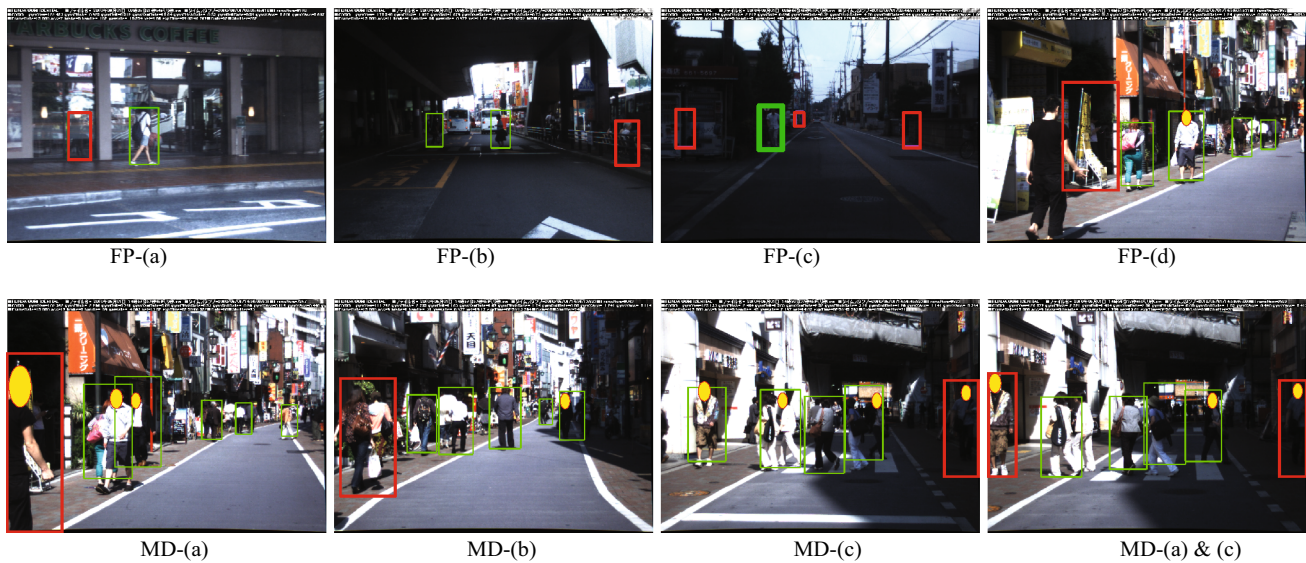
**d** STHOG detector. Here, STHOG detector won all sequences by producing the maximum true positive and minimum false alarms. The performance of HOGHOF and HOG looks similar and 3DHOG detector produced the maximum false alarms

the cost of producing many false alarms on the background regions which was caused by the changes of camera viewpoints. The STHOG detector was superior to the other detectors by producing the least false alarms while keeping high detection rate.

### 4.3 Analysis of false alarms and missed detection

Figure 21 shows some examples of the false positives (FP) or miss-detected (MD) pedestrians from the proposed STHOG detector under different scenes. In the top row, FP-(a) shows some examples that our system wrongly took the human shadow reflected in the glass as a pedestrian. That is because in the STHOG feature there is no difference between a real pedestrian and his reflected moving vertical shadow. In FP-(b), since a static seated person was not a pedestrian, our evaluation system considered such detection result as false

positive. In FP-(c), the camera motion parameters changed rapidly as: previous frame (yaw rate =  $-0.57$ ; pitch rate =  $-1.24$ ; roll rate =  $0.3$ ; current frame (yaw rate =  $0.32$ ; pitch rate =  $-4.7$ ; roll rate =  $0.44$ ). Therefore, the SfM became unstable as the pitch rate changed too rapidly in two frames, the image stabilization also worked poor as the motion of some background became strong and our STHOG detector wrongly took such area as the pedestrian because their shape also looked like person (containing many vertical gradients). As for FP-(d), there is an example that the background area was a combination of periodically waving flag and a supporter whose angle happened to look like people’s leg. In such case, such background happened to simultaneously satisfy our assumption in STHOG feature: the periodically waving flag produced some shape and motion features like the human torso and arms, the below supporter gave strong spatial gradient like our leg. Therefore, it is a relatively reason-



Examples of false positive and missed detection

**Fig. 21** Examples of false positive and missed detection. The *red rectangle* means the false positives or missed pedestrians. FP-(a): false positive on the human shadow reflected in the glass; FP-(b): static seated human is not marked as the ground truth; FP-(c): when the pitch rate of camera changed greatly, the SfM becomes unstable; FP-(d): flag

periodically waves like a person is walking. MD-(a): occluded pedestrian; MD-(b): multiple pedestrians in one detection window; MD-(c): extremely low contrast pedestrian. In FP-(d) a pedestrian was missing due to: (1) his depth distance to vehicle was shorter than 10m; (2) his foot was invisible and he could be considered as being occluded

able failure mode that the STHOG detector took such area as the human.

