IPSJ Transactions on Computer Vision and Applications Vol. 2 39–47 (Mar. 2010)

Research Paper

Co-occurrence Histograms of Oriented Gradients for Human Detection

Tomoki Watanabe,[†1] Satoshi Ito[†1]
and Kentaro Yokoi[†1]

The purpose of the work reported in this paper is to detect humans from images. This paper proposes a method for extracting feature descriptors consisting of co-occurrence histograms of oriented gradients (CoHOG). Including co-occurrence with various positional offsets, the feature descriptors can express complex shapes of objects with local and global distributions of gradient orientations. Our method is evaluated with a simple linear classifier on two well-known human detection benchmark datasets: "*DaimlerChrysler pedestrian classification benchmark dataset*" and "*INRIA person data set*". The results show that our method reduces the miss rate by half compared with HOG, and outperforms the state-of-the-art methods on both datasets. Furthermore, as an example of a practical application, we applied our method to a surveillance video eight hours in length. The result shows that our method reduces false positives by half compared with HOG. In addition, CoHOG can be calculated 40% faster than HOG.

1. Introduction

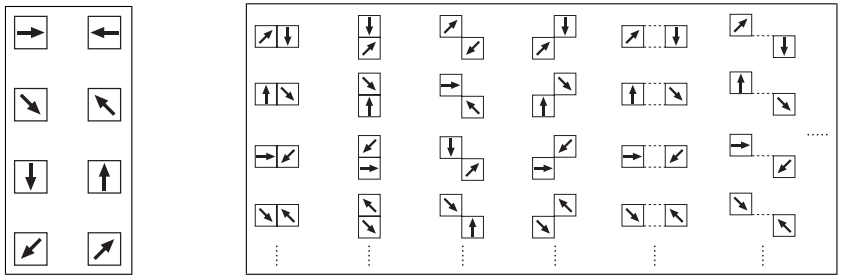
Detecting humans in images is essential in many applications such as automatic driver assistance, image surveillance, and image analysis. The extensive variety of postures and clothes of humans makes this problem challenging.

Many types of feature descriptors have been proposed for human detection. Gavrila, et al. proposed combining two feature descriptors [1]: templates of human contours with chamfer matching [2] and LRF (Local Receptive Fields) with a quadratic SVM classifier [3]. LRF are weight parameters of hidden layers of neural network that extract local features of humans. Viola, et al. proposed a motion feature descriptor and combined it with cascaded AdaBoost classifier [4]. Papageorgiou, et al. used SVM-based parts detectors with Haar wavelet feature and integrated them with SVM [5),6].

Recently, using gradient-orientation-based feature descriptors, such as SIFT (Scale Invariant Feature Transform) [7] and HOG (Histograms of Oriented Gradients) [8], is a trend in object detection [9),10]. Those feature descriptors are also used for human detection [8),11)–13]. Shashua, et al. employed body parts detectors using SIFT [11] and Mikolajczyk, et al. also used jointed SIFT with an SVM classifier [12]. Dalal, et al. proposed HOG and combined it with an SVM classifier [8], and also extended their method to motion feature descriptors [13].

Some multiple-edge-based feature descriptors also have been proposed. Wu, et al. proposed edgelet feature descriptor that expresses long curves of edges [14]. Sabzmeydani, et al. proposed shapelet feature descriptor based on edges selected by AdaBoost [15]. Since shapelets are combinations of edges, they can express more detailed shape information than SIFT/HOG feature descriptors can.

We propose a multiple-gradient-orientation-based feature descriptor named "Co-occurrence Histograms of Oriented Gradients (CoHOG)". CoHOG are histograms whose building blocks are pairs of gradient orientations. Since a pair of gradient orientations has more vocabulary than a single orientation as shown in **Fig. 1**, CoHOG can express shapes in more detail than HOG, which uses single gradient orientation. Benchmark results on two well-known datasets, namely, DaimlerChrysler pedestrian classification benchmark dataset and INRIA person data set, show the effectiveness of our method. Furthermore, as an example of a



(a) Single (b) Paired

Fig. 1 Vocabulary of gradient orientations. Though (a) a single gradient orientation has only eight varieties, (b) a pair of them has many more varieties than the single one.

