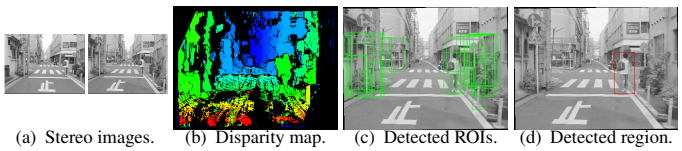
# Stereo-based Pedestrian Detection using Multiple Patterns

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

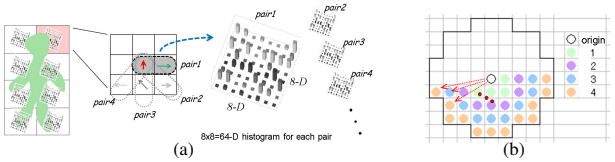
Detecting pedestrians from a moving vehicle is a challenging problem since the essence of the task is to search non-rigid moving objects with various appearances in a dynamic and outdoor environment. In order to alleviate these difficulties, we propose a new human detection framework which makes the 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, 2] 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.



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

Figure 1: Detection Overview.

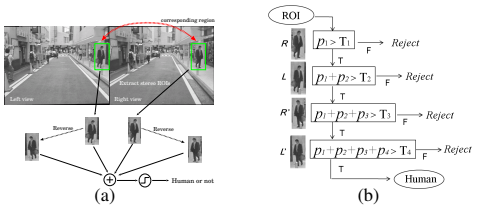
Figure 1 provides an overview of our stereo-based human detection system. Basically, uniform disparity regions are extracted as shown in Figure 1(c) illustrated with green rectangles and these regions are then fed to a pattern classifier which decides if the candidate regions contain a pedestrian or not. We adopt CoHOG or Co-occurrence Histograms of Oriented Gradients [2] 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 $8 \times 8$ dimensional histogram as shown in Figure 2(a). A different pair also creates another 64-D histogram. In our current implementation, we use 30 pairs whose Chebyshev distances from each other are up to four pixels as shown in Figure 2(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 $3 \times 6$ 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 [2].




(a) (b)

Figure 2: (a) CoHOG descriptor. (b) Typical 30 offset vectors.

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 3(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




(a) (b)

Figure 3: Detection via multiple monocular-based classifiers.

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 3(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. However, the cascade structure is effective to accelerate the processing since it makes processing times for subsequent steps smaller.

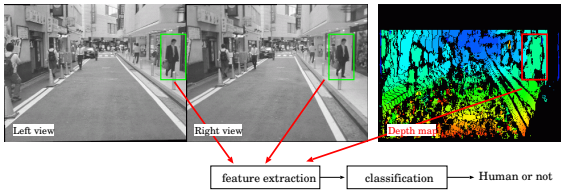



Figure 4: Pedestrian detection based on our stereo feature descriptor.

For more precise classification, the following describes an alternative method to utilize stereo ROIs. We introduce new feature descriptors which combine local appearances and stereo disparities. Figure 4 shows a 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. The detail of these two stereo feature descriptors are described in the paper. Table 1 summarizes the improvements on classification accuracy of HOG-based and CoHOG-based stereo features. Given the detection rate fixed at 95%, the false positive rate of the best Stereo CoHOG 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 1: Comparison of false positive rates at 95% detection rate.

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

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

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