In the bottom row, MD-(a) shows the example that the pedestrian is out of the image, and such a pedestrian could be considered as being occluded because only part of his body is visible. Our detection system failed in such case because only the holistic detection model was applied. MD-(b) gives an example that multiple persons are located inside one detection window. Our STHOG detector failed to detect such persons because as for our detector it looks like one person with two heads, which has been beyond our assumption that one person only has one head. MD-(c) refers to the examples that the contrast of a person is extremely low, where the person in dark clothes was walking in the shadow. Since we use the VS camera, when the pedestrian is almost invisible, the STHOG detector failed to find such person.

Currently, we run the proposed detection system (basically composed of two systems: image stabilization and STHOG detection system) on a normal desktop PC whose CPU is an Intel C2D 3.16 GHZ and 4GB memory. It will take 10 s to obtain the multiscale detection results from an  $800 \times 600$  pixel-sized image by scanning over the calculated ground plane at all positions and scales. The image stabilization system will take approximately 8 s and the STHOG detection process takes the rest 2 s. The main reason for such a slow processing speed lies in: (1) the applied SfM is not real time. We just applied the normal SfM for testing the performance of proposed system; (2) only a single-layer AdaBoost was

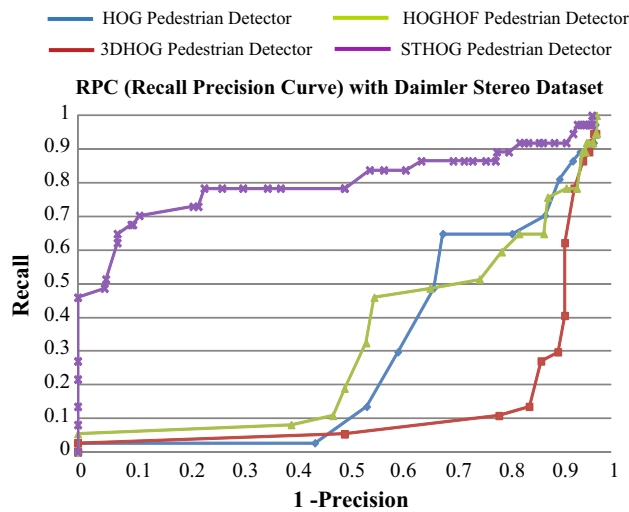
used; (3) only one thread of the CPU was applied while running our system, the other threads were wasted. Therefore, to tackle the aspect of computational cost of our system, more recent advanced online SfM algorithms (such as [7, 72]), cascade-structured AdaBoost algorithm for early rejection and the parallel processing software such as OpenMP should be applied.

#### 4.4 Comparative experiment with public dataset

Besides the dataset listed in Table 1, we also compared the performance of aforementioned four pedestrian detectors by the public Daimler Stereo Dataset [70], where useful vehicle parameters are included such as the velocity, pitch rate and steering. To compare the performance of proposed framework with the other systems conveniently, we choose the well-known Recall–Precision Curve (RPC) metric to describe their ability to detect pedestrians. In the RPC evaluation, each ground truth will only response to one prediction box. When the overlap rate between one ground truth and prediction box is over 50%, such prediction will be considered as a true positive, otherwise false alarm. When multiple prediction boxes intersect with one ground truth, only one prediction that satisfy the RPC metric will be considered as true positive, the others will be treated as false alarms.

Figure 22 shows the RPC of compared four kinds of pedestrian detectors by the Daimler Stereo Dataset. Here, the





**Fig. 22** The RPC evaluation of HOG, 3DHOG, HOGHOF and STHOG pedestrian detectors by the Daimler Stereo Dataset. In this experiment, the performance of STHOG detector is much superior to the other detectors. Corresponding experimental results could be found in Fig. 23

STHOG detector achieved the best performance and won the entire test. The HOGHOF detector got better performance than that of HOG detector. The performance of 3DHOG detector obtained the last ranking. The RPC ranking of this experiment is also consistent with that in Sect. 4.2.2, which implies that such ranking depends on the detection ability of compared algorithms but not the test scene.