†1 Corporate Research and Development Center, TOSHIBA Corporation

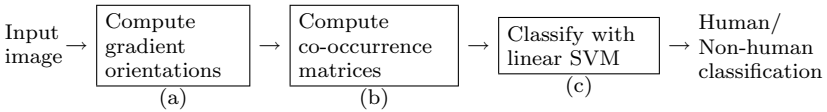


Fig. 2 Our classification process. We combine the strong feature descriptor CoHOG and a conventional simple classifier. Our classification process consists of three parts: (a) computation of gradient orientations from input images, (b) computation of CoHOG from gradient orientations, and (c) classification with linear SVM classifier that is fast at learning and classification.

practical application, we used CoHOG to detect humans from a long surveillance video.

The rest of this paper is organized as follows: Section 2 explains the outline of our human detection approach; Section 3 briefly explains HOG, and then describes our feature descriptor; Section 4 shows experimental results for two benchmark datasets and a surveillance video; the final section is the conclusion.

2. Outline of Our Approach

In most human detection tasks, classification accuracy is the most important requirement. The performance of the system depends on the effectiveness of feature descriptors and the accuracy of classification models.

In this paper, we focus on the feature descriptor. An overview of our human detection process is shown in **Fig. 2**. The first two parts of the process extract feature descriptors from input images, and then the last part classifies and outputs classification results. We propose a high-dimensional feature descriptor in Section 3. Our feature descriptor is effective for classification, because it contains building blocks that have an extensive vocabulary.

If the feature descriptor is informative enough, a simple linear classifier can detect humans accurately. We use a linear classifier obtained by a linear SVM [16] that works fast at learning and classification.

3. Gradient-orientation-based Feature Descriptor

3.1 Histograms of Oriented Gradients (HOG)

We briefly explain the essence of the HOG calculation process with **Fig. 3**. In order to extract HOG from an image, firstly, gradient orientations at every pixel are calculated (Fig. 3 (a)). Secondly, a histogram of each orientation in a small rectangular region is calculated (Fig. 3 (b)). Finally, the HOG feature vector is created by concatenating the histograms of all small regions (Fig. 3 (c)).

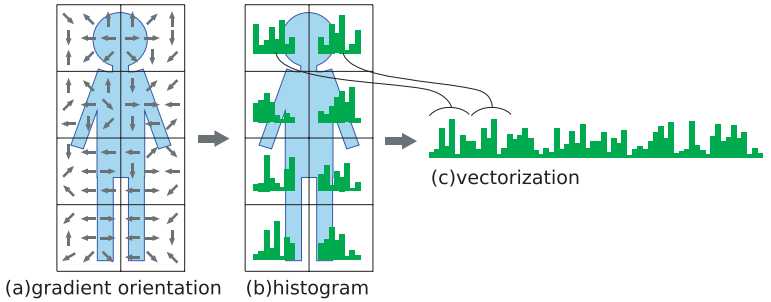


Fig. 3 Overview of HOG calculation.

HOG has two merits for human detection. One merit is the robustness against illumination variance because gradient orientations of local regions do not change with illumination variance. The other merit is the robustness against deformations because slight shifts and affine deformations make small histogram value changes.

3.2 Co-occurrence Histograms of Oriented Gradients (CoHOG)

We propose a high-dimensional feature "Co-occurrence Histograms of Oriented Gradients (CoHOG)". Our feature uses pairs of gradient orientations as units[⋆1], from which it builds the histograms. The histogram is referred to as the co-occurrence matrix, hereafter. The co-occurrence matrix expresses the distribution of gradient orientations at a given offset over an image as shown in **Fig. 4**. The combinations of neighbor gradient orientations can express shapes in detail. It is informative for human classification. Mathematically, a co-occurrence

⋆1 CoHOG can be defined with not only a pair of gradient orientations but also a set of multiple gradient orientations, in general. In this paper, we only explain the case of using two gradient orientations, because it is sufficient to contrast CoHOG with HOG and is easy to understand. In a similar way, we can use a set of multiple gradient orientations as a feature descriptor.

