Proceedings of the 4th World Congress on Electrical Engineering and Computer Systems and Sciences (EECSS'18)
Madrid, Spain – August 21 – 23, 2018
Paper No. MVML 100
DOI: 10.11159/mvml18.100

Pedestrian Detection based on Gaussian Mixture Model Multiresolution CoHOG

Shuto Higashi[1], Yuya Michishita[1], Shuichi Enokida[1], Masatoshi Shibata[2], Hideo Yamada[2]
[1]Kyushu Institute of Technology
Iiduka-shi, Fukuoka, Japan
[2]EQUOS RESEARCH Co., Ltd
Chiyoda-ku, Tokyo, Japan
higashi.shuto152@mail.kyutech.jp; michishita@mail.kyutech. jp; enokida@ai.kyutech.ac.jp

Abstract - Recently, Co-occurrence histograms of oriented gradients (CoHOG) describes image features to calculate the co-occurrence of pixels allocated at the local level and has attracted attention as an effective object detection method. However, this method has some problems. For feature descriptions that focus on individual pixels, calculation cost and the number of dimensions tend to increase exponentially with respect to the number of pixels. Multiresolution CoHOG (MRCoHOG) can suppress such exponential increases to linear increase without reducing the classification accuracy. This paper proposes a procedure in which a feature plane is divided using a Gaussian mixture model and a histogram is automatically divided to establish a less costly method for performing MRCoHOG. Experimental results demonstrate that the proposed procedure is more effective than conventional procedures.

***Keywords*:** People Detection, Jensen–Shannon Divergence, Gaussian Mixture model, Histograms of Oriented Gradients.

1. Introduction

Automatic object detection methods, particularly those based on deep learning [1], have attracted attention as important image processing techniques. In addition, handcrafted feature descriptors are important to realize compact sensing systems, such as ubiquitous sensing, which can be used in pedestrian recognition applications [2]. For example, a pedestrian detection system that utilizes dashcams to facilitate safe driving has been proposed in a previous study [3]. Generally, object detection algorithms comprise two phases. In the first phase, features are calculated from static images, while in the second phase, these features are used to recognize particular objects. The multiresolution co-occurrence histograms of oriented gradients (MRCoHOG) [4] utilizes a gradient histogram in a local area in a manner similar to that of other recent methods, for example, the histogram of oriented gradients (HOG), a descriptor proposed by Dalal and Triggs, is particularly efficient at detecting people [5], and Co-occurrence histograms of oriented gradients (CoHOG) [6] and feature interaction descriptor [7] describe high-dimensional features by calculating the co-occurrences of HOG features.

These co-occurrence features can express objects of complex shape by considering the relationship between the gradient orientations of pairs of pixels. However, calculating all co-occurrences involve vast numbers of combinations; thus, the following questions arise. Which pixels should be analyzed? Which computational algorithm should be utilized? Appropriate choices must be made because these factors influence classification accuracy. At the first of the method, CoHOG focuses on a pair of pixels with various offsets from various local regions in the image. The offsets comprise two pixels in a semicircle with a radius of four pixels (30 possible combinations). CoHOG can extract co-occurrence information over a wide region by expanding the radius of the semicircle; however, as the radius increases, the calculation cost and number of feature dimensions increase exponentially. Generally, in-vehicle and surveillance cameras used in object detection cannot perform these calculations in a realistic period of time. Thus, calculation costs must be reduced to achieve object detection with simple hardware. Therefore, we propose a feature description method based on the Gaussian mixture model MRCoHOG (GMM-MRCoHOG) to reduce calculation costs

2. Related Work

2.1. Co-occurrence Histograms of Oriented Gradients

The CoHOG extract local-level information about the co-occurrence of gradient orientations between two pixels and can classify objects with similar distributions of gradient orientations; thus, it suffers fewer false detections than HOG.

