Stereo-based Pedestrian Detection using Multiple Patterns

Hiroshi Hattori
kan.hattori@toshiba.co.jp
Akihito Seki
akihito.seki@toshiba.co.jp
Manabu Nishiyama
manabu.nishiyama@toshiba.co.jp
Tomoki Watanabe
tomoki8.watanabe@toshiba.co.jp

Research & Development Center,
TOSHIBA Corporation, JAPAN

Abstract

Detecting pedestrians from a moving vehicle is a challenging problem since the essence of the task is to search non-rigid moving objects with various appearances in a dynamic and outdoor environment. In order to alleviate these difficulties, we propose a new human detection framework which makes the most use of stereo vision. While the conventional stereo-based detection methods initially generate regions of interest or ROIs on one of stereo images, the proposed one defines the ROIs on both left and right images. This paper presents two different ways for utilizing the stereo ROIs. The first one is to classify the stereo ROIs individually and integrate the classification scores to obtain the final decision. The second one is to extend gradient-based local descriptors [1, 14] to multiple views and present new feature descriptors which we call *Stereo HOG* and *Stereo CoHOG*. Through experiments we show that both methods significantly reduce the false alarm rate while keeping the detection rate comparing with monocular-based methods.

1 Introduction

Detecting people in images has been a central issue in computer vision and long been investigated[1, 5, 10, 11, 13]. In the past few years, the task has attracted even more research attention and considerable advances have been made [2, 3, 7, 8]. Among many applications, automotive ones[4, 12] are particularly promising as they have the potential to dramatically improve traffic safety. This paper deals with pedestrian detection from a moving vehicle.

The conventional methods for pedestrian detection can roughly be classified into two categories; monocular-based and stereo-based. In the monocular-based ones, as the first stage towards finding people, certain texture-based [12] or motion-based [9] mechanism generates candidate regions of interest or ROIs. While this step makes a significant contribution to overall performance, it is rather difficult to create candidate rectangles whose scales and positions are reasonable with a single image alone. In the stereo-based methods, on the other hand, binocular disparities or depths are available to obtain possible human regions which

BMVC 2009 doi:10.5244/C.23.94

contain any vertical surfaces. The extracted regions are then fed to a pattern classifier which decides if the candidate regions contain a pedestrian or not.

Normally, ROIs are defined in a single view, not only in the monocular-based approach but also in the stereo-based one. It means that just a single image is used for the classification procedure. This paper describes a new human detection framework where ROIs are defined on both left and right images and we utilize them together for the pattern classification. There are two different methods to use a pair of ROIs on stereo images. The first one is to classify a pair of local image patterns individually and combine those outputs to obtain the final decision. The second one is to extend two gradient-based local descriptors, HOG [1] and CoHOG[14], to deal with stereo views and to present new stereo feature descriptors which we call *Stereo HOG* and *Stereo CoHOG*. Through experiments we show that both methods significantly reduce the false alarm rate while maintaining the detection rate comparing monocular-based detectors.

2 Overview of our stereo-based detection system

In this section, we introduce our stereo-based human detection system from a moving vehicle. Figure 1 provides an overview of our system whose detection examples in urban environments are shown in Figure 2. Our stereo-based detection system consists of three components as follows.

Stereo Disparity Estimation: We use an area-based stereo method suitable for road applications [6]. We adopt Sum of Absolute Differences(SAD) as matching criteria and a recursive correlation technique improves computational efficiency. And we use a geometric restriction for further reduction of the computational cost. In road scenes, humans are on a road surface and they have a certain maximum height. Therefore, we can define upper and lower boundary planes where SAD measures are estimated just for the limited space. Also, consistency check between left to right and right to left correspondence is performed in order to filter out mismatches. The combination of these processes produces a high quality dense disparity map at a reasonable computational cost. Figure 1(b) shows an example of a disparity map. More red intensity indicates larger disparities which mean closer areas while black indicates unknown disparities.