The corresponding experimental results of Fig. 22 are shown in Fig. 23, where the STHOG detector achieved the best result by keeping the high recall rate with the least false alarm. The performance of HOG and HOGHOF detectors is quite similar and the 3DHOG detector produced the maximum false alarms.

To evaluate the performance of proposed system with the conventional baseline pedestrian detector, given the same Daimler Stereo dataset, we also compared the proposed STHOG onboard pedestrian detector with the well-known HOG LinearSVM (hereafter called as HOG LinSVM) [25] detector which is illustrated in [13]. The HOG LinSVM detector was implemented by the OpenCV and trained with the same training samples as the proposed STHOG onboard detector. Through the RPC evaluation of compared detectors in Fig. 24, it is clear that the proposed STHOG onboard detection system outperforms the HOG LinSVM detector greatly. That is because that the proposed ground constraint and image stabilization process could help the proposed system to reject the unreasonable background area and compress the motion of static background objects caused by the camera movement. In this way, the background movement is greatly reduced and only the moving object could produce the meaningful motion information, which guarantees the

STHOG detector to efficiently locate the pedestrians. While the HOG LinSVM detector directly scans the input images at all position and scales, which will lead to the massive false alarms on the moving background area due to the fact that we could never prepare a training set to cover all the possible background variation.

#### 4.5 Detection rate and pedestrian height

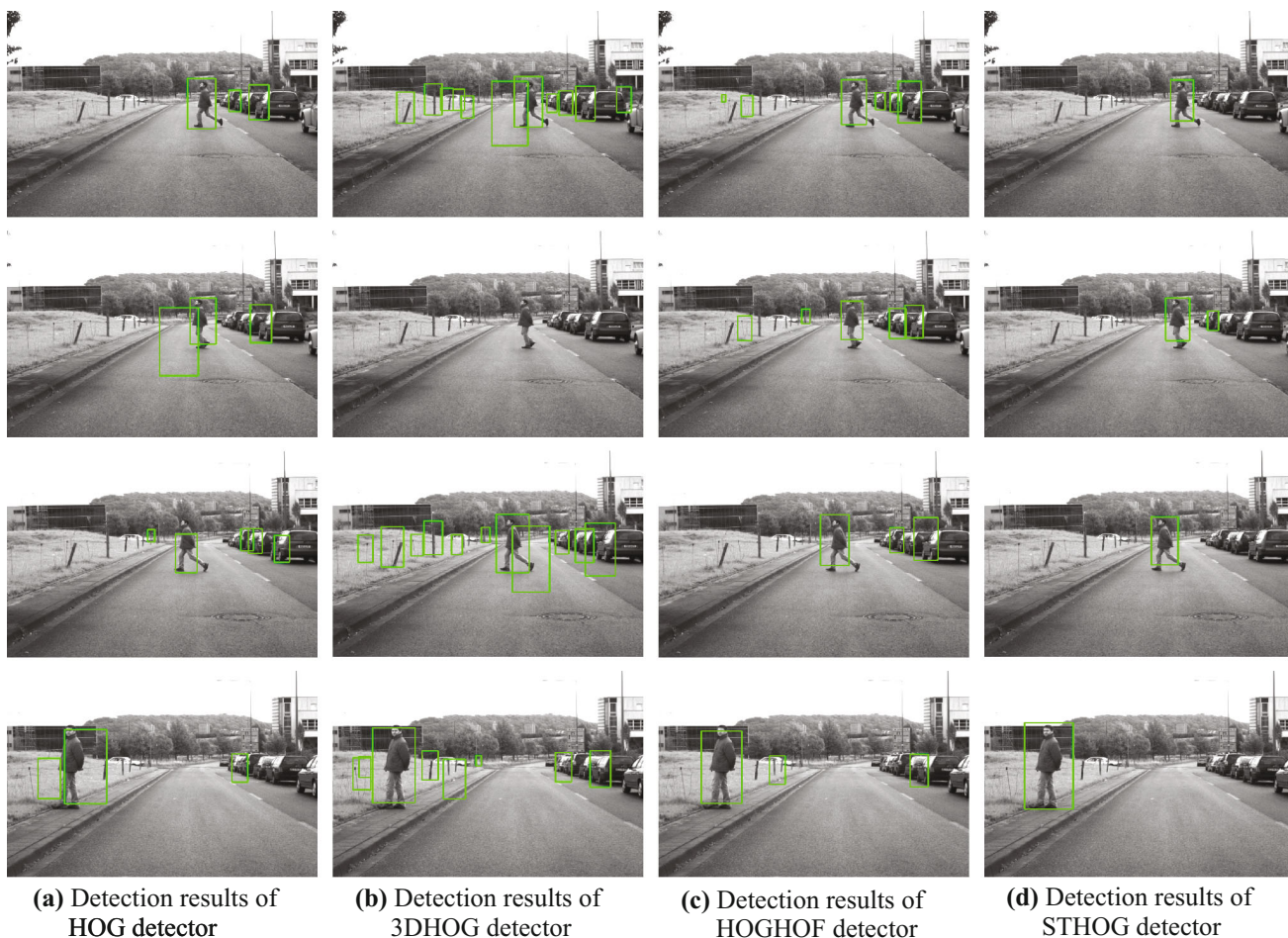
As well known, there is always a tradeoff between the detection rate and pedestrian height in a pedestrian algorithm. We investigate this tradeoff by analyzing the detection result of STHOG detector with the CityScene11 dataset. As shown in Fig. 25 (here the total recall and precision of test images are 90 and 92.6 %), the detection rate will increase as the pedestrian height becomes higher and higher. When the pedestrian height becomes higher than 46 pixels, the detection rate will be increased up to 100 %. Its detection rate slightly decreases to 90 % due to the rapid pitch changes of vehicle in this scene. This experiment proves that it is difficult for a pedestrian detector to find the people at very long distance. Similar conclusion has also been reported in [1].