2.1.1. Calculating the Magnitude and Orientation of a Gradient

A $w \times h$ [pixels] image is divided into blocks, and the magnitude and orientation of the gradient in each block is extracted. Let b_x [pixels] and b_y [pixels] be the horizontal and vertical dimensions of the block, respectively. The block number N_Blo is calculated as follows:

$$N\_blo = wb \times hb\left(wb = \frac{b_y}{b_x}, hb = \frac{h}{b_y}\right). \tag{1}$$

At pixel (x, y), the magnitudes of horizontal gradient $f_x(x, y)$ and vertical gradient $f_y(x, y)$ are as follows:

$$f_x(x, y) = L(x + 1, y) - L(x - 1, y), \tag{2}$$

$$f_y(x, y) = L(x, y + 1) - L(x, y - 1), \tag{3}$$

where $L(x, y)$ is brightness. Based on these magnitudes, the orientation of the gradient $\theta(x, y)$ at pixel (x, y) is calculated as follows:

$$\theta(x, y) = \arctan\frac{f_y(x, y)}{f_x(x, y)}. \tag{4}$$

Then, the gradient orientation θ is quantized into $N_\theta(= 8)$ directions, and the magnitude of the orientation $f_\theta(x, y)$ is calculated as follows:

$$\theta(x, y) = \sqrt{f_x(x, y)^2 + f_y(x, y)^2}. \tag{5}$$

2.1.2. Offset

The offset is calculated without redundancy for a pair of points, as shown in Fig. 1(a). The first point of interest is Xc, and the other point is selected from the semicircular region centered at Xc, as shown in Fig. 1(b). For example, there are 30 candidate points at coordinates 4 pixels away from Xc.

2.1.3. Vote on Co-Occurrence Matrix

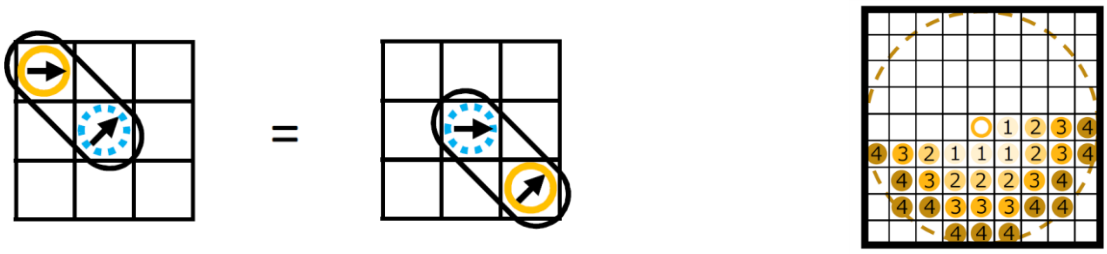
CoHOG can express complex shapes using a co-occurrence matrix. The co-occurrence matrix $C = (C_{x,y}(i, j))$ is calculated as follows:

$$\left(C_{x,y}(i, j)\right) = \sum_{p=1}^{w}\sum_{q=1}^{h}\begin{cases} 1, if\ \theta(p, q) = i \\ and\ \theta(p + x, q + y) = j \\ and\ f_\theta(p, q) > TH \\ \ and\ f_\theta(p + x, q + y) > \tau, \\ \\ 0, otherwise, \end{cases} \tag{6}$$

where i and j denote the orientations of the gradients and τ is a threshold gradient magnitude. Then, the magnitude and orientation of the gradient of zero-offset (Xc and Xc co-occurrence) is voted into another histogram with $N_\theta (= 8)$ bins. Then, the CoHOG is generated by concatenating all co-occurrence matrices into a vector, i.e., a high-dimensional feature. Figure 2 shows an overview of the CoHOG calculation. The dimension $Dim_{Co}$is calculated as follows:

$$Dim_{Co} = \left(N_{Mat} \times N_{Off_{Co}} + N_\theta\right) \times N\_Blo_{Co}, \tag{7}$$

where N_Mat denotes the number of bins in the co-occurrence matrix, N_Off_{Co} denotes the number of nonzero offsets, N_θ denotes the number of gradient orientations, and N_Blo_{Co} denotes the number of blocks. For example, when the region of interest is 30×60 pixels and the horizontal and vertical dimensions of the blocks are 10 and 10 pixels, respectively, the CoHOG vector comprises $\{(8 \times 8) \times 30 + 8\} \times 18 = 34{,}704$ dimensions.



(a) (b)

Fig. 1: (a) Redundant offset (right image) produces a result equivalent to another offset (left image). (b) Region of interest for calculating the co-occurrence matrix that describes a feature of the CoHOG [6].

