Pedestrian Detection by Combining the Spatio and Temporal Features

Chunsheng HUA†, Yasushi MAKIHARA†, and Yasushi YAGI†

† Department of Intelligent Medial, ISIR, Osaka University,8-1 Mihogaoka, Ibaraki, Osaka, 567-0047, JAPAN
E-mail: †huachunsheng@gmail.com, ††{makihara,yagi}@am.sanken.osaka-u.ac.jp

Abstract In this paper, we brought out a novel spatio-temporal HOG feature for the pedestrian detection under complex condition such as the cluttered background and various human appearances. Both the spatio and temporal image gradients are selected to set up the Spatio-Temporal HOG features (hereafter called as *STHOG*) for describing the human shape and movement character. Compared with the conventional HOG feature, the STHOG feature can effectively identify the human-like background regions because the movements of human and such regions are completely different in the video sequence. A STHOG-based pedestrian detector is learned from the training data by the Adaboost algorithm. Various experiments have confirmed the effectiveness and robustness of the proposed STHOG pedestrian detector.

Key words pedestrian detection, spatio-temporal feature, HOG

1. Introduction

Pedestrian detection has rapidly emerged as a challenging task in Computer Vision and attracted much attention from many researchers. It plays an important role in the surveillance system, smart vehicle and intelligent robot, etc. Although many powerful algorithms have been brought out, pedestrian detection still remains as a difficult task because of the variable human appearance caused by the articulated pose, clothing, illumination and cluttered background, etc. To make our work easy to be understood, in this paper, we refer the human/pedestrian to the moving people in the video sequence.

Although the computational cost of Sliding Window Detector (*SWD*) is always expensive, the *SWD* achieves impressive detection rate by scanning over the input image with various scales. The processing speed of *SWD* is significantly improved by either using the cascade-structure detector [1], [5] or restricting the search area with the camera geometry constraint [2], [6], [16] (such as, flat-world assumption, ground-plane-based object) and the prior knowledge about target (e.g. object height and aspect ratio).

The Haar wavelet-based cascade detector [1] is an efficient SWD method by introducing a cascade rejecter to select the most human-like region based on the Haar-like feature and spatio difference. The success of cascade structure is assuming that the overwhelming majority of detection window are the negative ones.

Histogram of Oriented Gradient (*HOG*) [3] uses the normalized histogram to describe the local shape and appearance of target object (similar to SIFT). Local gradients are binned according to their orientation, weighted by their magnitude within a spatial grid of cells. The most discriminative HOG features are selected by a linear SVM to achieve the pedestrian detection. Further improvement of combining HOG with optical flow has been reported in [4], where the optical flow is used for motion segmentation to filter out the background component far from the camera. Zhu *et al* [5] achieved almost real time pedestrian detection by training a cascade detector from the integral HOG feature while containing almost the same detection performance as [3].

Wu *et al* [6] proposed the Edgelet feature for pedestrian detection when partial occlusion happens. The affinity function of edgelet can be considered as a variation of Chamfer matching which can capture both the intensity and shape of the edge. Prior knowing the camera position and ground-plane, the partial occlusion problem is solved by maximizing the joint image likelihood with/without the occluded people.

Interest-Point based pedestrian detector [2], [16] can handle the partial occlusion by integrating multiple algorithms. Pedestrian hypothesis obtained from the interest-point detector (*ISM* detector) are further verified by the 3D information from stereo cameras, camera ego-motion flow and ground-plane geometry constrain.

To compress the affection of noisy background and partial occlusion, the HOG-LBP [15] detector is proposed by introducing the Local Binary Pattern (*LBP*) histogram into the conventional HOG. The histogram

produced by LBP is used to compress the random background noise and partial occlusion is detected by checking the inner product of a cell in the SVM classification. When partial occlusion happens, part-based HOG-LBP detector is applied to find the occluded person.

Good survey and other efforts for pedestrian detection have been proposed in [7]〜[10], [17]. As for the present algorithms, a common and stubborn problem is that there is no guarantee of the detection result under complicated condition where some background regions have the human-line shape/appearance or target contains strong texture like **Fig**.1. That is because they focus on detecting people frame by frame, the smoothness of human movement has been ignored, which can also be regarded as an important feature to tell human from human-like background.

