Hindawi Publishing Corporation
Abstract and Applied Analysis
Volume 2013, Article ID 436062, 11 pages
http://dx.doi.org/10.1155/2013/436062

Research Article

Analysis of Feature Fusion Based on HIK SVM and Its Application for Pedestrian Detection

Song-Zhi Su[1,2] and Shu-Yuan Chen[3]

[1] *School of Information Science and Technology, Xiamen University, Xiamen, Fujian Province 361005, China*
[2] *Fujian Key Laboratory of the Brain-Like Intelligent Systems (Xiamen University), Xiamen, Fujian Province 361005, China*
[3] *Department of Computer Science and Engineering, Yuan Ze University, Taoyuan 320, Taiwan*

Correspondence should be addressed to Song-Zhi Su; ssz@xmu.edu.cn

Received 5 March 2013; Accepted 1 April 2013

Academic Editor: Zhenkun Huang

This work presents the fusion of integral channel features to improve the effectiveness and efficiency of pedestrian detection. The proposed method combines the histogram of oriented gradient (HOG) and local binary pattern (LBP) features by a concatenated fusion method. Although neural network (NN) is an efficient tool for classification, the time complexity is heavy. Hence, we choose support vector machine (SVM) with the histogram intersection kernel (HIK) as a classifier. On the other hand, although many datasets have been collected for pedestrian detection, few are designed to detect pedestrians in low-resolution visual images and at night time. This work collects two new pedestrian datasets—one for low-resolution visual images and one for near-infrared images—to evaluate detection performance on various image types and at different times. The proposed fusion method uses only images from the INRIA dataset for training but works on the two newly collected datasets, thereby avoiding the training overhead for cross-datasets. The experimental results verify that the proposed method has high detection accuracies even in the variations of image types and time slots.

1. Introduction

Pedestrian detection is an active area of research in the field of computer vision [1, 2] and a preliminary task in various applications, including intelligent video surveillance, automotive robotics, content-based image annotation/retrieval, and management of personal digital images. Large variations in appearance caused by articulated body motion, viewpoint, lighting conditions, occlusions, and cluttered backgrounds present serious challenges. Hence, pedestrian detection in still images is more difficult than that of faces [3].

Most pedestrian detection methods use a pretrained binary classifier to find pedestrians in still images by scanning the entire image. Such a method is called the “sliding window method” (or scanning window). The classifier is “fired” if the image features inside the local search subwindow satisfy certain criteria. At the core of the sliding window framework are image descriptors and classifiers that are based on these descriptors. According to features used for pedestrian detection, these methods can be divided into three groups: holistic-based methods, part-based methods, and patch-based methods.

Holistic-based methods use global features, such as edge templates, histogram of oriented gradient [4], and Haar-like wavelet [5]. One popular holistic-based method is the histogram of oriented gradient (HOG) method, which has near-perfect classification performance when applied to the original MIT pedestrian database and is widely used in other computer vision tasks, such as scene classification, object tracking, and pose estimation. Part-based methods model a pedestrian as a set of parts, which include legs, torso, arms, and head. Hypotheses concerning these parts are generated by learning local features such as edgelet [6] and orientation features. These parts are then assembled to form a final human model based on geometric constraints. Accurate pedestrian detection depends on accurate part detection and pedestrian representation by parts. Though this approach is effective for dealing with partially occluded pedestrian detection, part detection is a difficult task. One example of a patch-based method is implicit-shape-model- (ISM-) based object detection, developed by Leibe et al. [7], which combines both detection and segmentation in a probability framework and requires only a few training images. However, constructing a smart and discriminative codebook from various perspectives remains an open problem.

Numerous descriptors used in pedestrian detection have recently been proposed. Zhao and Thorpe [8] proposed a pedestrian method by a pair of moving camera through stereo-based segmentation and neural network-based recognition. Dalal and Triggs [4] developed a descriptor similar to scale invariant feature transform (SIFT) [9], which encodes HOG in the detection window. HOG has been subsequently extended to describe histograms that present information on motion. Felzenszwalb et al. [10] recently applied HOG to their deformable part models and obtained promising results in the PASCAL VOC Challenge. Zhu et al. [11] implemented a cascade of rejecters based on HOG descriptors to achieve near-real-time performance. Cascade models have also been successfully used with other types of pedestrian descriptors, such as edgelet features and the region of covariance (COV) [12].

In order to integrate various pedestrian descriptors, many works have proposed fusing multiple features to detect pedestrians. Wojek et al. [13] combined the oriented histogram of flow with HOG or Haar on an onboard camera setup and concluded that incorporating motion information considerably enhances detection performance. Y. T. Chen and C. S. Chen [14] proposed a method for detecting humans in a single image, based on a novel cascade structure with metastages. Their method includes both intensity-based rectangular features and gradient-based 1D features in the feature pool of the Real AdaBoost algorithm for weak classifier selection. Wang et al. [15] combined HOG and local binary pattern, trained by a linear SVM, to solve the partial occlusion problem.

However, multicue pedestrian detection methods have the following disadvantages for detecting pedestrians in still images. First, optical flow information cannot be extracted from a single image. Second, edgelet extraction or the COV feature is computation-intensive. Finally, the AdaBoost has too many parameters to tune, and the cascading test is time-consuming and sensitive to occlusion. Therefore, this work uses HOG and LBP features, which can be extracted efficiently by integral images. An SVM with a linear kernel or HIK [16] has the advantage of ease of training in the training stage and fast prediction in the test stage [17].

Although many datasets have been collected for pedestrian detection, few are designed to detect pedestrians in cross-dataset, which is still a hot topic in computer vision. Vazquez et al. [18] proposed an unsupervised domain adaptation of virtual and real worlds for pedestrian detection. Jain and Learned-Miller [19] proposed an online approach for quickly adapting a pretrained classifier to a new test dataset without retraining the classifier. In this work, we collect two new pedestrian datasets—one for low-resolution visual images and one for near infrared images—to evaluate detection performance on various image types and at different times. This work proposes cross-dataset pedestrian detection by fusing integral channel features, which use only images from the INRIA dataset for training but are effective on the two newly collected datasets, thereby avoiding the training overhead for cross-datasets.

The remainder of this paper is organized as follows. Section 2 offers a description of the proposed method, including the features, classifiers, and fusion. Section 3 presents and offers a discussion of the relevant experimental results. Finally, Section 4 draws a conclusion and presents suggestions and directions for future work.

2. Proposed Pedestrian Detection Method

Sliding window-based object detection algorithms for static images consist of two components: feature extraction and classifier training. Feature extraction encodes the visual appearance of a detected object using object descriptors. Classifier training trains a classifier to determine whether the current searching window contains a pedestrian. In this section, we discuss the features and classifiers.

2.1. Feature Extraction. Several methods for describing pedestrians have recently been proposed. This work uses HOG and LBP as pedestrian descriptors. All of these features can be extracted using integral histogram techniques, accelerating the computation process. They are complementary because they encode gradient and texture information, respectively.