However, compared with the results in [1], this experiment has also proved that the image stabilization can greatly improve the detection rate of whole system. Such stabilization will help the vehicle to detect the walking pedestrians at longer distance, which means saving more time to avoid the collision with pedestrians.

## 5 Conclusion and future works

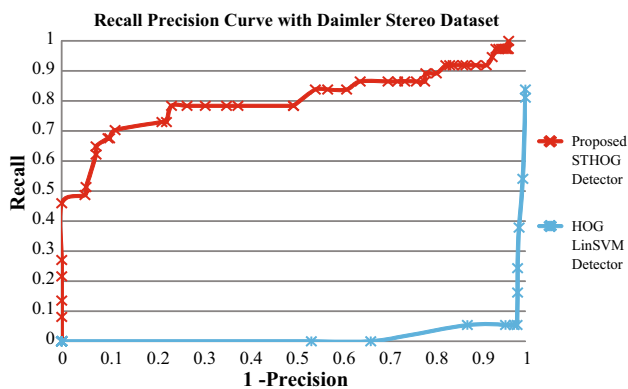
In this paper, we brought out a novel onboard pedestrian detection system which can work under the normal urban scene with a monocular camera. From the SfM, the geometric constraint process is applied for reducing not only the computational cost but also the false alarms in the unreasonable areas. To compress the motion of static background caused by the camera motion, a novel multiscale image stabilization method was also brought out according to homography transformation and camera geometry. The stabilized images are classified by the STHOG pedestrian detector which can separate a pedestrian from the background according to their shape and motion feature, simultaneously. Extensive experiments under various conditions proved the effectiveness of the proposed system.

The limitation of proposed system includes: (1) since the proposed pedestrian height estimation assumes the pitch rate of camera to be 0, when the pedestrian is at long distance to the vehicle (for example, 100 m), small pitch rate may still affect the estimation result, because at long distance the real pedestrian height in the image will be quite small and even small estimation error could still affect the accuracy of our

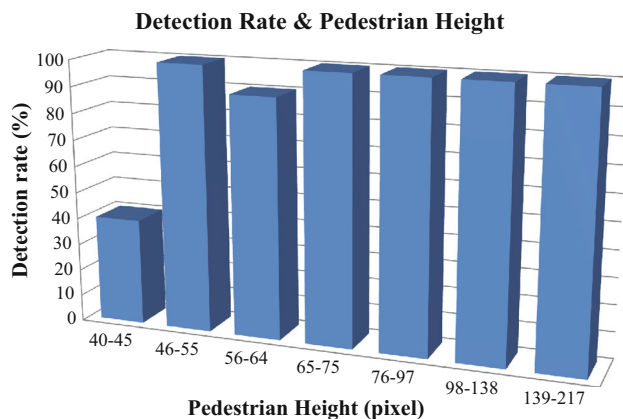


**Fig. 23** Experimental results of HOG, 3DHOG, HOGHOF and STHOG pedestrian detectors with the Daimler Stereo Dataset. Column **a**: results of HOG detector; Column **b**: results of 3DHOG detector; Column **c** results of HOGHOF detector; Column **d** results of STHOG

detector. Here, STHOG detector gets the best results with the least false alarms, the performance of HOGHOF and HOG is quite similar by producing almost the same true positive and false alarms. To keep high detection rate, 3DHOG detector produced the maximum false alarms