Stereo-based ROI Generation: The next step is a disparity-based ROI generation. Basically, uniform disparity regions are extracted as shown in Figure 1(c) illustrated with green rectangles. The dense disparity estimation makes this process straightforward and it determines proper scales of rectangles wherever the candidate objects appear. Also, road areas are easily detected as zero-height regions. The ability is significant as road regions occupy the

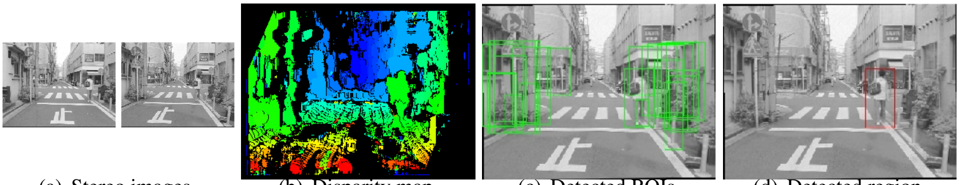

(a) Stereo images. (b) Disparity map. (c) Detected ROIs. (d) Detected region.

Figure 1: Detection Overview.

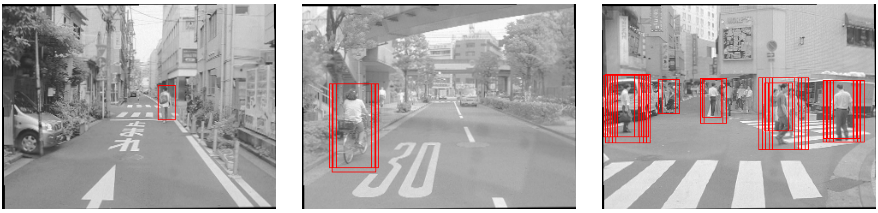


Figure 2: Detection examples. No post-processing is applied to merge multiple detections.

large part of an input image in most cases and those regions can be excluded in the following process. These tasks are rather difficult if we use a monocular camera.

Pattern Classification : The final step is to classify intensity patterns within the candidate regions. Figure 1(d) shows a detection example indicated with a red rectangle. A linear support vector machine is applied as a baseline classifier for simplicity and speed throughout this study. We basically adopt gradient-based local descriptors which proved to be appropriate for human detection. The next section provides an outline of our feature descriptor.

3 Co-occurence Histograms of Oriented Gradients

Histogram of Oriented Gradient(HOG) [1] is frequently selected as a feature descriptor for human detection. The outline procedure is as follows. The first step is to compute gradient orientations at every pixel $\mathbf{x} = (x, y)$. A gradient orientation θ is defined as $\theta(\mathbf{x}) = \tan^{-1} v(\mathbf{x})/h(\mathbf{x})$, where $v(\mathbf{x})$ and $h(\mathbf{x})$ are vertical and horizontal gradients, respectively. The orientation $\theta(\mathbf{x})$ is then coded into eight discrete labels. We divide a candidate rectangle into several spatial blocks and create a direction histogram for each block. Finally, all 8-D histograms are concatenated and the combined feature vector is used for classification. An orientation histogram H_i for each block is defined as $H_i = \#\{\mathbf{x} | \theta(\mathbf{x}) = i\}$, where # is the number of elements in the set and i is the direction label.

We adopt CoHOG or Co-occurrence Histograms of Oriented Gradients [14] as a standard feature descriptor. The CoHOG feature can be thought as the extension of the HOG feature. The basic idea is to handle gradient orientations *in pairs* instead of individually. The CoHOG indicates the joint histogram of oriented gradients of two pixels at a certain displacement. For example, a pair of two horizontally adjacent pixels generates 8×8 dimensional histogram as