HOG. The HOG proposed by Dalal and Triggs [4] has been widely used in the computer vision field, including object detection, recognition, and classification. HOG is similar to edge orientation histograms, shape context, and the SIFT descriptor, but it is computed on a dense grid of uniformly titled cells. Overlapping local contrast normalization in blocks is conducted to improve accuracy. HOG implementation involves dividing search windows into small-connected regions, called cells, for which the histogram of gradient directions is computed (Figure 1(a)). In this work, an HOG descriptor is implemented using the following parameters. Image derivatives in x and y directions are obtained by applying the masks $[-1\ \ 0\ \ 1]$ and $[-1\ \ 0\ \ 1]^T$, respectively. The gradient orientation is linearly voted into nine orientation bins in the range 0°–180°. A block size is 16 × 16; a cell size is 8 × 8; blocks overlap half of a cell in each direction; Gaussian is weighting with $\sigma = 4$ using an L2-norm for the feature vector in a block. The final vector consists of all normalized block histograms, yielding 3780 dimensions.

LBP. Various applications have applied the local binary pattern (LBP) extensively, which is highly effective in texture classification and face recognition because it is invariant to monotonic changes in the gray level [20]. Wang et al. [15] noted that HOG performs poorly when the background is cluttered with noisy edges and LBP is complementary when it exploits the uniform pattern concept (Figure 2). In this work, we adopt eight sample points and require bilinear interpolation to find the red points in Figure 2(a) with a radius of one and take the l_∞ distance as the distance to the central pixel. The number of 0/1 transitions is no more than

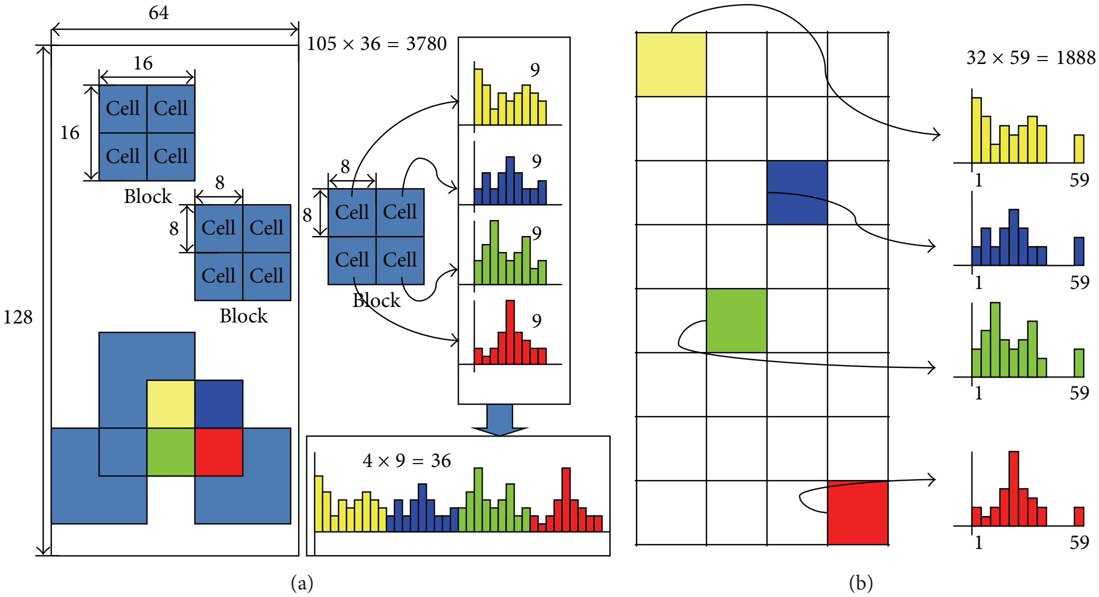


Figure 1: Feature extraction using HOG and LBP.

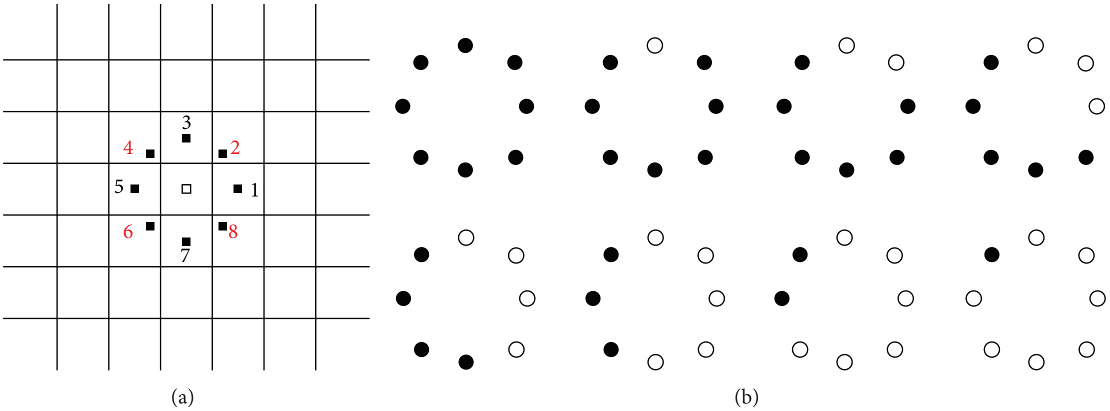


Figure 2: (a) Eight sample points of the local binary pattern; (b) examples of the uniform local binary pattern.

two. Different uniform patterns are counted into different bins, and all nonuniform patterns are voted into one bin. A cell includes 59 bins, and the L2-norm is adopted to normalize the histogram. We used the procedure by Wang et al. to extract the LBP feature and to directly establish pattern histograms in the cells (16 × 16, without overlap, as shown in Figure 1(b)). LBP histograms in the 32 cells are then concatenated into a feature vector with dimensions of 59 × 32 = 1888 to describe the texture in the current search window.

2.2. Classifier Training. Linear SVM and AdaBoost are widely used for detecting pedestrians. This work focuses on an SVM with different kernel functions because it is easy to train in the training stage and can make rapid predictions in the test stage. Linear SVMs learn the hyperplane that separates pedestrians from the background in the original feature space. Extended versions of SVM, such as RBF kernel SVMs, transform data to a high and potentially infinite number of dimensions. However, the extensions are seldom used in pedestrian detection because more dimensions lead to computational overload.

Maji et al. [16] recently approximated the histogram intersection kernel of SVM (HIKSVM) to accelerate prediction, and Wu [17] proposed a fast dual method for HIKSVM learning. Section 3 describes experiments conducted to compare the performance of a linear SVM with that of HIKSVM. The experimental results show that HIKSVM outperforms the linear SVM. A brief introduction of HIKSVM follows.