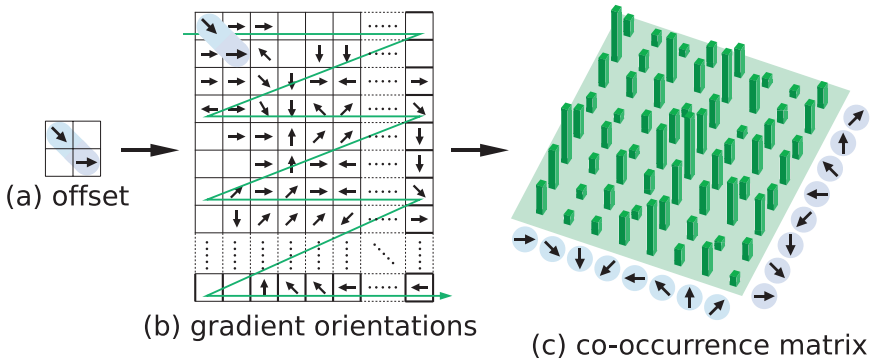


Fig. 4 Co-occurrence matrix of gradient orientations. It calculates sums of all pairs of gradient orientations at a given offset.

matrix C is defined over an $n \times m$ image, parameterized by an offset (x, y), as:

$$C_{x,y}(i,j) = \sum_{p=1}^{n} \sum_{q=1}^{m} \begin{cases} 1, & \text{if } I(p,q) = i \text{ and } I(p+x, q+y) = j \\ 0, & \text{otherwise,} \end{cases} \quad (1)$$

where I denotes a gradient orientation image, and i and j denote gradient orientations. CoHOG has robustness against deformation and illumination variance for the same reasons as HOG, because CoHOG is a gradient-based histogram feature descriptor.

We describe the process of CoHOG calculation shown in **Fig. 5**. Firstly, we compute gradient orientations from an image by

$$\theta = \arctan \frac{v}{h}, \quad (2)$$

where v and h are vertical and horizontal gradients, respectively, calculated by Sobel filter, Roberts filter, etc. We label each pixel with one of eight discrete orientations or as no-gradient (Fig. 5 (a)). All 0° – 360° orientations are divided into eight orientations per 45°. No-gradient means $\sqrt{v^2 + h^2}$ is smaller than a threshold. Secondly, we compute co-occurrence matrices by Eq. (1) (Fig. 5 (b)). The offsets we used are shown in **Fig. 6**. By using short-range and long-range offsets, the co-occurrence matrices can express local and global shapes. We do not use half of the offsets, because they behave the same as the others in calculation of co-occurrence matrix as shown in **Fig. 7**. The dashed-circle is the maximum range of offsets. We can get 31 offsets including a zero offset. The co-occurrence matrices are computed for each small region (Fig. 5 (c)). The small rectangular regions are tiled $N \times M$, such as 3×6 or 6×12, with no overlapping. Finally, the components of all the co-occurrence matrices are concatenated into a vector (Fig. 5 (d)).

Since CoHOG expresses shapes in detail, it is high-dimensional. The dimension

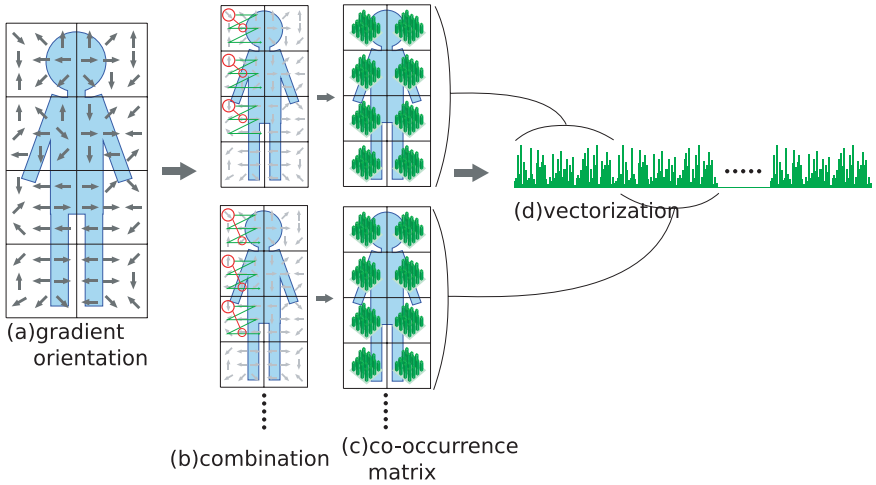


Fig. 5 Overview of CoHOG calculation.