**Fig. 24** The RPC evaluation of proposed STHOG pedestrian detectors and HOG LinSVM detector [25] by the Daimler Stereo Dataset. The performance of STHOG detector is much superior to the HOG LinSVM method. It illustrates that without the proposed stabilization system, directly applying pedestrian detector to the mobile camera could not achieve desirable detecting performance



**Fig. 25** The investigation of tradeoff between detection rate and pedestrian height of STHOG detector with the CityScene11 dataset. This experiment has proved that the combination of image stabilization and detection algorithm can greatly improve the detection rate of whole system when the pedestrian height is low



system; (2) detecting the occluded pedestrian still remains as a challenging task in our system due to the fact that only the holistic detector is applied to detect the pedestrians; and (3) how to accurately identify a standing people at long distance is also one open issue to our system.

Therefore, the future works of our system should include:

- More accurate stabilization and geometric constraint. Although the SfM algorithm [65] applied in this paper requires high computational cost, it could be possible to realize a good tradeoff between the computational cost and false alarm by introducing more recent advanced online SfM algorithm such as [72]. Since this paper pays more attention on the aspect of precision and false alarm, we would like to tackle the aspect of computational cost in our future research. Meanwhile, other physical stabilization systems (e.g., gyroscope) that could reduce the camera pitch movement could also be helpful in improving the geometric constraint of proposed method.
- Occlusion problem. As occlusion always happens in the images captured from a vehicle, part-based or other flexible pedestrian models like [36] are good choice for solving this problem.
- Temporal tracking through the detection results. Based on the fact that the consistency of a pedestrian is stronger than that of an abrupt false positive, the after-process to track pedestrians through the detection results will greatly improve the detection rate and reduce the random false alarms. And this method has been proved to be useful and effective in [29,30]

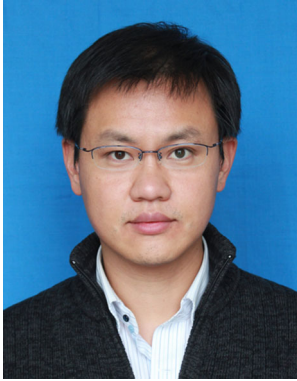
**Acknowledgments** This work was partly supported by the National Natural Science Foundation of China project 61433016, JSPS KEKENDHI Grant Number 21220003, “R&D Program for Implementation of Anti-Crime and Anti-Terrorism Technologies for a Safe and Secure Society”, Strategic Funds for the Promotion of Science and Technology of the Ministry of Education, Culture, Sports, Science and Technology, the Japanese Government, and the JST CREST “Behavior Understanding based on Intention-Gait Model” project.