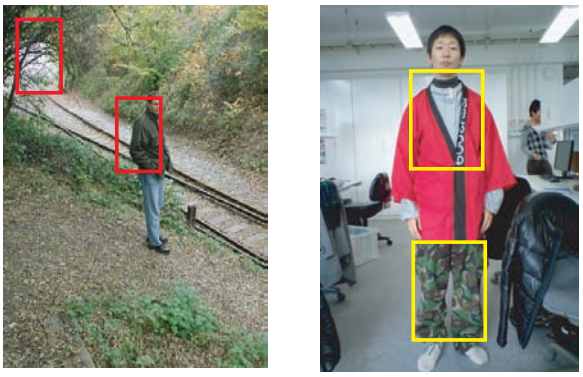


Fig. 1 Cluttered background or strong texture makes the detection difficult.

In this paper, a novel Spatio-Temporal HOG (called as the *STHOG*) feature is brought out for pedestrian detection even the background is cluttered. To efficiently discriminate the human from background, the human motion feature is embedded into HOG feature because the human movement is always so unique and different from the background component. The human movement is described by the histogram of spatio-temporal gradient from the adjacent frames. Since the movement of human head-shoulder, arm and legs is so different from that of the background component, the STHOG feature has the strong ability to distinguish the human from the noisy background. Various experiments under complex condition have confirmed the effectiveness of the proposed algorithm.

2. Spatio-Temporal HOG (*STHOG*)

2.1 Related works in action recognition

Since the last few years, the spatio-temporal feature has been proved to be effective in patch-based image segmentation [11], [12] and action recognition [13], [14], [19]〜[22]. Among those related works, the 3D HOG [21] and HOGHOF [19], [20] feature descriptors are close to the proposed STHOG method.

In this paper, the motivation to apply spatio-temporal feature (which can describe motion and shape) to pedestrian detection is that: although some background regions is cluttered or may have human-like shape, a background region will not contain the human-like shape and motion simultaneously. Therefore, the spatio-temporal feature which is produced by combining the motion and shape character is believed to have the strong power in locating pedestrian in the cluttered sequence.

2.1.1 3D HOG

Klaser *et al* [21] achieved the action recognition with a 3D HOG feature descriptor, which can describe both the shape and motion feature with a spatio vector. For a given 3D patch which is divided into $n_x \times n_y \times n_t$ cells, the gradients calculated from x, y and t are combined together to produce a spatio vector. The orientation of this spatio vector is quantized with an n-side polyhedron where n can be 4, 6, 8, 12 or 20. The corresponding 3D HOG descriptor concatenates gradient histograms of all cells and is normalized.

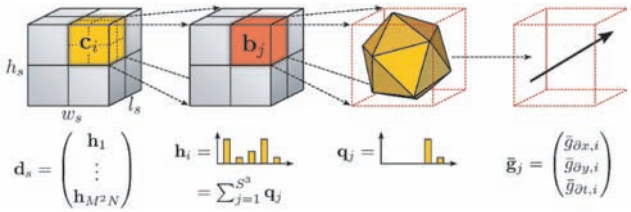


Fig. 2 Illustration of the 3D HOG descriptor from [21].

2.1.2 HOGHOF

Laptev used the shape and motion features for locating and recognizing action in [20] and clearly brought out the word HOGHOF (where HOF means the histogram of optical flow) in [19]. To characterize the local motion and appearance, histograms of the oriented gradient and optical flow accumulated in the space-time are concatenated to produce the HOGHOF descriptor. To improve the recognition rate, the Space-Time Interest Point (STIP) detector is applied to extract feature points and HOGHOF descriptor is used around those detected points.

2.1.3 Problems of 3D HOG and HOGHOF

Although the performance of 3D HOG is highly evaluated in action recognition, it is regarded as being unsuitable for pedestrian detection. That is because the spatio vector in 3D HOG is combined from the derivates of x, y and t. In the case of pedestrian detection, the

occluded background region in frame $t-1$ will appear in the next frame t, thus the derivate of t will be easily affected by such suddenly appeared region. Correspondently, the spatio vector will be changed, which may lead to the histogram shape changed. The successful action recognition of 3D HOG in the Hollywood Dataset lies in using the STIP detector to extract interest points. After stably extracting the corresponding feature points, 3D HOG descriptor is applied around the detected points.