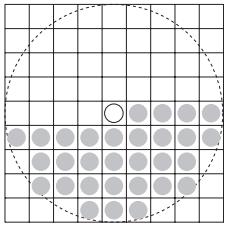


Fig. 6 Offsets of co-occurrence matrices. Offsets are smaller than the large dashed-circle. The center small white-circle and the other 30 dark-circles are paired. We calculate 31 co-occurrence matrices with different offsets including zero offset.

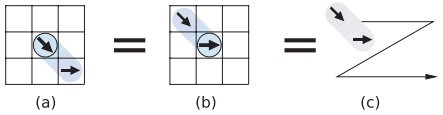


Fig. 7 Offset value of (a) $(1, 1)$ is equivalent to that of (b) $(-1, -1)$ in the calculation of co-occurrence matrix.

```
given I: an image of gradient orientation
initialize H ← 0
for all positions (p, q) inside of the image do
    i ← I(p, q)
    k ← the small region including (p, q)
    for all offsets (x, y) such that corresponds neighbors do
        if (p + x, q + y) is inside of the image then
            j ← I(p + x, q + y)
            H(k, i, j, x, y) ← H(k, i, j, x, y) + 1
        end if
    end for
end for
```

Fig. 8 Implementation of CoHOG calculation. The bins of histogram H are initialized to zero before voting. All pixels in the gradient orientation image I are scanned, and bins of H corresponding to pixels are incremented.

is 34,704, when the small regions are tiled 3×6. From one small region, CoHOG obtains 31 co-occurrence matrices. Each co-occurrence matrix has 64 components (Fig. 4 (c)). The co-occurrence matrix calculated with zero offset has only eight effective values because off-diagonal components are zero. Thus CoHOG obtains $(64 \times 30 + 8) \times (3 \times 6) = 34{,}704$ components from an image. In fact, the effective values are fewer than 34,704, because co-occurrence matrices have multiple zero-valued components. Nevertheless, CoHOG is a more powerful feature descriptor than HOG because CoHOG has more effective values than HOG.

The implementation of CoHOG is simple. An example of CoHOG implementation is shown in **Fig. 8**. We can calculate CoHOG by only iterating to increment the components of co-occurrence matrices, whereas HOG calculation includes more procedures, such as orientation weighted voting, histogram normalization, and region overlapping. CoHOG can achieve high performance without those complex procedures.

4. Experimental Results

We evaluated our method by two experiments. First, we compared the performance of our method with the state-of-the-art methods by using benchmark datasets. Second, we applied our method to a practical application that detects humans from a surveillance video.

4.1 Benchmark Datasets

We evaluated the performance of CoHOG by applying our method to two human image datasets, namely, the DiamlerChrysler dataset [3] and the INRIA dataset [8], which are widely used as human detection benchmark datasets. The DaimlerChrysler dataset contains human images and non-human images cropped to 18×36 pixels. The INRIA dataset contains human images cropped to 64×128 pixels and non-human images of various sizes. The details of those datasets are shown in **Table 1**, and some samples of the datasets are shown in **Fig. 9**.

Because the images differe in size, in our method we divided the DiamlerChrysler dataset images into 3×6 small regions, and the INRIA dataset images into 6×12 small regions. Thus, the dimensions of our features are 34,704 on the DiamlerChrysler dataset, and quadruple that on the INRIA dataset. We used Roberts filter, the filter size of which is 2×2, on the DiamlerChrysler dataset and Sobel filter, the filter size of which is 3×3, on the INRIA dataset, because the image size of the DiamlerChrysler dataset is small. We used a linear SVM classifier trained with LIBLINEAR [17] that solves linear SVM learning problems much faster than previous solvers such as LIBSVM [18] and SVMLight [19].

We compared our method with five previous methods [3),8),15),20),21)]. All the methods use different features and classifiers: Dalal, et al. used HOG and RBF kernel SVM and linear SVM [8]; Gavrila, et al. used local receptive fields (LRF) and quadratic SVM [3]; Dollar, et al. used Haar wavelet and AdaBoost [20]; Sabzmeydani, et al. used shapelet and AdaBoost [15]; and Maji, et al. used multi-level oriented edge energy features and intersection kernel SVM (IKSVM) [21].