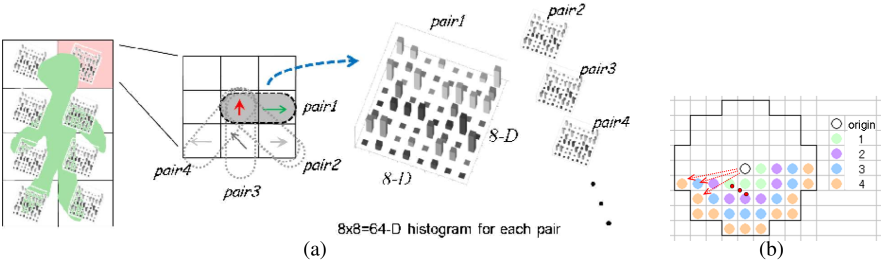
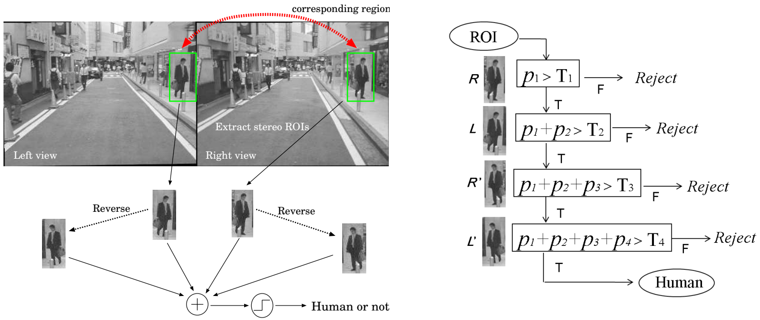


Figure 3: (a) Overview of CoHOG descriptor. (b) Typical 30 offset vectors.



(a) (b)

Figure 4: Detection via multiple monocular-based classifiers.

shown in Figure 3(a). A different pair also creates another 64-D histogram. Assuming **d** to be a certain 2D displacement vector, the (i, j)-th element of the histogram is defined as

$$H_{ij} = \#\{\mathbf{x} | \theta(\mathbf{x}) = i, \theta(\mathbf{x} + \mathbf{d}) = j\}. \tag{1}$$

In our current implementation, we use 30 pairs whose Chebyshev distances from each other are up to four pixels as shown in Figure 3(b). These pairs generate a 1920 (= 64×30) dimensional feature. Combined with a HOG histogram, we define a 1928-D histogram for each block and concatenate them to generate a CoHOG descriptor. Typically we divide a candidate rectangle into 3 × 6 blocks which create a 34704 (=1928×3×6) dimensional feature descriptor in total. CoHOG feature descriptors have extensive vocabulary and outperform HOG features for human detection as reported in [14].

4 Detection with Multiple classifiers

In addition to our baseline detection system stated above, we introduce detection methods to use corresponding regions on left and right images as shown in Figure 4(a) in order to improve classification performance. Let the right view be reference one where original ROIs are generated. The corresponding ROIs are defined on the left image using dominant disparity values within the original ROIs. The simplest way to utilize those pairs of ROIs is to evaluate each candidate rectangle individually and integrate the classification scores to determine if it is a human or not. As shown in Figure 4(a), we also employ left and right reversed image patterns besides left and right patterns. Four identical classifiers, each of which is learned from *monocular* training samples, evaluate the four intensity patterns individually and the total value of the outputs is then computed to determine if it is a pedestrian or not.

Although we can expect the improvement in classification accuracy as it employs multiple intensity patterns instead of a unique pattern, the computational cost becomes higher as it needs to classify more than once. Let N be the number of original ROIs in the reference image and M be the number of image patterns per an original ROI. The number of classification operations becomes up to $M \times N$. However, the cascade structure is effective to accelerate the processing since it makes processing times for subsequent steps smaller. Thus, it makes the total number of times of classification less than $M \times N$ as follows.

We refer to each of the four classifiers as i-th classifier whose output is $p_i (i = 1 \sim 4)$. The output p_i is normalized from 0 to 1 where 1 represents positive class. In the following description, we call the normalized score p_i classification probability. A candidate region is classified as positive if $\sum_{i=1}^{4} p_i > T_{total}$, where T_{total} is a predefined threshold. Figure 4(b) shows the detection cascade and we describe below the cascade procedure assuming $T_{total} = 3.2$, which means that the candidate region is classified as positive if the average probability is higher than $0.8(= T_{total}/4)$. The first classifier evaluates all candidate regions and rejects a large number of negative examples. The sum of subsequent classification probabilities, $\sum_{i=2}^{4} p_i$, is 3.0 at most. Therefore, if $p_1 \leq T_1$ where $T_1 = 0.2(= T_{total} - 3.0)$, we can reject the samples without further procedures. Then, the second classifier evaluates those candidate regions selected by the first classifier. As the sum of subsequent classification probabilities, $p_3 + p_4$, is 2.0 at most, we can discard the samples as negative without further procedures if $p_1 + p_2 \leq T_2 = 1.2(= T_{total} - 2.0)$. Analogously, the third classifier examines remaining candidates and eliminates the samples whose total probability $p_1 + p_2 + p_3$ is less than $T_3 = 2.2(= T_{total} - 1.0)$. The fourth classifier examines selected candidates and decides whether they are human or not by $\sum_{i=i}^{4} p_i > T_{total}$.