HOGHOF can be independently applied for pedestrian detection and its performance is confirmed to be better than the normal HOG method. That is because the HOF component can distinguish the human movement from that of background. However, in Section 4., in the outdoor sequence, we will show that HOGHOF method is sensitive to the illumination and camera view point. In the fine day, the human shadow will move together with pedestrian and produce the similar optical flow to human, which makes HOGHOF confused. When the pedestrian walking direction is different from the training set due to different view point, the performance of HOGHOF will also be degraded.

2. 2 Our STHOG descriptor

To avoid the problem of 3D HOG and HOGHOF and contain their good property, a novel spatio-temporal HOG descriptor is brought out for pedestrian detection.

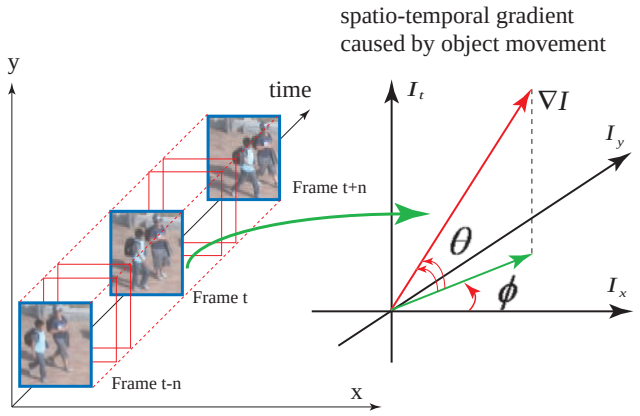


Fig. 3 Illustration of Spatio-Temporal gradient produced by movement.

As illustrated in **Fig**.3, as for image $I(x, y, t)$, the orientation of spatio-temporal gradient caused by the object movement can be calculated as:

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

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

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

where, I_x, I_y and I_t represents the gradient in x, y directions and time sequence, respectively. The spatio gradient caused by human movement is defined as θ, while ϕ represents the 2D image gradient as the normal HOG feature. Since the human movement is different from that of background, the θ from human motion will also be unique. Even if the abrupt background gradient affects θ, the shape feature ϕ can still work well.

To compress the affect of movement from background component that happens to move like human body (such as moving leaves), in this paper, the object shape and motion are described by the histograms of θ and ϕ, where both orientations are calculated in the similar to the well known HOG method. The normalized histogram of θ in a 5×5 cell from image $I(x, y, t)$ is computed from three continuous images at $t-1$, t and $t+1$, where the orientation of θ is also divided into 9 bins and the magnitude of bins is the sum of gradient strength. The histogram of ϕ is just calculated in the same way as HOG.

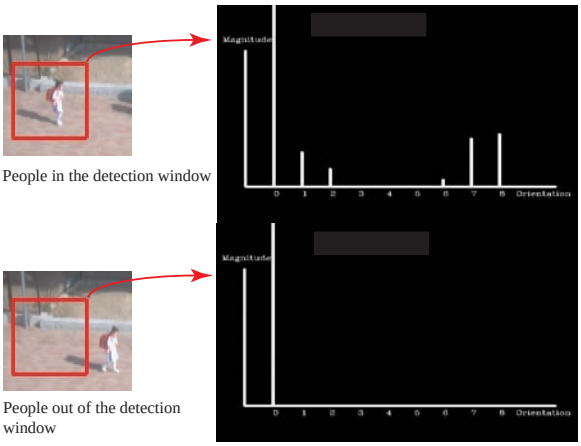


Fig. 4 Histogram of θ with/without people.

As shown in **Fig**.4, in a detection window, the histogram shape of spatio gradient ϕ with/without people will change greatly due to the special movement caused by human parts such as the arm, head-shoulder and legs. Such a phenomenon indicates that the histogram of spatio gradient can tell us if there is human-like movement or not in a detection window. **Fig**.4 shows that when there is no human in the detection window, the histogram of θ will keep bin 0 large value while other bins as zero, which indicates that there no moving objects in the sequence. The histogram changes of θ caused by other things such as the waving leaves or moving shadows will also be different from that caused by human parts.

The Spatio-Temporal HOG (*STHOG*) feature is produced by concatenating the histograms of normal image

gradient ϕ and spatio gradient θ together. Even if the problems that 3D HOG suffers from happens and the histogram of θ is affected, in STHOG, the histogram of ϕ can still work well. Since both the motion and shape features are integrated together, the STHOG feature naturally has the ability to tell human from those human-like background regions because of their different movement in the video sequence.