The comparison of their performances is shown in **Fig. 10**. The results of pre-

Table 1 Human detection benchmark datasets.

(a) DaimlerChrysler dataset

Dataset Name	DaimlerChrysler Pedestrian Classification Benchmark Dataset
Distribution site	http://www.science.uva.nl/research/isla/downloads/pedestrians/
Training data	4,800 × 3 human images 5,000 × 3 non-human images
Test data	4,800 × 2 human images 5,000 × 2 non-human images
Image size	18 × 36 pixels

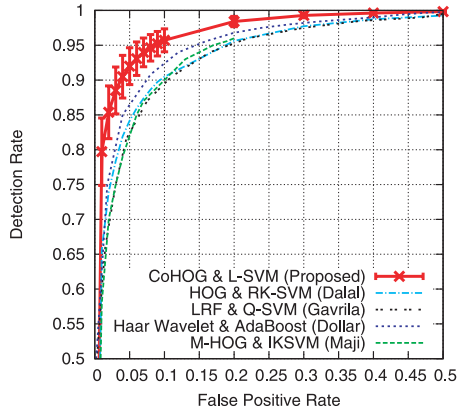
(b) INRIA dataset

Dataset Name	INRIA Person Data Set
Distribution site	http://pascal.inrialpes.fr/data/human/
Training data	2,716 human images 1,218 non-human images (10 regions are randomly sampled per image for training.)
Test data	1,132 human images 453 non-human images
Image size	Human images are 64 × 128 pixels Non-human images are of various size (214 × 320 – 648 × 486 pixels)

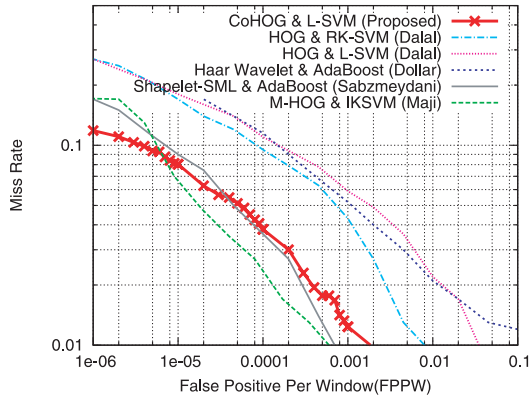
(a) DaimlerChrysler dataset

(b) INRIA dataset

Fig. 9 Thumbnails of (a) DaimlerChrysler dataset and (b) INRIA dataset. Upper rows are images of humans and lower rows are images of non-humans in each dataset.



(a)



(b)

Fig. 10 Performance of our method on (a) DaimlerChrysler dataset and (b) INRIA dataset. We compared our method with several previous methods. On the DaimlerChrysler dataset, our method shows the best performance. Our method reduces the miss rate 40% compared with the state-of-the-art method at a false positive rate of 0.05. On the INRIA dataset, our method decreases the miss rate by 30% compared with that of the state-of-the-art method at a FPPW of 10^{-6}. Our method reduces the miss rate by half compared with HOG on both datasets.

vious methods are traced from the original papers except the performance of HOG on the DaimlerChrysler dataset, because it is not shown by Dalal, et al. We show it based on the result of our experiment. The parameters of HOG are

as follows: nine gradient orientations in 0°–180°, cell size of 3×3 pixels, block size of 2×2 cells, L2Hys normalized, and the classifier is an RBF-kernel SVM. In Fig. 10 (a), ROC (Receiver Operating Characteristic) curves on the Daimler-Chrysler dataset are shown. An ROC curve further toward the top-left of the diagram means better performance. The results show that our method achieved the best detection rate at every false positive rate. Our method reduced the miss rate (= 1 − detection rate) by about 40% from the state-of-the-art method at a false positive rate of 0.05; the miss rate of our method is 0.08 and that of Dollar, et al., the second best, is 0.14.