Figure 5 shows a detection example. Totally, 151 ROIs are generated as depicted by green windows. Out of these candidates, 13 regions are finally classified as pedestrians as shown in Figure 5(b). No post-processing is performed for merging overlapping detections. The numbers of candidates which the $1st$, $2nd$, $3rd$ and $4th$ classifier evaluate are 151, 61, 24 and 19, respectively. The first classifier rejects 90 samples as negative, which are about 60% of the original candidates. The number of classification processes is $255(= 151 + 61 + 24 + 19)$ in total while it is 151 when we use a single classifier alone. Therefore, even if we use four classifiers, the cascade structure allows us to detect humans at about 1.7 times more computational cost in this scene. The number of times of classification varies mainly depending on the threshold T_{total} and how many people there are. In our experiment, the cascade detection using four patterns needs roughly less than twice as much computational cost as the monocular-based detection does.

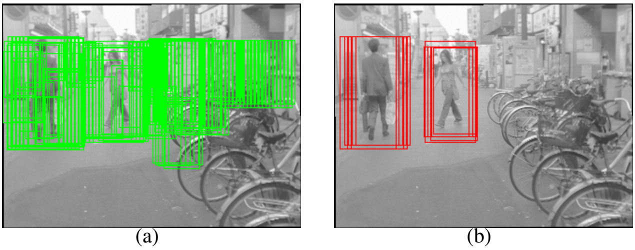
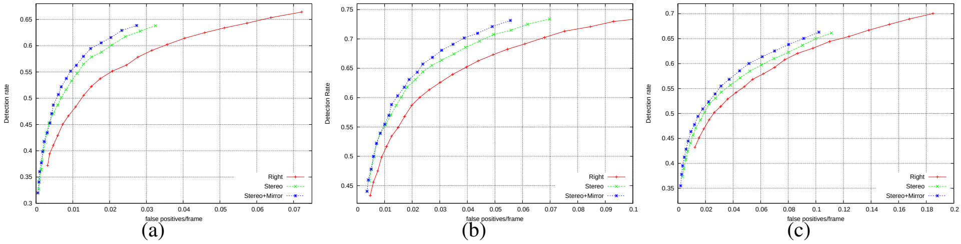

(a) (b)

Figure 5: (a) Extracted ROIs. (b) Detected pedestrians.

4.1 Experimental Results

We have performed some experiments in which the proposed method was applied to outdoor stereo images. The stereo images were acquired with a pair of cameras mounted on a car driven in urban environments. We use CCD cameras with 12.5mm lenses. The stereo baseline is about 70cm and each camera is at the height of about 115cm. We have tested our method on three sequences, each of which contains 20,000 frames roughly equivalent to 11 minutes. Each detection is counted as correct if it overlaps with an annotation by more than



(a) (b) (c)

Figure 6: ROC curves on three different stereo sequences.

$1/3$ using the intersection-over-union measure [2] or IoU given by

$$IoU = \frac{area(R_d \cap R_g)}{area(R_d \cup R_g)} = \frac{area(R_d \cap R_g)}{area(R_d) + area(R_g) - area(R_d \cap R_g)}, \tag{2}$$

where R_d and R_g are a detected region and a ground truth annotation respectively.

Figure 6 shows qualitative results on the three sequences. The y-axis corresponds to the detection rate and the x-axis to the number of false positives per frame. Higher curves are better. Red curves indicate the detection performance using a unique intensity pattern extracted from the right image. Green ones represent results using a pair of ROIs on stereo views. Blue plots corresponds to results using four patterns; stereo patterns and left-right mirrored patterns of them. These results show a clear advantage of the pedestrian detection using stereo image regions and the performance becomes better when the left-right reflections are combined with them. Table 1 shows the total number of false detections of “Sequence (a)” when the detection rate is fixed at 60%. The stereo patterns reduce about 40% of false detections which further decreases by half when mirrored patterns are taken into account.