3. Pedestrian Detection with STHOG

Because of the lack of explicit models, finishing the pedestrian detection task always requires the use of machine learning techniques, where an implicit representation is learned from the mass training examples. Here, we select the Adaboost algorithm to train the STHOG-based pedestrian detector.

The training set for STHOG is composed of the temporal feature/image sets. One set of STHOG feature for training contains the STHOG histograms calculated from three continuous frames, and each set of the positive and negative training samples also share the same structure. The trained STHOG pedestrian detector will scan the image to locate human in the video sequence.

Since the motion and shape feature are integrated together in STHOG, we can get the fair good STHOG-based pedestrian detector from Adaboost with small training set. In our experiments, the detector is trained from 832 sets of positive samples and 927 sets of negative ones.

4. Experiment and Discussion

To evaluate the performance of proposed pedestrian detector, the comparative experiments were taken among the HOG detector, HOGHOF detector, Sequential HOG (SeHOG) detector and our STHOG method. In order to get the fair and accurate evaluation, all the detectors were trained by Adaboost with the same training datasets. Because the STHOG can only work with the video sequence, the test data included the PETS04, PETS09 dataset and three elemental school sequences taken by ourselves.

Since the STHOG method needs the spatio-temporal sequence for detection and training, while the HOG only needs static samples for training, the STHOG detector was trained with positive and negative sample sets (each set contained three continuous samples), and the HOG detector was trained with the middle sample extracted from each STHOG training set. The optical flow of HOGHOF was calculated by the pyramid Lucas-Kanade method from OpenCV. Sequential HOG (SeHOG) detector is constructed by combining three HOG histograms in parallel to set up a cubic HOG filter.

Here, the recall-precision curve (RPC) was selected to describe the accuracy of pedestrian detectors. The motivation and details of producing plots in RPC could be found in [18].

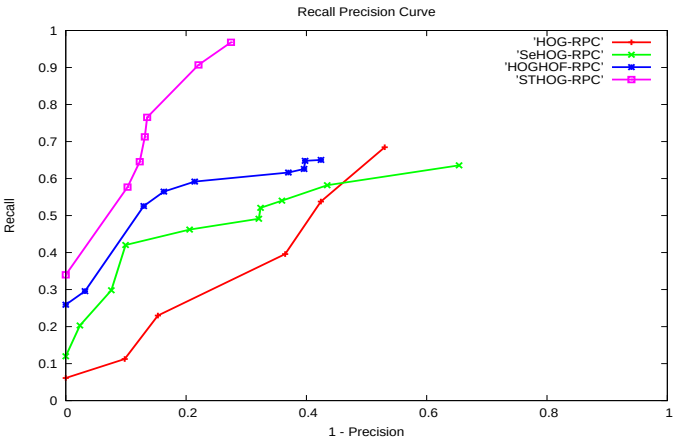


Fig. 5 RPC of STHOG, HOG and HOGHOF pedestrian detector in the sunshine day sequence.

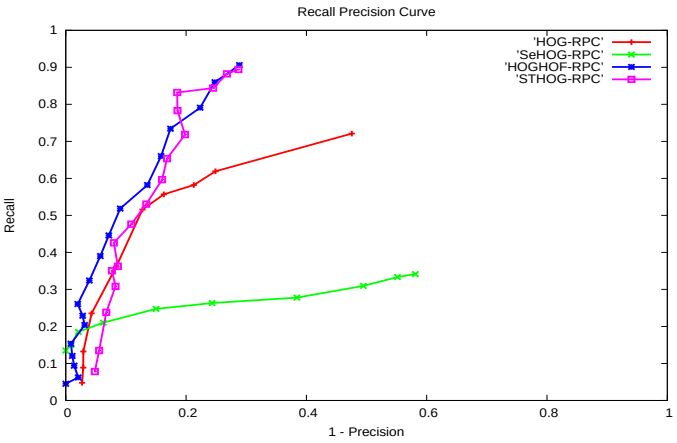


Fig. 6 RPC of STHOG, HOG and HOGHOF pedestrian detector in the rainy day sequence.

Fig.5 shows the RPC of each detectors in the fine day, where STHOG detector achieved the best detection rate while kept the false alarm (the "*1 - Precision*" in the RPC) low. HOGHOF detector was inferior to STHOG in this sequence because of the noisy optical flow caused by the moving human shadow. Since the shadow moved at the same speed and direction as human, sometimes HOGHOF detector was confused and wrongly took the shadow as human. SeHOG detector also worked better than HOG detector because it could filter the detection window three times.