In Fig. 10 (b), DET (Detection Error Tradeoff) curves on the INRIA dataset are shown. A DET curve further toward the bottom-left of the diagram means better performance. The results show that the performance of our method is better than the state-of-the-art methods or at least comparable. On condition that FPPW is low, our method reduced the miss rate by about 30% compared with the state-of-the-art method; the miss rate of our method is 0.12 and that of Maji, et al. is 0.17 at a FPPW of 10^{-6}. On condition that FPPW is high, the miss rates of our method is comparable to that of Maji, et al.; the miss rate of our method is 0.017 and that of Maji, et al. is 0.01 at a FPPW of 5×10^{-4}. Even though the miss rate of our method is slightly higher than that of Maji, et al. the difference of the miss rates is only 0.007.

The performance at low FPPW is more important than that at high FPPW, because low FPPW is necessary for machine-aided monitoring systems that are a typical application field of human detection, such as drive assistance system for automobile [2] and highway-railway grade crossing monitoring system. The alarm of the drive assistance system is activated when a human stands in front of the automobile. If the system's alarm is unnecessarily activated often, the user will be irritated and turn off the system. Even though the miss rate is high at low FPPW, it can be recovered using multiple images taken in a short period of time.

Furthermore, they show the stability of our method; the performance of the method of Dollar, et al. is not good on the INRIA dataset and the method of Maji, et al. is not good on the DaimlerChrysler dataset, whereas, the performance of our method is consistently good on both datasets. Though our method uses a linear classifier that is simpler than an RBF-kernel SVM classifier used with HOG, the miss rate of our method is less than half that of HOG.

Table 2 Surveillance video.

Length	First day: 7 hours and 25 minutes Second day: 7 hours and 46 minutes
Frame rate	2 frames per minute
Image size	320×240 pixels
Training data	3,751 human images from the first day and 45,406 non-human images from another dataset. The image sizes of them are 42×42 pixels.
Test data	932 images from the second day that include 3,864 humans.

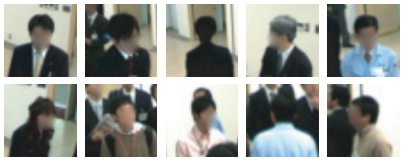


Fig. 11 Positive samples.

4.2 Surveillance Video

We detected humans from a surveillance video taken by a camera hung from the ceiling. The specification of the video is shown in **Table 2**. We used two videos for training data and test data that were taken on different days. In this experiment, we do not use the whole body of a human but use only the upper body to detect humans, because a lower body of the human who is near the camera is invisible. The ground truths that are given by hand are shown in **Fig. 11**. (The faces in the images are masked in this publication for reasons of privacy. A non-masked version is used for the experiment.) The images used for training are regularized to 42×42 pixels. In calculation of CoHOG feature descriptor, we divided the images into 6×6 small regions, and thus the dimension of our feature is 69,408.

We defined correct/incorrect detection by using the distance $D_{r,t}$ between the detection result r and the ground truth t that is defined as

$$D_{r,t} = \frac{A(r \cap t)}{A(r \cup t)}, \tag{3}$$

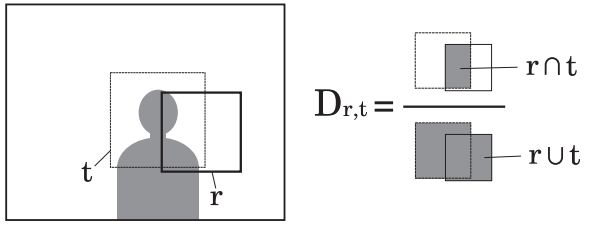


Fig. 12 Distance between detected result and ground truth.

where $A(\cdot)$ denotes the area (see **Fig. 12**). If $\exists t D_{r,t} \geq 1/3$, then the detection result is correct, otherwise the detection result is incorrect.

The result of human detection by our method is shown in **Fig. 13**. True positives, false positives, and false negatives are represented by red, yellow, and blue squares, respectively. As shown in Fig. 13, even when there are many humans in the image, almost all the humans are detected correctly (true positive); and undetected humans (false negative) and non-human regions detected as humans (false positive) are few.