Swain and Ballard [21] first proposed the HIK, which is widely used as a measure of similarity between histograms. Researchers have proven that HIK is positive definite and can be used as a discriminative kernel function for SVMs. However, the HIK requires memory and computation time that is linearly proportional to the number of support vectors because it is nonlinear. Maji et al. presented HIKSVMs with a runtime complexity, that is, the logarithm of the number of

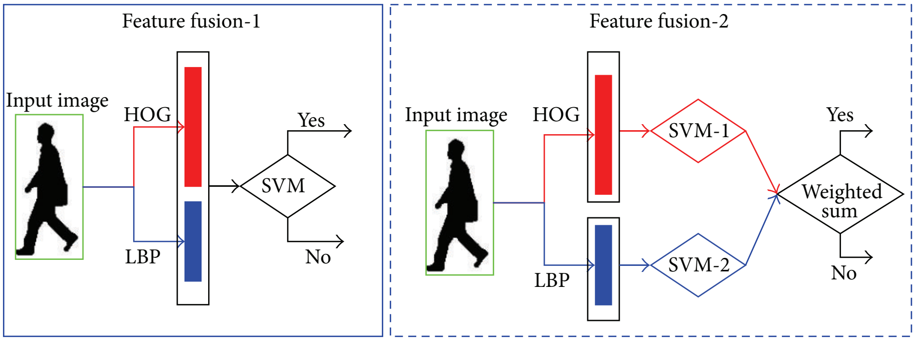


Figure 3: Examples of two feature fusion methods using the SVM as a classifier.

support vectors. Based on precomputing auxiliary tables, an approximate classifier can be constructed with runtime and space requirements that are independent of the number of support vectors.

For feature vectors $\mathbf{x}, \mathbf{y} \in R_{+}^{n}$, the HIK can be expressed as follows:

$$k_{\mathrm{HI}}(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{n} \min(x(i), y(i)), \tag{1}$$

and the corresponding discriminative function for a new input vector $\mathbf{x}$ is

$$\begin{aligned} h(\mathbf{x}) &= \sum_{l=1}^{m} \alpha_l y_l k(x, x_l) + b \\ &= \sum_{l=1}^{m} \alpha_l y_l \left(\sum_{i=1}^{n} \min(x(i), x_l(i)) \right) + b. \end{aligned} \tag{2}$$

Maji et al. noticed that for intersection kernels, the summations in (2) can be reformed as follows:

$$\begin{aligned} h(\mathbf{x}) &= \sum_{i=1}^{n} \left(\sum_{l=1}^{m} \alpha_l y_l \min(x(i), x_l(i)) \right) + b \\ &= \sum_{i=1}^{n} h_i(x(i)) + b, \end{aligned} \tag{3}$$

where $h_i(s) = \sum_{l=1}^{m} \alpha_l y_l \min(s, x_l(i))$. Consider the functions $h_i(s)$ at fixed point i; $\overline{x}_l(i)$ represents the sorted values of $x_l(i)$ in increasing order with corresponding values of α and labels given by $\overline{\alpha}_l$ and $\overline{y}_l$. According to the HIK, let r be the largest integer, such that $\overline{x}_r(i) \leq s$; therefore,

$$\begin{aligned} h_i(s) &= \sum_{l=1}^{m} \alpha_l y_l \min(s, x_l(i)) \\ &= \sum_{1 \leq l \leq r} \overline{\alpha}_l \overline{y}_l \overline{x}_l(i) + s \sum_{r < l \leq m} \overline{\alpha}_l \overline{y}_l \\ &= A_i(r) + s B_i(r), \end{aligned} \tag{4}$$

where $A_i(r) = \sum_{1 \leq l \leq r} \overline{\alpha}_l \overline{y}_l \overline{x}_l(i)$, $B_i(r) = \sum_{r < l \leq m} \overline{\alpha}_l \overline{y}_l$. Clearly, (4) is piecewise linear, and functions A_i and B_i are independent of the input data. Therefore, s can be precomputed by first finding the position of $s = x(i)$ in the sorted list by binary search, with a runtime complexity of $O(\log m)$. Although the runtime complexity of computing $h(x)$ is $O(n \log m)$, it necessitates to double the storage that is required by the standard implementation because the modified version must store $\overline{x}_l$ and $h_i(\overline{x}_l)$.

Maji et al. found that the support distributions in each dimension tend to be smooth and concentrated. Therefore, the $h(x)$ is relatively smooth and can be approximated by simpler functions, greatly reducing the required storage and accelerating the prediction. In this work, $h_i(s)$ is computed using a lookup table with a piecewise constant approximation.

2.3. Feature Fusion. The two main feature fusion methods (Figure 3) are concatenated fusion (FF1) and weighted sum (FF2). Concatenated fusion concatenates different feature descriptors and then feeds the concatenated results into the classifier. The weighted sum feeds different features into individual classifiers and then combines classification scores using a weighted sum.

This work fuses HOG and LBP features for detecting pedestrians because both can be implemented by integral histogram approaches, accelerating the subsequent prediction process, as described in Section 3. Let the output scores of the individual SVM classifiers using HOG and LBP features be f_{HOG} and f_{LBP}, respectively. For the FF2 fusion method, the final output score is then defined by the weighted sum

$$f = \alpha f_{\mathrm{HOG}} + (1 - \alpha) f_{\mathrm{LBP}}, \quad 0 < \alpha < 1. \tag{5}$$

The values of α to $\alpha \in \{\alpha \mid \alpha = 0.1K, K = 1, 2, \ldots, 9\}$ are herein. Section 3 verifies that FF1 performance is superior to FF2 for all of the values of α, and FF2 has the best performance when $\alpha = 0.5$. Hence, this work fuses HOG, LBP, and Haar using HIKSVM by the FF1 method because this method is highly accurate, as confirmed in Section 3.

3. Experimental Results

The accuracies achieved using various integral channel features, different kernels of support vector machines, and two feature fusion methods for detecting pedestrians are extensively compared. Random noise blocks are added to

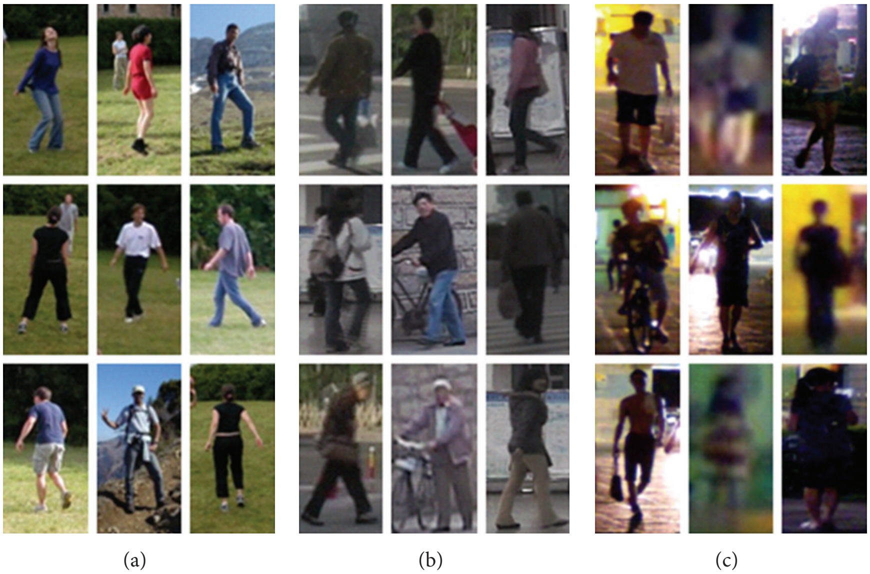


Figure 4: Examples of pedestrian images: (a) INRIA; (b) XMU-VIS; (c) XMU-NIR.