Table 1: A comparison of the total number of false detection at detection rate of 60%.

	single	stereo	stereo+mirrored
# of FPs	695	405	324
Rate of reduction	-	41.7%	53.4%

5 Stereo Feature Descriptors

For more precise classification, this section describes an alternative method to utilize stereo ROIs. We introduce new feature descriptors which combine local appearances and stereo disparities. Figure 7 shows a pedestrian detection overview using our stereo feature descriptors. Our stereo feature descriptor is obtained from a pair of input images and its disparity measurements. We have investigated both HOG-based and CoHOG-based stereo descriptors which we call *Stereo HOG* and *Stereo CoHOG*, respectively. In the following subsections, we describe the detail of these two stereo feature descriptors.

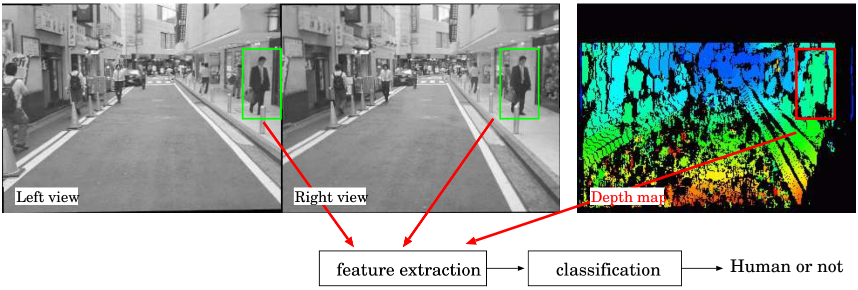


Figure 7: Pedestrian detection based on our stereo feature descriptor.

5.1 Histogram of Relative Disparities

In our stereo descriptor, features are extracted from a disparity map as well as from stereo images. Although the original HOG features are extracted from oriented gradients, it is difficult to calculate oriented gradients from disparity maps as they have regions of unknown disparities as shown in Figure 7. Also, since an absolute disparity value corresponds to the distance between stereo cameras and target objects, it is not an appropriate feature to discriminate between humans and the other objects. Instead of oriented gradients of disparities or raw disparity values, we use relative disparity values given as,

$$\Delta d(x,y) = \begin{cases} occ. & \text{if } d(x,y) = occ. \\ \min(|d(x,y) - d_{ref}|, T) & \text{otherwise} \end{cases}, \quad (3)$$

where $d(x,y)$ represents a disparity value of an image point (x,y), d_{ref} indicates the dominant disparity within a candidate rectangle. The label $occ.$ represents unknown disparities and T is a pre-defined threshold for a saturating operation. Based on the relative disparity $\Delta d(x,y)$, histograms of relative disparities or HRD are created in the same way as for the standard HOG and co-occurrence histograms of relative disparities or CoHRD are generated in the same procedure as for the monocular CoHOG. Figure 8(a) shows an example of histogram of relative disparities or HRD. Foreground and background features are roughly divided in the histogram.

5.2 Foreground/background separation

In general, the presence of cluttered background makes human detection even more difficult. When we refer to disparity values, it is possible to roughly separate foreground regions from background ones. Background regions normally have rather different disparities from the majority of disparities or d_{ref}. Figure 8(b) shows a pair of foreground regions obtained from disparity estimates. Now we have two options for utilizing this foreground/background separation. The first one is to remove background image regions and generate a feature descriptor of foreground image areas exclusively. We call feature descriptors of foreground regions alone F-HOG and F-CoHOG. The second one is to remain feature descriptors of background regions and combine them with foreground feature descriptors. As foreground/background segmentation is not perfect, some critical features may lie on background regions as a result. We call combined feature descriptors of foreground and background regions FB-HOG or FB-CoHOG.