## References

1. Dollar, P., Wojek, C., Schiele, B., Perona, P.: Pedestrian detection: an evaluation of the state of the art. In: *IEEE Trans. Patt. Anal. and Mach. Intell.* (2011)
2. Geronimo, D., Lopez, A.M., Sappa, A.D., Graf, T.: Survey of pedestrian detection for advanced driver assistance systems. *IEEE Trans. Patt. Anal. Mach. Intell.* **32**(7), 1239–1258 (2010)
3. Gandhi, T., Trivedi, M.M.: Pedestrian detection systems: issues, survey and challenges. *IEEE Trans. Intell. Transp. Syst.* **8**(3), 413–430 (2007)
4. Knoll, P., Hoefflinger, B.: HDR vision for driver assistance. In: *High-Dynamic-Range (HDR) Vision*. Springer, Berlin, pp. 123–136 (2007)
5. Dang, T., Hoffmann, C.: Stereo calibration in vehicles. In: *IEEE Transactions on Intelligent Transportation Systems*, pp. 268–273 (2004)
6. Broggi, A., Bertozzi, M., Fascioli, A.: Self-calibration of a stereo vision system for automotive applications. In: *IEEE International Conference on Robotics and Automation*, pp. 3698–3703 (2001)
7. Hoiem, D., Efros, A., Heber, M.: Putting objects in perspective. *IEEE Conf. Comp. Vis. Patt. Recognit.* **2**, 2137–2144 (2006)
8. Labayrade, R., Aubert, D., Tarel, J.: Real time obstacle detection in stereovision on non flat road geometry through ‘V-Disparity’ representation. *IEE Trans. Intell. Transp. Syst.* **2**, 17–21 (2002)
9. Sappa, A., Dornaika, F., Ponsa, D., Geronimo, D., Lopez, A.: An efficient approach to onboard stereo vision system pose estimation. *IEE Trans. Intle. Transp. Syst.* **9**(3), 476–490 (2008)
10. Ess, A., Leibe, B., VanGool, L.: Depth and appearance for mobile scene analysis. In: *IEEE International Conference on Computer Vision* (2007)
11. Ess, A., Leibe, B., Schindler, K., VanGool, L.: A mobile vision system for robust multi-person tracking. In: *IEEE Conference on Computer Vision and Pattern Recognition* (2008)
12. Ess, A., Leibe, B., Schindler, K., VanGool, L.: Robust multiperson tracking from a mobile platform. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, No. 10, pp. 1831–1846 (2009)
13. Keller, C.G., Dang, T., Feritz, H., Joos, A., Rabe, C., Gavrila, D.M.: Active pedestrian safety by automatic braking and evasive steering. *IEE Trans. Intell. Transp. Syst.* **12**(4), 1292–1304 (2011)
14. Agawal, S., Roth, D.: Learning a sparse representation for object detection. In: *European Conference on Computer Vision* (2002)
15. Munder, S., Gavrila, D.M.: An experimental study on pedestrian classification. *IEEE Trans. Patt. Anal. Mach. Intell.* **28**(11), 1863–1868 (2006)
16. Tuzel, O., Porikli, F., Meer, P.: Human detection via classification on Riemannian manifolds. In: *IEEE Conference on Computer Vision and Pattern Recognition* (2007)
17. Dollar, P., Wojek, C., Schiele, B., Perona, P.: Pedestrian Detection: A Benchmark. In: *IEEE Conference on Computer Vision and Pattern Recognition* (2009)
18. Enzweiler, M., Gavrila, D.M.: Monocular pedestrian detection survey and experiments. *IEEE Trans. Patt. Anal. Mach. Intell.* **31**(12), 2179–2195 (2009)
19. Dollar, P., Belongie, S., Perona, P.: The fastest pedestrian detection in the west. In: *Briti. Mach. Visi. Conf.* (2010)
20. Bourdev, L., Maji, S., Brox, T., Malik, J.: Detecting people using mutually consistent poselet activations. In: *Euro. Conf. Computer Vision* (2010)
21. Gavrial, D.M., Munder, S.: Multi-cue pedestrian detection and tracking from a moving vehicle. *Int. J. Comput. Vis.* **73**(1), 41–59 (2007)
22. Gavrial, D.M., Giebel, J., Munder, S.: Vision-based pedestrian detection: the PROCTOR system. In: *IEEE Intell. Veh. Sysmp.*, pp. 13–18 (2004)
23. Papageorgiou, C., Poggio, T.: A trainable system for object detection. *Int. J. Comput. Vis.* **38**(1), 15–33 (2000)
24. Viola, P., Jones, M.: Robust real-time object detection. *Int. J. Comput. Vis* (2001)
25. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: *IEEE Conference on Computer Vision and Pattern Recognition* (2005), pp. 886–893
26. Zhu, Q., Avidan, S., Yeh, M., Cheng, K.: Fast human detection using a cascade of histograms of oriented gradients. In: *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1491–1498 (2006)
27. Yamauchi, Y., Fujiyoshi, H.: People detection based on co-occurrence of appearance and spatiotemporal features. In: *International Conference on Pattern Recognition* (2008)