Table 1: INRIA training and test sets, XMU-VIS test sets, and XMU-NIR datasets.

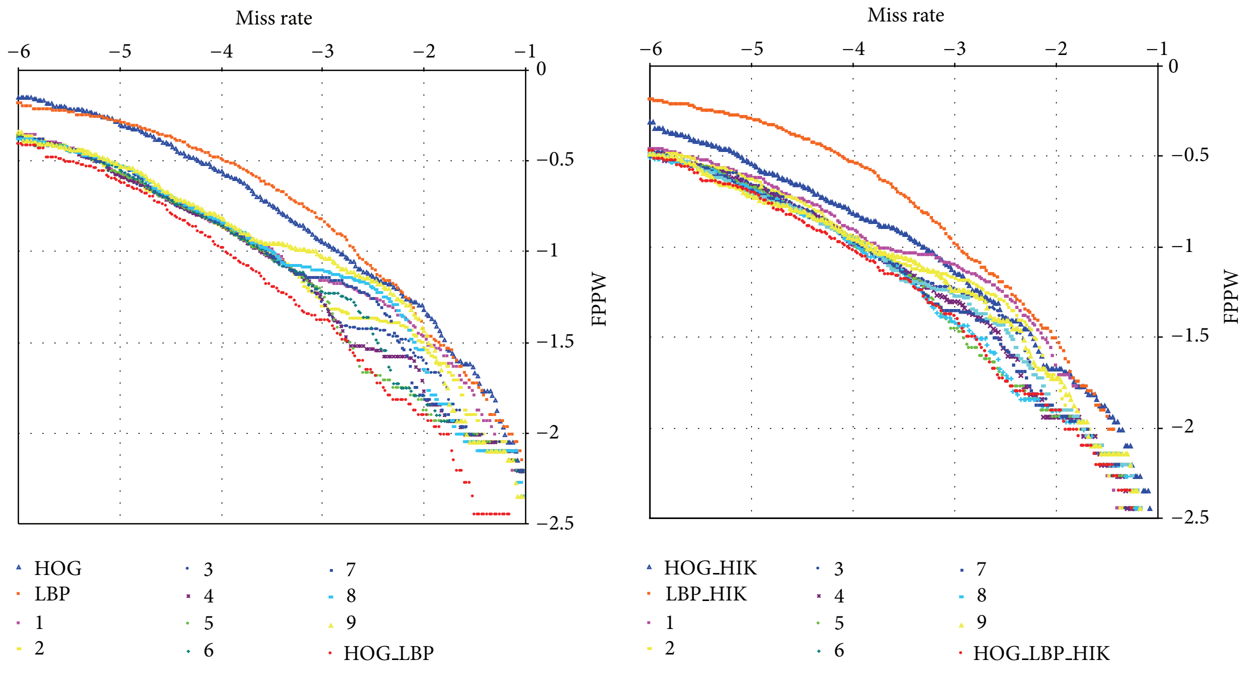
	Training				Test			
	Pedestrians		Nonpedestrians		Pedestrians		Nonpedestrians	
	#imgs	#win	#imgs	#win	#imgs	#win	#imgs	#win
INRIA	615	2416	1218	22111	288	1126	453	4484965
XMU-VIS	—	—	—	—	4207	10154	413	1834994
XMU-NIR	—	—	—	—	1057	2596	—	—

the pedestrian image to test the robustness achieved using various features and classifiers. Experimental results obtained using the INRIA person dataset and two newly collected Xiamen databases indicate that the combined HOG and LBP features by the concatenated-fusion method using the SVM with the HIK as a classifier yield the highest accuracy. The multiple feature combination outperforms single features, and the HIK consistently outperforms the linear SVM.

3.1. Dataset and Performance Evaluation Measures. This work evaluates the performance of pedestrian detection using three databases: the INRIA person database [4] and two new databases collected at Xiamen University, called XMU-VIS and XMU-NIR, respectively. The INRIA dataset contains human images taken from several viewing angles under various lighting conditions both indoors and outdoors. Figure 4(a) shows samples of the INRIA dataset. INRIA images fall into three groups, which are further divided into training and testing sets. The first group is composed of 615 full-size positive images containing 1208 pedestrian instances for training and 288 images containing 566 instances, for testing. The second group comprises scale-normalized crops of humans sized 64 × 128, including 2416 positive images for training and 1126 positive images for testing. The third group comprises full-size negative images including 1218 images for training and 453 images for testing.

This work used 2416 scale-normalized crops of human images as positive training samples and randomly sampled 22111 subimages from 1218 person-free training photographs as negative training samples. All of the training images are from the INRIA dataset, including the situations of test images from XMU-VIS or XMU-NIR datasets, to show cross-dataset human detection. For the INRIA dataset, the 1126 cropped images of pedestrians were used for testing. The negative test samples were obtained by scanning the 453 testing images in steps of eight pixels in the x- and y-directions using five scales (0.8, 0.9, 1.0, 1.1, and 1.2) of image size, yielding 4484965 negative cropping windows.

The XMU-VIS dataset was collected at various places around Xiamen University and at different time. The size of each pedestrian image in the XMU-VIS dataset is 640 × 480 smaller than that of INRIA in 720 × 576. The goal is to simulate images captured by onboard cameras in intelligent vehicles for detecting pedestrians in low resolution. The XMU-VIS test set is composed of 4207 pedestrian images with 10154 cropped images and 413 negative images with 1834994 cropped images. The XMU-NIR dataset was also collected at various locations around Xiamen University and at different times. The images captured by near-infrared sensors were sized 1280 × 720. The XMU-NIR dataset consists of 1057 pedestrian images, in which 2596 are pedestrians. Table 1 summarizes the three datasets.



(a)

(b)

Figure 5: Comparison of accuracies achieved by combining HOG and LBP features through FF1 and FF2 methods on the INRIA dataset. The classifiers are linear SVM and HIKSVM in (a) and (b), respectively.

Feature fusion performance is measured by plotting the number of false positives per window (FPPW) versus the miss rate, as proposed by Dalal and Triggs [4] in Section 3.2. This only measures classification performance and excludes nonmaximum suppression and other postprocessing steps. The FPPW miss rate curves are plotted in log-log space. To avoid sampling bias, negative samples are selected as a fixed set and are not boosted by bootstrap learning.

Pedestrian detection performance over cross-datasets is measured by precision and recall curves, described in Section 3.3. This is a measure of both classification and location performance, including nonmaximum suppression and various postprocessing steps. Efficiency and robustness to occlusion of the proposed method are also discussed in Sections 3.4.