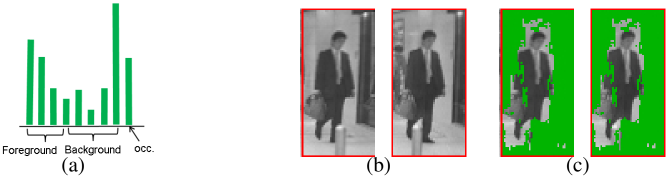


Figure 8: (a) Histogram of relative disparities. (b) A pair of ROIs. (c) Its foreground regions.

5.3 Experimental Results

We have performed some qualitative evaluations of various combinations of different features. For this experiment, we use 4,472 positive and 6,230 negative samples for training. The test data consists of 3,103 human and 5,011 non-human image regions. As our feature descriptors are extracted from stereo images and its disparity estimates, each training or test sample is a triplet of corresponding image regions and its disparity estimates.

HOG: Figure 9 shows the performance of the various HOG-based descriptors on our data sets. The monocular HOG is the original HOG descriptor which uses a pair of image regions independently for both training and test phases. The concatenated HOG is a feature vector which we just concatenate a pair of HOGs from stereo views. It results in poor performance and this means that the feature descriptor may overfit the training data. The average HOG or the average vector of the left and right HOGs is more distinctive than the monocular HOG. The HRD descriptor, which is extracted from disparity estimates alone, is more powerful than the average HOG. The F-HOG, extracted from foreground regions exclusively, is more discriminative than the HRD. The FB-HOG is a concatenated vector of the F-HOG and B-HOG generated from background regions. The FB-HOG is more discriminative than the F-HOG. It shows that critical features exist on background regions since the foreground and background separation is not accurate. The FB-HOG+HRD is a combined vector of the FB-HOG and the HRD, which achieves the highest performance among various HOG-based descriptors. Given the detection rate fixed at 95%, the false positive rate of the FB-HOG+HRD is 2.4% while the rate of the original or monocular HOG is about 24.5%. It

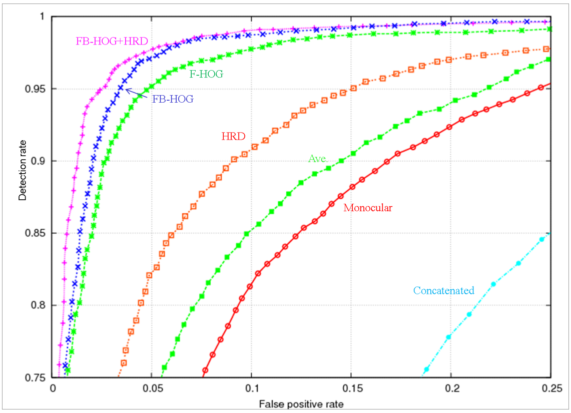


Figure 9: Results by various HOG descriptors.

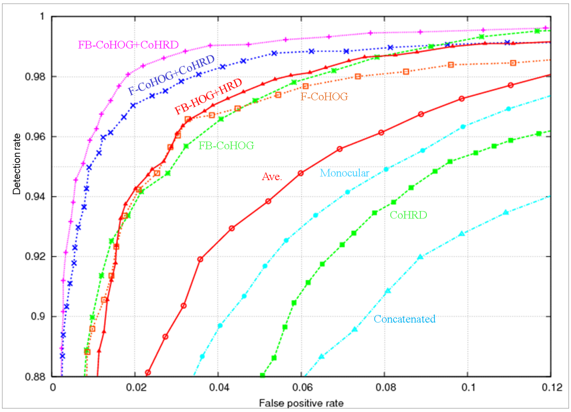


Figure 10: Experimental results by various CoHOG-based descriptors. Note that the scales are different from those of Figure 9.

shows that our stereo-based feature descriptor decreases the false alarm rate by an order of magnitude comparing to the monocular-based HOG descriptor.