We compared the performance of our method and HOG by using the surveillance video. The parameters of HOG are as follows: Nine gradient orientations in 0°–180°, cell size of 3×3 pixels, block size of 2×2 cells, L2Hys normalized, and the classifier is a linear SVM. Thus, the dimension of HOG is 6,084. The DET curves are shown in **Fig. 14**. At every miss rate, our method reduced FPPF (False Positives Per Frame) by half compared with HOG.

The processing times per frame are shown in **Table 3**. To detect humans in an image, 7,577 ROIs are processed. The minimum size of the ROI is 42×42 pixels. Our method is 40% faster than HOG, even though CoHOG is a higher-dimensional feature descriptor than HOG; because CoHOG is simple to calculate as mentioned in Section 3.2. The result means that a high-dimensional feature descriptor is not always disadvantageous in terms of processing time.

Several methods which reduce the processing time of HOG are proposed [22),23)], such as feature selection, classifier cascading, integral image, and feature pool. The speed-up methods can be utilized to reduce the processing time of CoHOG too, because the fundamental structure of HOG used by the speed-up methods is common to CoHOG.

Fig. 13 Detection results: Red, yellow, and blue squares represent true positive, false positive, and false negative, respectively.

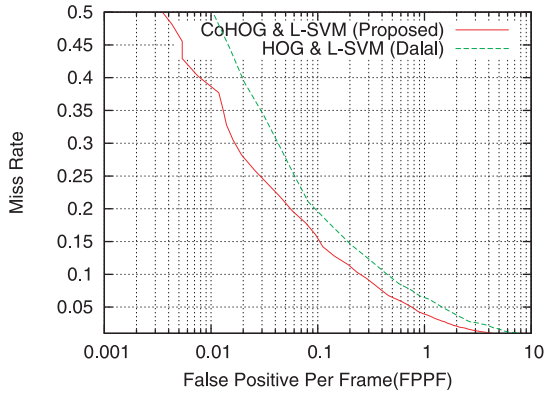


Fig. 14 Performance of human detection on surveillance video. Our method reduces FPPF by half compared with HOG at every miss rate.

Table 3 Processing time per a frame. Our method is 40% faster than HOG.

	Processing time
CoHOG	0.77 sec/frame
HOG	1.25 sec/frame

5. Conclusion

In this paper, we proposed a high-dimensional feature descriptor "Co-occurrence histograms of oriented gradients (CoHOG)" for human detection. Our feature descriptor uses pairs of gradient orientations as units, from which it builds histograms. Since the building blocks have an extensive vocabulary, our feature descriptor can express local and global shapes in detail. We compared the classification performance of our method and several previous methods on two well-known datasets. The experimental results show that the performance of our method is better than that of the state-of-the-art methods or at least comparable, and consistently good on both datasets. The miss rate (i.e., the rate of human images classified as non-human) of our method is less than half that of HOG. Furthermore, as an example of a practical application, we applied our method to a surveillance video eight hours in length. The result shows that our method reduces false positives by half compared with HOG. In addition, CoHOG can be calculated 40% faster than HOG.