Fig.7 shows the corresponding experimental result of the RPC in **Fig**.5. Clearly, the third row shows that the

HOGHOF pedestrian detector sometimes wrongly took the human shadow as the pedestrian. That is because such shadow contained not only the human-look shape but also the same optical flow as human body. While the STHOG detector became insensitive to the shadow because the spatio gradient of shadow was different from that of the training samples.

Fig.6 shows the RPC of tested detectors in the rainy day. In the test sequence, the illumination was dark and people were taking umbrellas which were not included in the training dataset. In this sequence, the HOGHOF pedestrian detector achieved the best performance because the human shadow did not appear and the optical flow caused by human movement was in the same direction as the training set. However, as the false alarm increases, the performance of STHOG detector becomes as good as that of HOGHOF detector.

Fig.8 shows the real comparative experimental results of the RPC in **Fig**.6. The performance of STHOG pedestrian detector is a little degraded because some human legs became almost invisible due to the illumination changes and the summer clothes, thus the spatio-temporal gradients in the leg region were not so clear. The SeHOG detector worked worst because the training dataset did not include the umbrellas and the cubic filter in SeHOG was too strict with the target shape. . The difference between the performance of STHOG and HOGHOF detectors is not so large and as the false alarm increases their performance is almost the same as each other.

Fig.9 shows the performance of STHOG pedestrian detector in the other sequences, e.g. unicycle, PETS04 and PETS09. Although, the tested pedestrians were not included in the training data, the STHOG detector could still work well because both the shape and motion features were considered.

Generally, it took our PC about 2 seconds to process a VGA test image, where the CPU is C2D 3.16GHZ and 4 GB memory. Since both motion and shape features are processed, we achieved the training with small training set, which contains 832 and 927 positive/negative sample sets.

5. Conclusion

In this paper, we brought out a novel algorithm for pedestrian detection. Both the motion and shape features are described in the novel STHOG feature, which makes it able to locate human in the cluttered background condition. Compared with the present shape-based pedestrian detector (HOG, SeHOG,etc), the STHOG detector becomes more insensitive to the human-like background region, variation of pedestrian appearance and clothes, etc. Compared with available motion-shape detector (like HOGHOF), our STHOG method is robust against the illumination and viewpoint changes. Since both the motion and shape feature are integrated together, the STHOG achieves getting good detector from the small training dataset by Adaboost algorithm.

In the future, we would like to focus on the multi-scale and pedestrian occlusion problems.

Acknowledgment

This work was supported by Grant-in-Aid for Scientific Research(S) 21220003.

References

[1] P.Viola, M.Jones and D.Snow, "Detecting Pedestrians Using Patterns of Motion and Appearance", *IJCV*, Vol.63, No.2, pp.13-161, 2005.

[2] B.Leibe, N.Cornelis and L.V.Gool, "Dynamic 3D Scene Analysi from a Moving Vehicle", *CVPR*, 2007.

[3] N.Dalal and B.Triggs, "Histograms of Oriented Gradients for Human Detection", *CVPR*, pp.886-893, 2005

[4] N.Dalal, B.Triggs and C.Schmid, "Human Detection Using Oriented Histograms of Flow and Appearance", *ECCV*, pp.428-441, 2006

[5] Q.Zhu, S.Avidan, M.Yeh and K.Cheng, "Fast Human Detection Using a Cascade of Histograms of Oriented Gradients", *CVPR*, pp.1491-1498, 2006

[6] B.Wu, R.Nevatia, "Detection and Tracking of Multiple, Partially Occluded Humans by Bayesian Combination of Edgelet Based Part Detectors", *IJCV*, Vol.75, No.2, pp.247-266, 2007

[7] D.M.Gavrial and S.Munder, "Multi-Cue Pedestrian Detection and Tracking from a Moving Vehicle", *IJCV*, Vol.73, No.1, pp.41-59, 2007

[8] T.Gandhi and M.M.Trivedi, "Pedestrian Detection Systems: Issues, Survey and Challenges", *IEEE Trans. Intel. Transportation Systems*, Vol.8, No.3, pp.413-430, 2007

[9] S.Munder and D.M.Gavrila, "An Experimental Study on Pedestrian Classification", *PAMI*, Vol.28, No.11, pp.1863-1868, 2006

[10] O.Tuzel, F.Porikli and P.Meer, "Human Detection via Classification on Riemannian Manifolds", *CVPR*, 2007