3.2. Performance Evaluation of Feature Fusion. As mentioned in Section 2.3, the two main feature fusion methods are FF1 and FF2. This experiment was conducted to compare the performances of FF1 and FF2. Both the linear SVM and HIKSVM are applied on the INRIA dataset. HIKSVM is approximated as 20 linear segments with a piecewise constant function. Experimental results show that the FF1 method outperforms FF2 for all of the values of α and FF2 has the best performance when $\alpha = 0.5$ (Figure 5). Therefore, FF1 is selected by default for feature fusion hereafter.

The experimental results show that combining HOG and LBP features through the FF1 method using the HIKSVM classifier yields the best performance. Figure 6 shows a comparison of the results obtained by applying combined features (single features or combining HOG and LBP features) and different SVMs (HIKSVM or linear SVM) on the INRIA and XMU-VIS datasets, respectively. Figure 6(a) shows that applying feature HOG to the INRIA dataset is better than applying feature LBP. In contrast, Figure 7(b) shows that applying feature LBP is better than applying HOG on the XMU-VIS dataset. The HIKSVM outperforms the linear SVM, regardless of the features used. Combining HOG and LBP features through the FF1 method with HIKSVM as a classifier yields the best performance, regardless of the INRIA or XMU-VIS datasets. Therefore, the proposed method fuses HOG and LBP features through the FF1 method and uses the HIKSVM as a classifier. The method is then applied to test images using the sliding window strategy to evaluate pedestrian detection performance over cross-datasets in Section 3.3.

3.3. Performance Evaluation of Pedestrian Detection over Cross-Datasets. As shown in [2], the per-window measure for pedestrian classification is flawed and fails to predict full image performance for pedestrian detection. Therefore, the proposed method is also evaluated on full images using the PASCAL criteria in this section. The details are described as follows. The proposed pedestrian detection, fusing HOG and LBP features through the FF1 method with the HIKSVM as a classifier, is used to find pedestrians in an image by scanning the entire image with a fixed size rectangle. A denoted window, labeled as a rectangle in Figure 7, presents the framework of the proposed HIKSVM-based pedestrian detection with sliding window scanning on full images, called sliding window scanning. Various sized windows are scanned to detect multiscale humans.

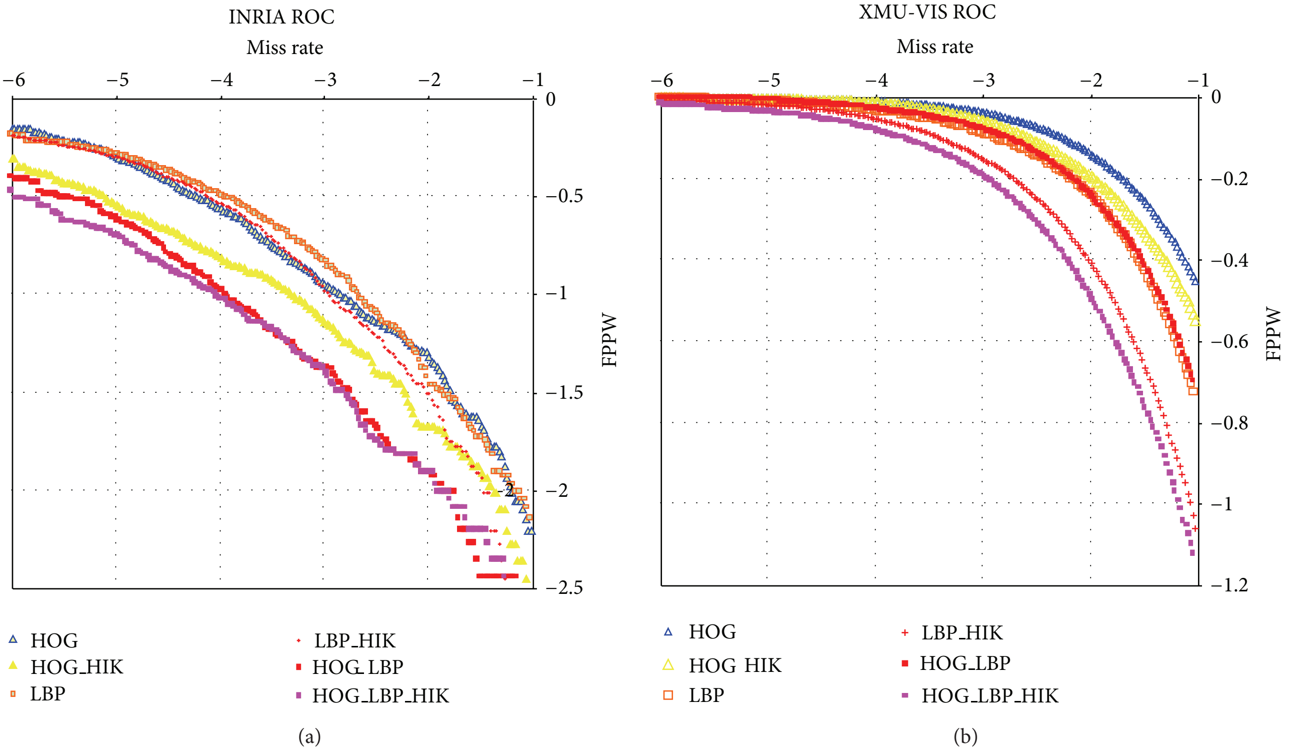


Figure 6: Comparing accuracies achieved by applying single features (HOG and LBP) and fusing features (HOG + LBP) using HIKSVM and FF1 to INRIA and XMU-VIS datasets.