28. Wu, B., Nevatia, R.: Detection and tracking of multiple, partially occluded humans by bayesian combination of edgelet based part detectors. *Int. J. Comput. Vis.* **75**(2), 247–266 (2007)
29. Leibe, B., Cornelis, N., Gool, L.V.: Dynamic 3D scene analysis from a moving vehicle. In: *IEEE Conference on Computer Vision and Pattern Recognition* (2007)
30. Andriluka, M., Roth, S., Schiele, B.: People-tracking-by-detection and people-detection-by-tracking. In: *IEEE Conference on Computer Vision and Pattern Recognition* (2008)
31. Wang, H.Z., Han, X., Yan, S.C.: An HOG-LBP Human Detector with Partial Occlusion Handling. In: *IEEE International Conference on Computer Vision* (2009)
32. Park, D., Ramanan, D., Fowlkes, C.: Multiresolution models for object detection. In: *European Conference on Computer Vision* (2010)
33. Barinova, O., Lempitsky, V., Kohli, P.: On detection of multiple object instances using hough transforms. In: *IEEE Conference on Computer Vision and Pattern Recognition* (2009)
34. Roth, P.M., Sternig, S., Grabner, H., Bischof, H.: Classifier grids for robust adaptive object detection. In: *IEEE Conference on Computer Vision and Pattern Recognition* (2009)
35. Sternig, S., Roth, P.M., Bischof, H.: Learning of scene-specific object detectors for classifier co-grids. In: *Intern. Confer. on Advan. Vid. and Sigal-Based Survei.* (2010)
36. Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part-based models. In: *IEEE Trans. Patt. Anal. and Mach. Intell.*, vol. 32, No. 9, pp. 1627–1645 (2010)
37. Felzenszwalb, P.F., Girshick, R.B., McAllester, D.: Cascade object detection with deformable part models. In: *IEEE Conference on Computer Vision and Pattern Recognition* (2010)
38. Walk, S., Majer, N., Schindler, K., Schiele, B.: New features and insights for pedestrian detection. In: *IEEE Conference on Computer Vision and Pattern Recognition* (2010)
39. Marin, J., Vazquez, D., Geronimo, D., Lopez, A.M.: Learning appearance in virtual scenarios for pedestrian detection. In: *IEEE Conference on Computer Vision and Pattern Recognition* (2010)
40. Wang, M., Wang, X.G.: Automatic adaption of a generic pedestrian detector to a specific traffic scene. In: *IEEE Conference on Computer Vision and Pattern Recognition* (2011)
41. Rodríguez, M., Laptev, I., Sivic, J.: Jean-Yves Audibert: density-aware person detection and tracking in crowds. In: *IEEE International Conference on Computer Vision* (2011)
42. Breitenstein, M.D., Reichlin, F., Leibe, B., Koller-Meier, E., Van-Gool, L.: Online multi-person tracking-by-detection from a single, uncalibrated camera. *IEEE Trans. Patt. Anal. Mach. Intell.* **33**(9), 1820–1833 (2011)
43. Viola, P., Jones, M., Snow, D.: Detecting pedestrians using patterns of motion and appearance. *Int. J. Comput. Vis.* **63**(2), 13–161 (2005)
44. Laptev, I., Perez, P.: Retrieving actions in movie. In: *IEEE International Conference on Computer Vision* (2007)
45. Scovanner, P., Ali, S., Shah, M.: A 3-dimensional sift descriptor and its application to action recognition. In: *International Conference on Multime*, pp. 357–360 (2007)
46. Laptev, I., Marszalek, M., Schmid, C., Rozenfeld, B.: Learning realistic human actions from movies. In: *IEEE Conference on Computer Vision and Pattern Recognition* (2008)
47. Klaser, A., Marszalek, M., Schmid, C.: A spatio-temporal descriptor based on 3D-gradient. In: *Briti. Mach. Visi. Conf.*, pp. 995–1004 (2008)
48. Daniel, W., Mustafa, O., Pascal, F.: Making action recognition robust to occlusions and viewpoint changes. In: *European Conference on Computer Vision*, pp. 635–648, Berlin Heidelberg (2010)
49. Rodriguez, M., Sivic, J., Laptev, I., Audibert, J.Y.: Data-driven crowd analysis in videos. In: *IEEE International Conference on Computer Vision* (2011)
50. Sun, J., Wu, X., Yan, S.C., Cheong, L.F., Chua, T.S., Li, J.T.: Hierarchical spatio-temporal context modeling for action recognition. In: *IEEE Conference on Computer Vision and Pattern Recognition* (2009)
51. Wang, H., Ullah, M.M., Klaser, A., Laptev, I., Schmid, C.: Evaluation of local spatio-temporal features for action recognition. In: *British Mach. Visi. Conf.* (2009)
52. Murai, Y., Fujiyoshi, H., Kanade, T.: Combined Object Detection and Segmentation by Using Space–Time Patches. *Asian Conference on Computer Vision*, pp. 915–924 (2009)
53. Klaser, A.: Will Person Detection help Bag-of-features Action Recognition? Technical Report. INRIA, Spe (2010)
54. Dalal, N., Triggs, B., Schmid, C.: Human Detection using oriented histograms of flow and appearance. In: *European Conference on Computer Vision*, pp. 428–441 (2006)
55. Horn, B.K.P., Schunk, B.G.: Determining Optical Flow. *Artif. Intell.* **17**, 185–203 (1981)
56. Liu, Y., Shan, S., Chen, X., Heikkila, J., Gao, W., Pietikainen, M.: Spatial-temporal granularity-tunable gradients partition (STGGP) descriptor for human detection. In: *European Conference on Computer Vision* (2010)
57. Cheng, Y.Z.: Mean shift, mode seeking, and clustering. *IEEE Trans. Patt. Anal. Mach. Intell.* **17**(8), 790–799 (1995)
58. Goubet, E., Katz, J., Porikli, F.: Pedestrian tracking using thermal infrared imaging. In: *SPIE Conference Infrared Technology and Applications*, pp. 797–808 (2006)
59. Fardi, B., Schuener, U., Wanielik, G.: Shape and motion-based pedestrian detection in infrared images: a multi sensor approach. *IEEE Trans. on Intle. Transp. Sys.*, pp. 18–23 (2005)
60. Milch, S., Behrens, M.: Pedestrian detection with radar and computer vision. In: *Proceedings of the Conference on Progress in Automobile Light* (2001)
61. Bertozzi, M., Broggi, A., Felisa, M., Vezzoni, G., DellRose, M.: Low-level pedestrian detection by means of visible and far infrared tera-vision. In: *IEEE Trans. on Intle. Transp. Sys.*, pp. 231–236 (2006)
62. Marchal, P., Dehesa, M., Gavrilu, D., Meinecke, M.-M., Skellern, N., Viviguerra, R.: SAVE-U. Final Report, Technical Report, Inform. Sco. Techn. Program. of the EU (2005)
63. Harris, C., Stephens, M.J.: A combined corner and edge detector. *Alvey Visi. Confe.*, pp. 147–152 (1988)
64. Lucas, B., Kanade, T.: An iterative image registration technique with an application to stereo vision. In: *Seventh International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 674–679 (1981)
65. Hartley, R., Gupta, R., Chang, T.: Stereo from uncalibrated cameras. In: *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 761–764 (1992)
66. Fischler, M.A., Bolles, R.C.: Random sample consensus: a paradigm for mobile fitting with applications to image analysis and automated cartography. *Commun. ACM* **24**(6), pp. 381–395
67. Triggs, B., McLauchlan, P., Hartley, R., Fitzgibbon, A.: Bundle adjustment—a moden synthesis vision algorithm: theory & practice. In: Triggs, B., Zisserman, A., Szeliski, R. (eds.) Springer, LNCS, Berlin, 1883 (2000)
68. Richar, H., Zisserman, A.: *Multiple View Geometry in Computer Vision*. Cambridge University Press, Cambridge (2003)
69. Viola, P., Jones, M.: Robust real-time face detection. *Int. J. Comput. Vis.* **57**, 137–154 (2004)
70. Keller, C., Enzweiler, M., Gavrilu, D.M.: A New Benchmark for Stereo-based Pedestrian Detection. In: *Proc. of the IEEE Intell. Veh. Symp, Baden-Baden, Germany* (2011)