References

1) Gavrila, D.M. and Munder, S.: Multi-cue Pedestrian Detection and Tracking from a Moving Vehicle, *Int. J. Comput. Vision*, Vol.73, No.1, pp.41–59 (2007).
2) Gavrila, D. and Philomin, V.: Real-Time Object Detection for "Smart" Vehicles, *7th IEEE International Conference on Computer Vision*, Vol.1, Los Alamitos, CA, USA, pp.87–93, IEEE Computer Society (1999).
3) Munder, S. and Gavrila, D.M.: An Experimental Study on Pedestrian Classification, *IEEE Trans. Pattern Anal. Mach. Intell.*, Vol.28, No.11, pp.1863–1868 (2006).
4) Viola, P., Jones, M.J. and Snow, D.: Detecting Pedestrians Using Patterns of Motion and Appearance, *9th IEEE International Conference on Computer Vision*, Washington, DC, USA, pp.734–741, IEEE Computer Society (2003).
5) Mohan, A., Papageorgiou, C. and Poggio, T.: Example-Based Object Detection in Images by Components, *IEEE Trans. Pattern Anal. Mach. Intell.*, Vol.23, No.4, pp.349–361 (2001).
6) Papageorgiou, C. and Poggio, T.: A Trainable System for Object Detection, *Int. J. Comput. Vision*, Vol.38, No.1, pp.15–33 (2000).
7) Lowe, D.G.: Distinctive Image Features from Scale-Invariant Keypoints, *Int. J. Comput. Vision*, Vol.60, No.2, pp.91–110 (2004).
8) Dalal, N. and Triggs, B.: Histograms of Oriented Gradients for Human Detection, *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Vol.1, pp.886–893 (2005).
9) Mikolajczyk, K. and Schmid, C.: A performance evaluation of local descriptors, *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp.257–263 (2003).
10) Winder, S.A.J. and Brown, M.: Learning Local Image Descriptors, *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp.1–8 (2007).
11) Shashua, A., Gdalyahu, Y. and Hayun, G.: Pedestrian detection for driving assistance systems: single-frame classification and system level performance, *IEEE Intelligent Vehicles Symposium*, pp.1–6 (2004).
12) Mikolajczyk, K., Schmid, C. and Zisserman, A.: Human Detection Based on a Probabilistic Assembly of Robust Part Detectors, *European Conference on Computer Vision*, Vol.1, pp.69–82 (2004).
13) Dalal, N., Triggs, B. and Schmid, C.: Human Detection Using Oriented Histograms of Flow and Appearance, *European Conference on Computer Vision*, pp.428–441 (2006).
14) Wu, B. and Nevatia, R.: Detection of Multiple, Partially Occluded Humans in a Single Image by Bayesian Combination of Edgelet Part Detectors, *10th IEEE Inter-*

national Conference on Computer Vision, Vol.1, Washington, DC, USA, pp.90–97, IEEE Computer Society (2005).
15) Sabzmeydani, P. and Mori, G.: Detecting Pedestrians by Learning Shapelet Features, *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp.1–8 (2007).
16) Cortes, C. and Vapnik, V.: Support-Vector Networks, *Mach. Learn.*, Vol.20, No.3, pp.273–297 (1995).
17) Hsieh, C., Chang, K., Lin, C., Keerthi, S. and Sundararajan, S.: A Dual Coordinate Descent Method for Large-scale Linear SVM, *the 25th Annual International Conference on Machine Learning*, McCallum, A. and Roweis, S. (eds.), pp.408–415, Omnipress (2008).
18) Hsu, C.W., Chang, C.C. and Lin, C.J.: A Practical Guide to Support Vector Classification, Technical report, Taipei (2003).
19) Joachims, T.: Training Linear SVMs in Linear Time, *the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp.217–226 (2006).
20) Dollar, P., Tu, Z., Tao, H. and Belongie, S.: Feature Mining for Image Classification, *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp.1–8 (2007).
21) Maji, S., Berg, A.C. and Malik, J.: Classification using Intersection Kernel Support Vector Machines is Efficient, *IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (2008).
22) Jia, H.-X. and Zhang, Y.-J.: Fast Human Detection by Boosting Histograms of Oriented Gradients, *Image and Graphics, International Conference on*, pp.683–688 (2007).
23) Zhu, Q., Yeh, M.-C., Cheng, K.-T. and Avidan, S.: Fast Human Detection Using a Cascade of Histograms of Oriented Gradients, *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp.1491–1498, IEEE Computer Society (2006).

(Received March 31, 2009)
(Accepted September 16, 2009)
(Released March 11, 2010)

(Communicated by *Fay Huang*)

Tomoki Watanabe was born in 1981. He received his B.E. and M.I. degrees from Hokkaido University in 2004 and 2006, respectively. He has been with the Corporate Research & Development Center of Toshiba Corporation since 2006. His current research interests are object detection, object recognition, and machine learning. He is a member of IEICE.

Satoshi Ito was born in 1982. He received his B.E. and M.I. degrees from the University of Tokyo in 2005 and 2007, respectively. He has been with the Corporate Research & Development Center of Toshiba Corporation since 2007. His current research interests are object detection, object recognition, and machine learning. He is a member of IEICE.

Kentaro Yokoi was born in 1973. He received his B.E. and M.E. degrees from Kyoto University in 1995 and 1997, respectively. He has been with Toshiba Corp. since 1997 and is a research scientist at the Corporate Research & Development Center. His current research interests are object detection, object tracking, and surveillance system. He is a member of IEICE.