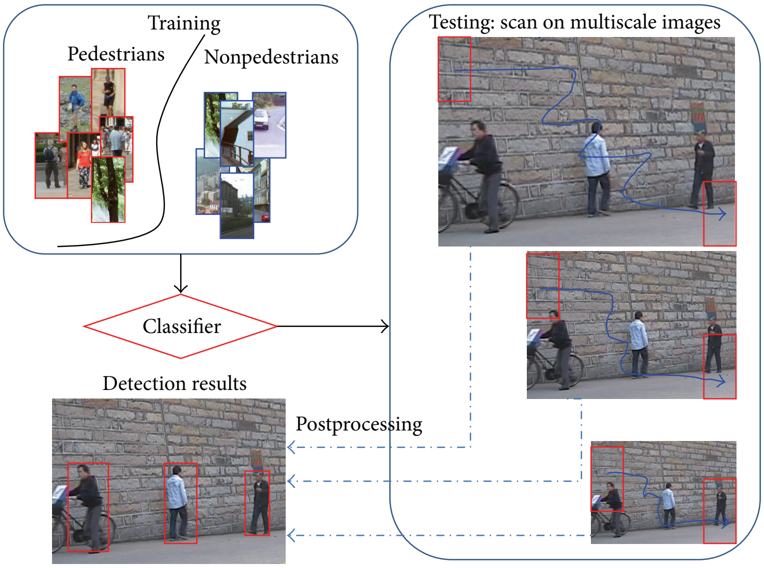
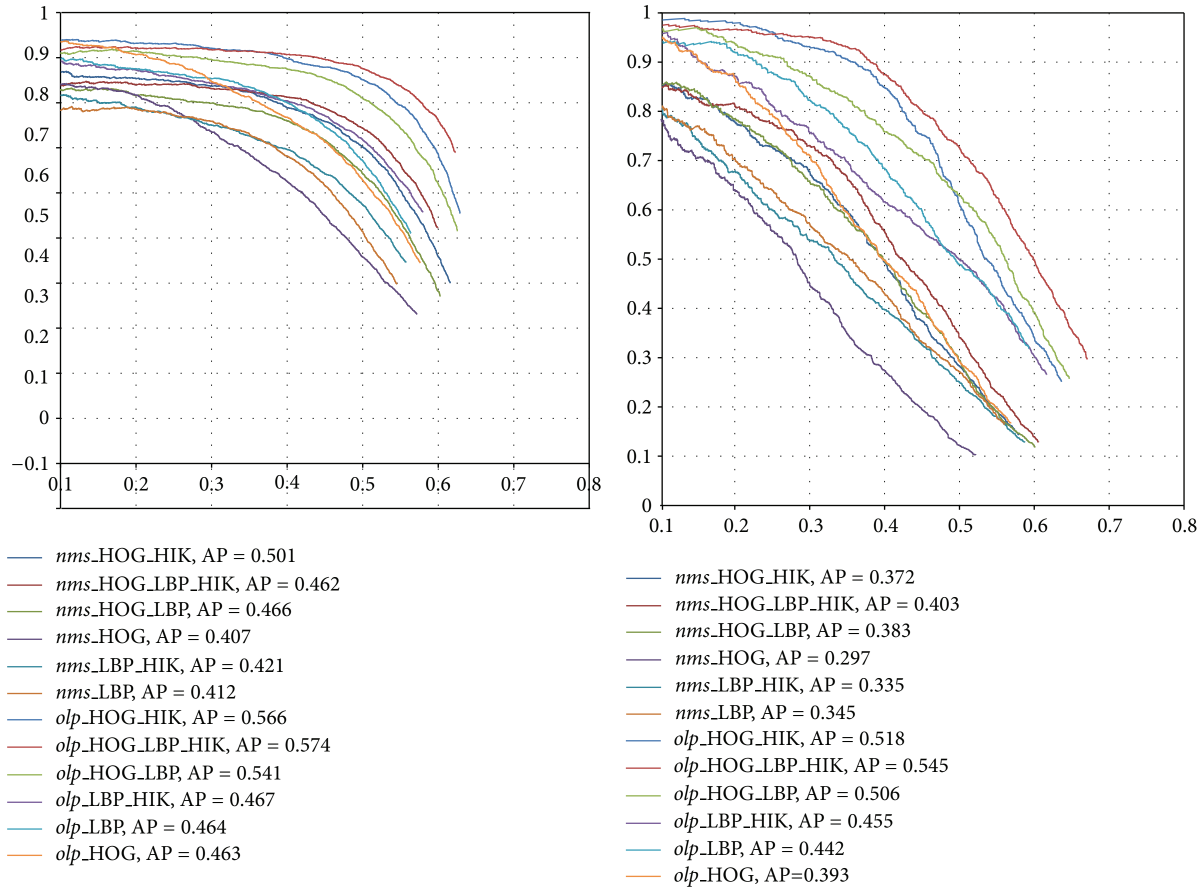


Figure 7: Framework of the proposed HIKSVM-based pedestrian detection with sliding window scanning.

The local block features of each window are fed to the HIKSVM-based pedestrian classifier to determine whether a human exists in the window. Windows determining whether a human exists are considered as candidate windows. After performing multiscale sliding window scanning, candidate windows of various sizes may overlap each other, specifically surrounding authentic humans. Overlapping windows should be postprocessed to locate humans with an accurate position. Two typical postprocessing methods, mean-shift location and window overlapping handling, denoted by nms and olp, respectively, are used and compared to determine the proper postprocessing methods. Experimental results show that the proposed pedestrian detection, fusing HOG and LBP features through the FF1 method with the HIKSVM classifier and window overlap postprocessing, is superior (Figure 8).



(a)

(b)

Figure 8: Comparison of precision and recalls achieved by using single features (HOG and LBP) and fusing HOG and LBP through various postprocessing: (a) XMU-VIS; (b) XMU_NIR.

3.4. Robustness Evaluation to Partial Occlusion. An experiment was conducted to show that the proposed method by concatenating HOG and LBP features through the FF1 method with an HIKSVM classifier is typically robust to partial occlusion (Figures 9 and 10). The experiment was designed to randomly add one to five random blocks of size 16 × 16 to the 1126 test-cropped images of pedestrians in the INRIA dataset (Figure 11). Figure 11 shows that when random blocks are added to the test-cropped images, the number of missed pedestrians increases, regardless of the features and SVMs used. The number of missed pedestrians increases when more random blocks are added. Figure 12 shows that the number of missing pedestrians for HOG and LBP is lower than that when using a single feature, regardless of the SVM that is used. In this experiment, a test sample is considered to include a pedestrian when the SVM output score exceeds 0.5.

4. Conclusion

This work systematically compares integral channel features, fusion methods, and kernels of SVM. The experimental results show that fusing HOG and LBP features through concatenation with the HIKSVM classifier yields the best performance, even for cross-datasets. The comparison is conducted using the INRIA person dataset for training and two newly collected Xiamen databases, XMU-VIS and XMU-NIR, combined with INRIA for testing. The results are as follows. First, directly concatenating various features as the final feature for classification is better than the weighted fusion of individual classifier results. Second, combining HOG and LBP features outperforms using a single feature, regardless of whether HIKSVM or linear SVM is used. As to kernel mapping, there are also some non-linear kernels [22], such as RBF and Chi2 kernel, which have reported obtaining better performance than HIK. But non-linear kernels are time-consuming in testing state; so, in this paper, we only discuss the linear kernels for pedestrian detection. Third, HIKSVM consistently outperforms linear SVM, even when noise blocks are added that cause the occlusion problem. Fourth, for the postprocessing method, window-overlap-based postprocessing outperforms the mean-shift-based postprocessing. Finally, the proposed method is effective to detect pedestrian locations, even for cross-datasets collected in Xiamen University and captured by low-resolution visual sensors or near-infrared sensors. However, the method proposed in this work has certain limitations. Therefore, future works should extend