[11] Y.Yamauchi and H.Fujiyoshi, "People Detection Based On Co-occurrence of Appearance and Spatiotemporal Features", *ICPR*, 2008

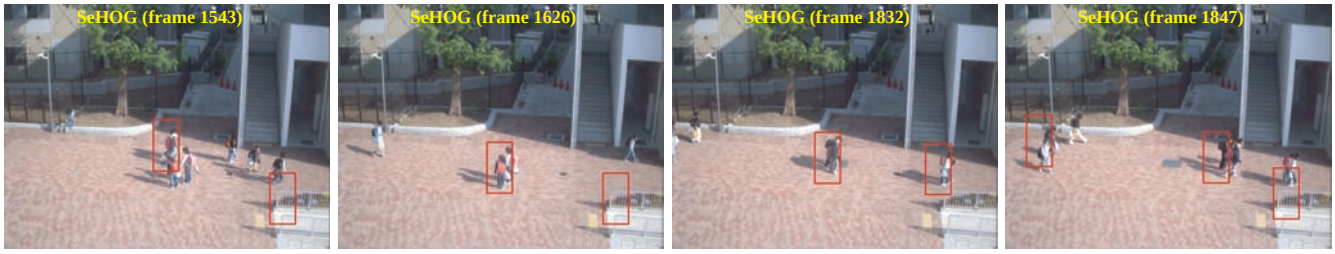
[12] Y.Murai, H.FUjiyoshi and T.Kanade, "Combined Object Detection and Segmentation by Using Space-Time Patches", *ACCV*, pp.915-924.2009

[13] J.Sun,X.Wu,S.C.Yan,L.F.Cheong,T.S.Chua and J.T.Li, "Hierarchical Spatio-Temporal Context Modeling for Action Recognition", *CVPR*,2009

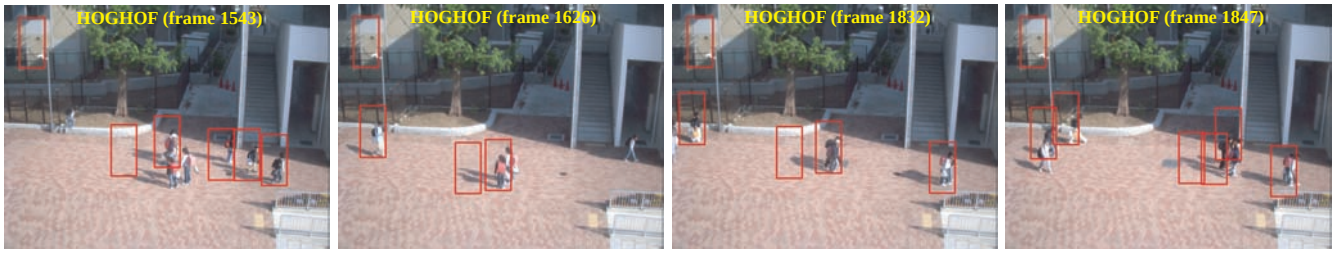
[14] H.Wang, M.M.Ullah, A.Klaser, I.Laptev and C.Schmid, "Evaluation of local spatio-temporal features for action recognition", *BMVC*, 2009

[15] H.Z.Wang, X.Han and S.C.Yan, "An HOG-LBP Human Detector with Partial Occlusion Handling",

Detection result of HOG pedestrian detector.



Detection result of SeHOG pedestrian detector.



Detection result of HOGHOF pedestrian detector.

Detection result of STHOG pedestrian detector

Fig. 7 Comparative experimental results among HOG, SeHOG, HOGHOF and STHOG detector in the sunshine day sequence. Here, HOGHOF pedestrian detector sometimes suffered from the moving human shadow that contained almost the same optical flow as human body and human-like shape. HOG and SeHOG detectors sometimes took the human-like static background as pedestrian. STHOG pedestrian detector worked best by checking both motion and shape feature to locate pedestrian through the video.