CoHOG: Figure 10 shows the performance of the various CoHOG-based descriptors on the same stereo data sets. ROC curves move leftward and upward as the original CoHOG by itself is more distinctive than the original HOG. Again, it is clear that stereo disparities and the foreground/background separation contribute to accurate classification. FB-CoHOG and F-CoHOG features are more discriminative than the average vector of left and right CoHOGs. Although the disparity descriptor or CoHRD is not discriminative comparing to the monocular or original CoHOG, CoHRD makes a great deal of contribution when combined with FB-CoHOG. The best combined descriptor is FB-CoHOG+CoHRD which is a concatenated vector of the foreground and background CoHOGs and the CoHRD from stereo disparities. Given the detection rate fixed at 95%, the false positive rate of the FB-CoHOG+HRD is about 0.75% while the rate of the original CoHOG is about 8%. Our stereo-based feature descriptor, therefore, decreases the false alarm rate by an order of magnitude comparing to the monocular-based descriptor also in case of CoHOG. Table 2 summarizes the improvements on classification accuracy of HOG-based and CoHOG-based stereo features.

Table 2: Comparison of false positive rates at 95% detection rate.

	monocular	stereo
HOG	24.5%	2.4%
CoHOG	8.0%	0.75%

6 Summary and conclusions

In this paper we have proposed a new stereo-based pedestrian detection framework which uses pairs of ROIs defined on both left and right view instead of monocular ROIs. This paper introduces two different approaches. Firstly, we use multiple classifiers to evaluate multiple patterns individually and combine those outputs to obtain the final decision. In this scheme, the simple cascade detector is effective to reduce computational cost. Secondly, we extend gradient-based local descriptors to multiple views and propose new stereo feature descriptors or Stereo HOG and Stereo CoHOG which combine local appearances and stereo disparities. The quantitative results show that both approaches significantly reduces the false alarm rate while maintaining the high detection rate comparing with monocular-based detectors. In future work, we will investigate how to combine various detectors for computational efficiency and how to combine motion descriptors with stereo ones.

References

[1] N. Dalal and B.Triggs. Histograms of oriented gradients for human detection. In *CVPR*, volume 2, pages 886–893, 2005.

[2] A. Ess, B. Leibe, and L.van Gool. Depth and appearance for mobile scene analysis. In *ICCV*, pages 1–8, 2007.

[3] A. Ess, B. Leibe, K. Schindler, and L. van Gool. A mobile vision system for robust multi-person tracking. In *CVPR*, pages 1–8, 2008.

[4] D. M. Gavrila and S. Munder. Multi-cue pedestrian detection and tracking from a moving vehicle. *IJCV*, 73(1):41–59, 2007.

[5] D. M. Gavrila and V.Philomin. Real-time object detection for "smart" vehicles. In *ICCV*, pages 87–93, 1999.

[6] H. Hattori and N. Takeda. Dense stereo matching in restricted disparity space. In *Proc. of the IEEE Intelligent Vehicles Symposium*, pages 118–123, 2005.

[7] B. Leibe, N. Cornelis, K. Cornelis, and L.van Gool. Dynamic 3D scene analysis from a moving vehicle. In *CVPR*, pages 1–8, 2007.

[8] B. Leibe, K. Schindler, and L. Van Gool. Coupled detection and trajectory estimation for multi-object tracking. In *ICCV*, pages 1–8, 2007.

[9] P. Kanter M. Enzweiler and D. M. Gavrila. Monocular pedestrian recognition using motion parallax. In *Proc. of the IEEE Intelligent Vehicles Symposium*, pages 792–797, 2008.

[10] S. Munder and D. M. Gavrila. An experimental study on pedestrian classification. *PAMI*, 28(11):1863–1868, 2006.

[11] C. Papageorgiou and T. Poggio. A trainable system for object detection. *IJCV*, 38(1): 15–33, 2007.

[12] A. Shashua, Y. Gdalyahu, and G. Hayun. Pedestrian detection for driving assistance systems: Single-frame classification and system level performance. In *Proc. of IEEE Intelligent Vehicles Symposium*, pages 1–6, 2008.

[13] D. Snow, M.J. Jones, and P. Viola. Detecting pedestrians using patterns of motion and appearance. In *ICCV*, pages 734–741, 2003.

[14] T. Watanabe, S. Ito, and K. Yokoi. Co-occurrence histograms of oriented gradients for pedestrian detection. In *PSIVT*, pages 37–47, 2009.