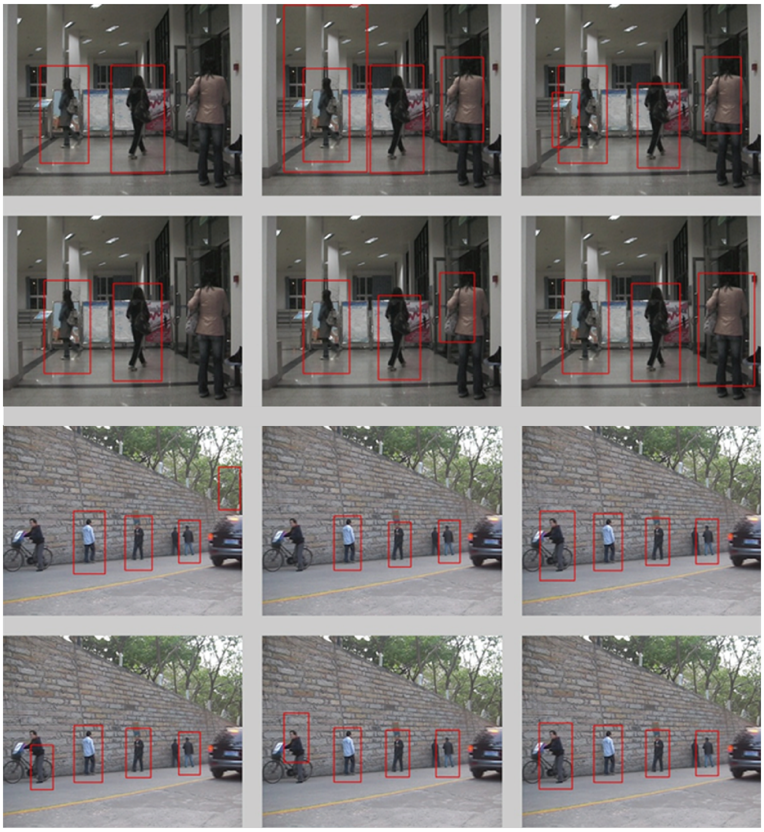


Figure 9: Detection results on XMU-VIS. From left to right and top to down, the classifiers are HOG_lin, LBP_lin, HOG_LBP_lin, HOG_HIK, LBP_HIK, and HOG_LBP_HIK.

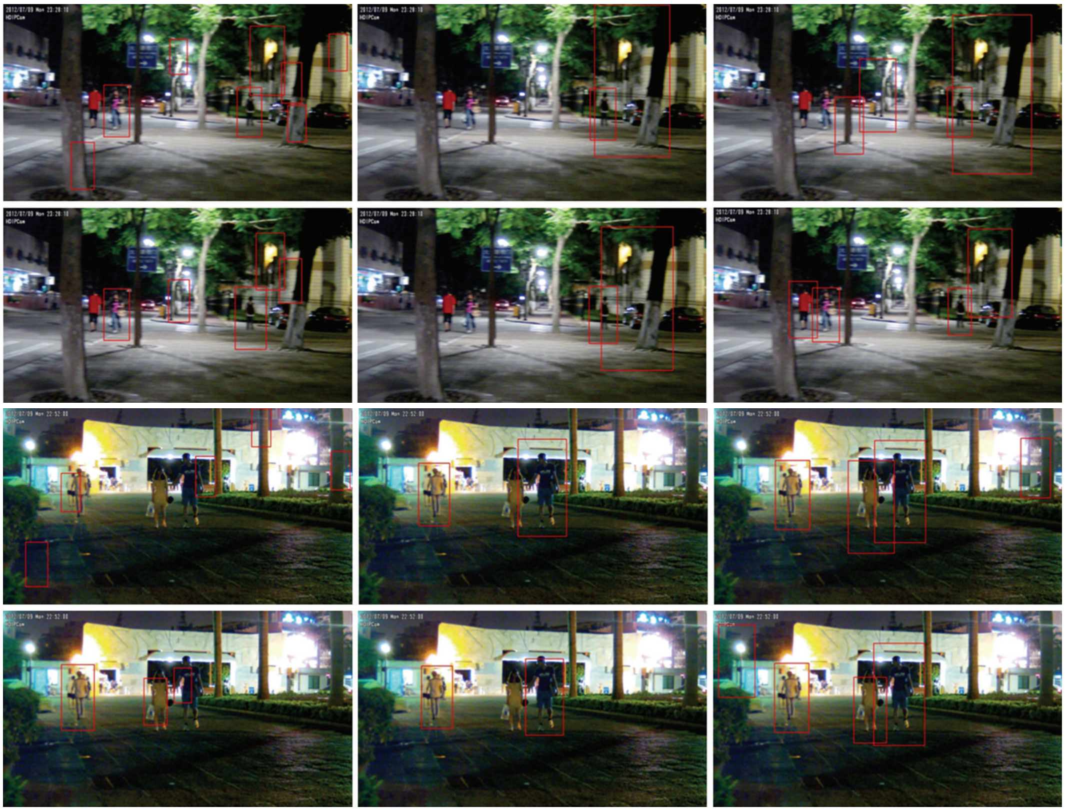


Figure 10: Detection results on XMU-NIR. From left to right and top to down, the classifiers are HOG_lin, LBP_lin, HOG_LBP_lin, HOG_HIK, LBP_HIK, and HOG_LBP_HIK.

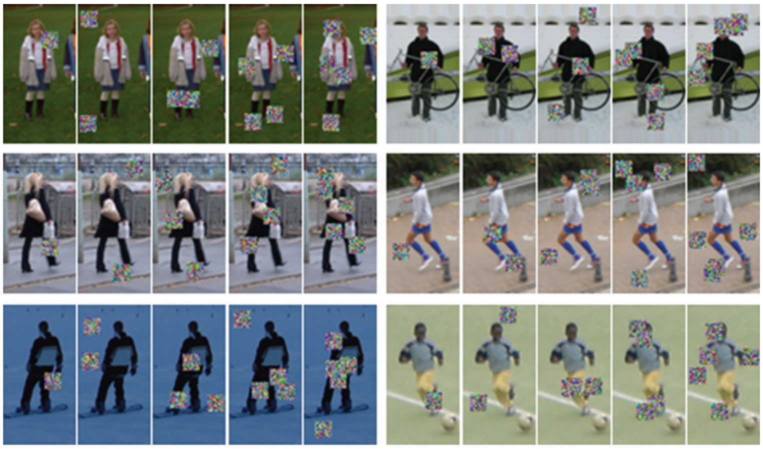


Figure 11: Examples of adding random blocks of size 16 × 16 to test-cropped images in the INRIA dataset.