ICCV, 2009

[16] M.Andriluka, S.Roth and B.Schiele, "People-Tracking-by-Detection and People-Detection-by-Tracking", *CVPR*,2008

[17] M.Enzweiler, D.M.Gavrila, "Monocular Pedestrian Detection Survey and Experiments", *PAMI*, Vol.31, No.12, pp.2179-2195, 2009

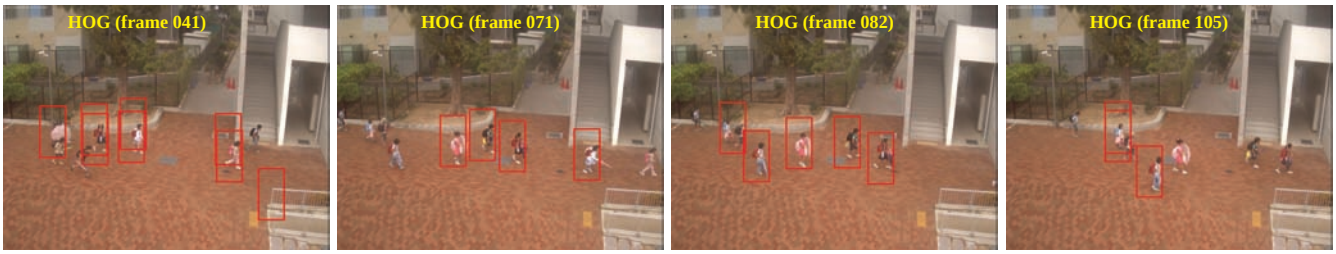
[18] S.Agawal and D.Roth "Learning a Sparse Representation for Object Detection", *ECCV*, 2002

[19] Ivan Laptev, M.Marszalek, C.Schmid, B.Rozenfeld, "Learning realistic human actions from movies", *CVPR*, 2008

[20] I.Laptev, P.Perez, "Retrieving actions in movie", *ICCV*, 2007

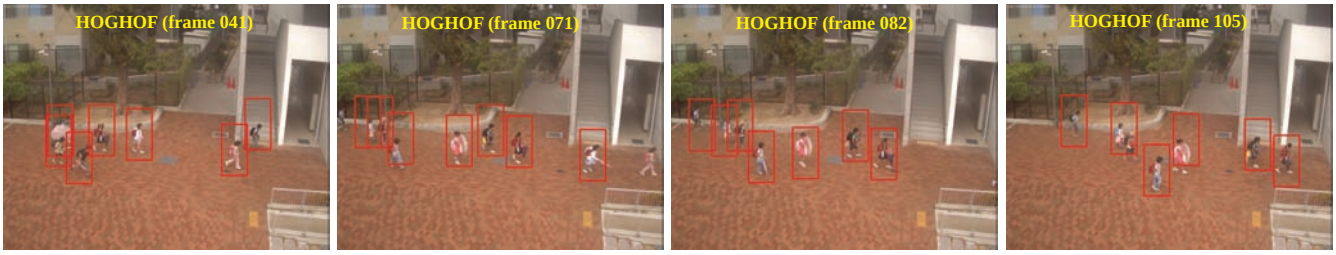
[21] A.Klaser, M.Marszalek, C.Schmid, "A Spatio-Temporal Descriptor Based on 3D-Gradient", *BMVC*, pp.995-1004, 2008

[22] H.Wang, M.M.Ullah, A.Klaser, I.Laptev, C.Schmid, "Evaluation of local spatio-temporal features for action recognition", *BMVC*, 2009

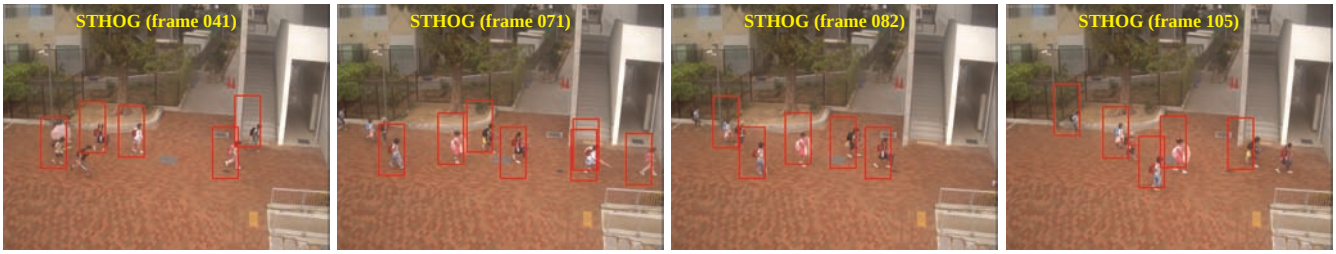


Detection result of HOG pedestrian detector.

Detection result of SeHOG pedestrian detector.