71. Hua, C.S., Makihara, Y., Yagi, Y.: Pedestrian detection by using a spatio-temporal histogram of oriented gradients. In: *IEICE Trans. on Inform. & Sys.*, vol. E96-D, No. 6, pp. 1376–1386 (2013)
72. Ryan, K., Balzano, L., Wright, S.J., Taylor, C.J.: Online algorithms for factorization-based structure from motion. [arXiv:1309.6964](https://arxiv.org/abs/1309.6964). <http://arxiv.org/abs/1309.6964>



**Chunsheng Hua** received his B.E. degree in electronic engineering from Shenyang University of Technology in 2001. He received his M.S. 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 State Key Laboratory of Robotics, 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.

versity. 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 IEICE, RSJ, and IEEE.



**Shun Iwasaki** is engaged in the development of safety technology in Honda R&D Co., Ltd., Automobile R&D Center, Tochigi, Japan. He received the B.S. and M.S. degrees from Shibaura Institute of Technology in 2006 and 2008. He developed a pedestrian recognition algorithm using a monocular camera.



**Yasushi Makihara** 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 Associate Professor of the Institute of Scientific and Industrial Research, Osaka University. His research interests are computer vision, pattern recognition, and image processing including gait recognition, pedestrian detection, morphing, and temporal super resolution. He is a member of IPSJ, RJS, and JSME.



**Keisuke Miyagawa** is Assistant Chief Engineer, Honda R&D Co., Ltd. Automobile R&D Center, Tochigi, Japan. He received the B.S. and M.S. degrees from Tokyo University of Agriculture and Technology in 1999 and 2001. He developed a faster recognition algorithm for pedestrian collision warning system using a monocular camera.



**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 Uni-



**Bo Li** did his Bachelor's degree between 2003 and 2007 in the Department of International Development Engineering in the Tokyo Institute of Technology. Between 2007 and 2009, he did his Master's degree in the Department of Information Processing in Tokyo Institute of Technology, and between 2009 and 2011, worked as an Engineer in Honda R&D Co., Ltd.