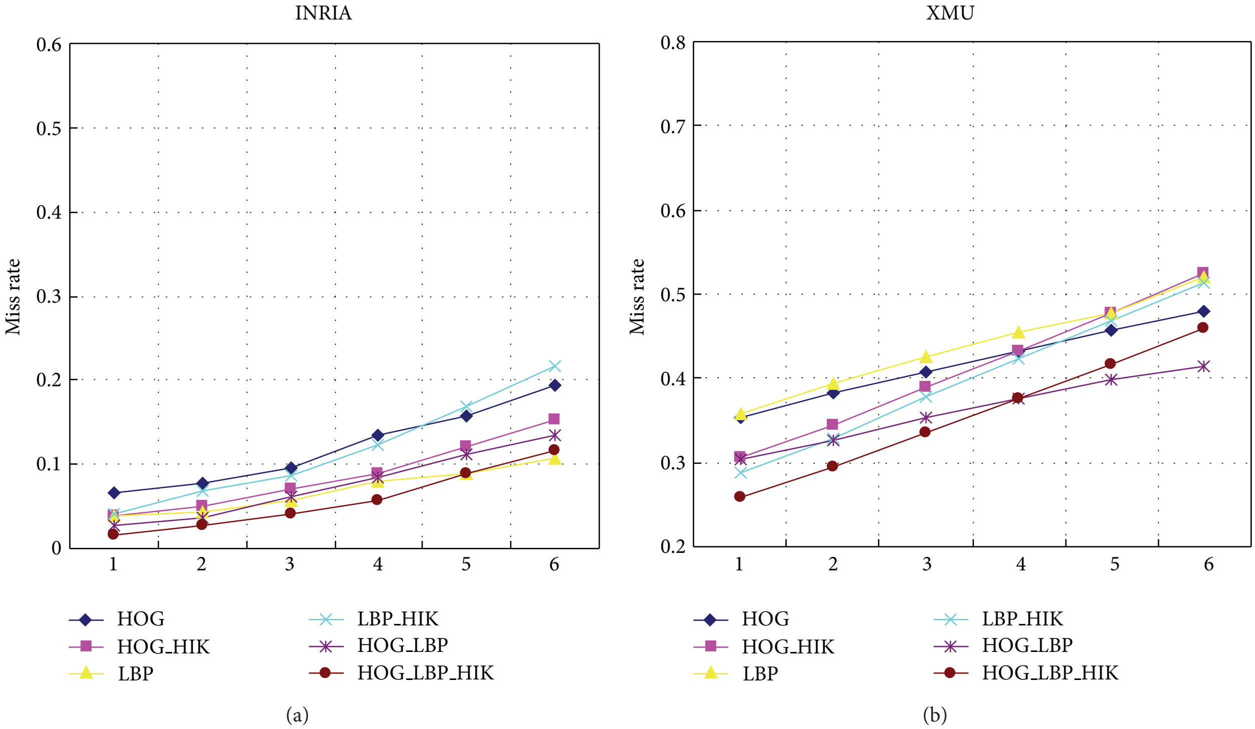


Figure 12: Comparison of missing rates using various combinations of features (single features or feature fusion) when random blocks were added to test-cropped images.

the proposed method to construct a practical pedestrian detection system for videos that integrates additional motion features and scene geometry information.

Acknowledgments

The authors wish to thank the Intel Open Source Computer Vision Library (OpenCV), fiksvm [16], and libSVM [23], for making their implementation convenient. This work was supported by the National Natural Science Foundation of China (no. 61202143) and the National Science Council of Taiwan (no. NSC 101-2221-E-155-060).

References

[1] D. Gerónimo, A. M. López, A. D. Sappa, and T. Graf, “Survey of pedestrian detection for advanced driver assistance systems,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 7, pp. 1239–1258, 2010.

[2] P. Dollar, C. Wojek, B. Schiele, and P. Perona, “Pedestrian detection: an evaluation of the state of the art,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 4, pp. 743–761, 2012.

[3] P. Viola and M. J. Jones, “Robust real-time face detection,” *International Journal of Computer Vision*, vol. 57, no. 2, pp. 137–154, 2004.

[4] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '05)*, pp. 886–893, San Diego, Calif, USA, June 2005.

[5] A. Mohan, C. Papageorgiou, and T. Poggio, "Example-based object detection in images by components," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 4, pp. 349–361, 2001.

[6] B. Wu and R. Nevatia, "Detection and tracking of multiple, partially occluded humans by Bayesian combination of edgelet based part detectors," *International Journal of Computer Vision*, vol. 75, no. 2, pp. 247–266, 2007.

[7] B. Leibe, A. Leonardis, and B. Schiele, "Robust object detection with interleaved categorization and segmentation," *International Journal of Computer Vision*, vol. 77, no. 1–3, pp. 259–289, 2008.

[8] L. Zhao and C. E. Thorpe, "Stereo- and neural network-based pedestrian detection," *IEEE Transactions on Intelligent Transportation Systems*, vol. 1, no. 3, pp. 148–154, 2000.

[9] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[10] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1627–1645, 2010.

[11] Q. Zhu, M. C. Yeh, K. T. Cheng, and S. Avidan, "Fast human detection using a cascade of histograms of oriented gradients," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '06)*, vol. 2, pp. 1491–1498, June 2006.

[12] O. Tuzel, F. Porikli, and P. Meer, "Pedestrian detection via classification on Riemannian manifolds," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 10, pp. 1713–1727, 2008.

[13] C. Wojek, S. Walk, and B. Schiele, "Multi-Cue onboard pedestrian detection," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPR '09)*, pp. 794–801, June 2009.

[14] Y. T. Chen and C. S. Chen, "Fast human detection using a novel boosted cascading structure with meta stages," *IEEE Transactions on Image Processing*, vol. 17, no. 8, pp. 1452–1464, 2008.

[15] X. Wang, X. Han, and S. Yan, "A HOG-LBP human detector with partial occlusion handling," in *Proceedings of the IEEE 12th International Conference on Computer Vision (ICCV '09)*, pp. 32–39, Kyoto, Japan, October 2009.

[16] S. Maji, A. C. Berg, and J. Maliks, "Classification using intersection kernel support vector machines is efficient," in *Proceedings of the 26th IEEE Conference on Computer Vision and Pattern Recognition (CVPR '08)*, pp. 1–8, Anchorage, Alaska, USA, June 2008.

[17] J. X. Wu, "Efficient HIK SVM learning for image classification," *IEEE Transactions on Image Processing*, vol. 21, no. 10, pp. 4442–4453, 2012.

[18] D. Vazquez, A. M. Lopez, and D. Ponsa, "Unsupervised domain adaptation of virtual and real worlds for pedestrian detection," in *Proceedings of the 21st International Conference on Pattern Recognition (ICPR '12)*, pp. 3492–3495, Tsukuba, Japan, 2012.

[19] V. Jain and E. Learned-Miller, "Online domain adaptation of a pre-trained cascade of classifiers," in *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR '11)*, 2011.

[20] T. Ahonen, A. Hadid, and M. Pietikäinen, "Face description with local binary patterns: application to face recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 12, pp. 2037–2041, 2006.

[21] M. J. Swain and D. H. Ballard, "Color indexing," *International Journal of Computer Vision*, vol. 7, no. 1, pp. 11–32, 1991.

[22] A. Vedaldi and A. Zisserman, "Efficient additive kernels via explicit feature maps," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 3, pp. 480–492, 2012.

[23] C. C. Chang and C. J. Lin, "LIBSVM: a library for support vector machines," 2001, http://www.csie.ntu.edu.tw/~cjlin/libsvm.