Detection result of HOGHOF pedestrian detector.



Detection result of STHOG pedestrian detector

Fig. 8 Comparative experimental results among HOG, SeHOG, HOGHOF and STHOG detector in the rainy day sequence. As the illumination changes, some people's legs become difficult to be observed and STHOG detector suffers a little from this problem. HOGHOF detector works best here, because there is no human shadow and people's motion is similar to the training dataset.

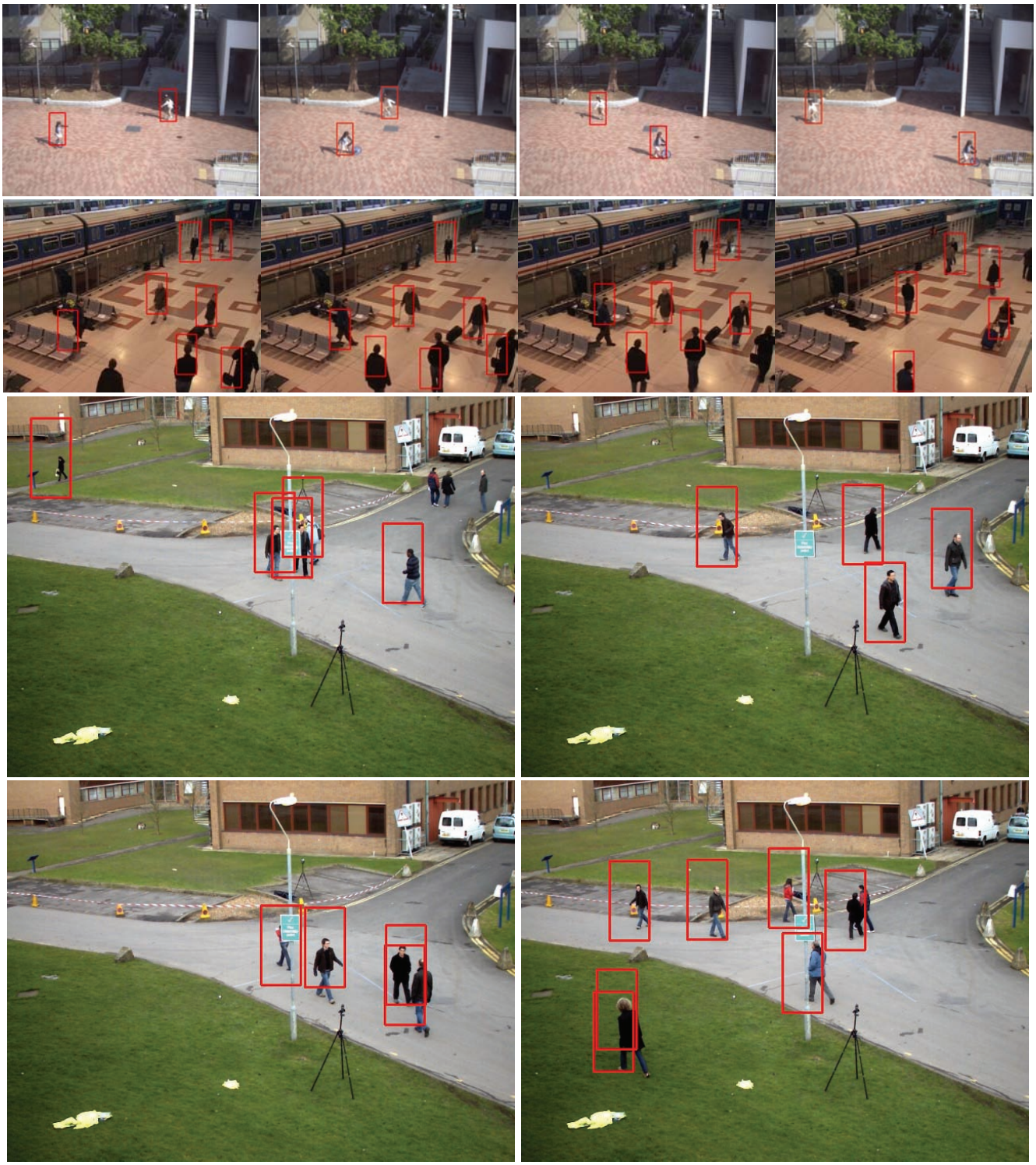


Fig. 9 Experimental results of STHOG detector in various sequences.