2.2. Multiresolution Co-occurrence Histograms of Oriented Gradients

Typically, to achieve efficient and accurate object detection, a large semicircle must be considered to select the best candidate to extract the offset. However, the computational cost of the CoHOG increases exponentially with an increase in the semicircle's radius. In contrast, the computational cost of MRCoHOG increases linearly with an increase in the radius because the number of offsets increases linearly. In the following, we present the details of MRCoHOG with respect to describing features.

2.2.1. Region of Offset in MRCoHOG

MRCoHOG uses only adjacent pixels in multiresolution images as candidates to extract the offset, as shown in Fig. 2(a). MRCoHOG uses four adjacent pixels rather than using the offsets extracted from a semicircle centered on the original image. The process to obtain multiresolution images and offsets for MRCoHOG are shown in Fig. 2(b).

2.2.2. Relation between the Range of the Region of Offset and Calculation Cost

In CoHOG, the number of offsets increases exponentially as the range expands (e.g., 10, 18, and 30 offsets for ranges of 2, 3, and 4, respectively), and the calculation cost increases in proportion to the number of offsets. As mentioned previously, the number of offsets for MRCoHOG increases linearly with respect to the radius of the semicircle; thus, computation cost increases linearly. In addition, as the block size is invariant, low resolutions require few blocks, which reduces the calculation cost.

2.2.3. Dimension of MRCoHOG

The dimension $\mathrm{Dim_{MR}}$ of the MRCoHOG vector is calculated as follows:

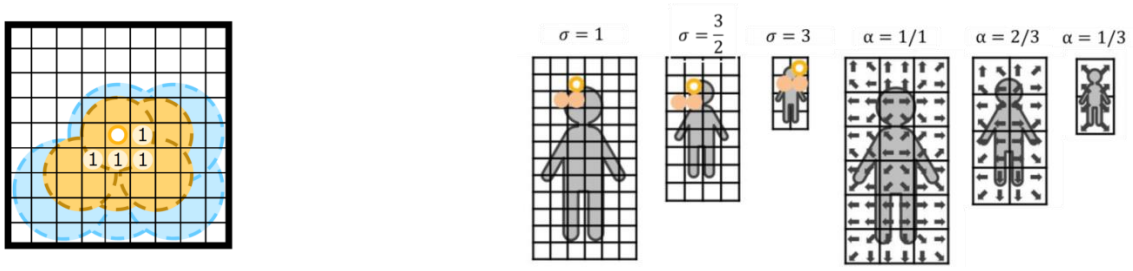
$$\mathrm{Dim_{MR}} = (N\_Mat \times N\_Off_{MR} + N\theta) \times N\_Blo_{MR,} \tag{8}$$

where N_Mat denotes the number of bins in the co-occurrence matrix, N_Off_{MR} denotes the number of nonzero offsets, $N\theta$ denotes the number of gradient orientations, and N_Blo_{MR} denotes the total number of blocks in all resolutions. For example, when the region of interest is 30×60 pixels and the horizontal and vertical dimensions of the block are 8 and 6 pixels,

respectively, the MRCoHOG vector comprises $\{(8 \times 8) \times 4 + 8\} \times (64 + 16 + 4) = 22{,}176$ dimensions. Note that this is ~80% of the dimensions of the CoHOG vector.

3. Proposed Method

With an MRCoHOG algorithm, features are described by histograms with uniform resolution that is determined by the number of bins in each block. Consequently, some bins remain unused; thus resulting in sparse feature values. This means that memory must be allocated for unused bins. This paper proposes GMM-MRCoHOG to determine the optimal width and location of the bins in each block automatically. With the proposed method, feature spaces of the MRCoHOG are approximated using the GMM. The optimal number of bins and their locations for feature description are generated automatically using the distributions of positive and negative samples. Compared with conventional MRCoHOG, describing features based on automatically generated bins reduces dimensionality without reducing recognition accuracy.



(a) (b)

Fig. 2: (a) Region of interest for calculating co-occurrence matrix used to describe a feature in MRCoHOG. (b) Resizing images in MRCoHOG.

3.1. Gaussian Mixture Model

When a probability density distribution is multicrestedness, approximation with normal distribution is difficult; therefore, multicrestedness distributions are approximated using the weighted linear sums of multiple normal distributions. This model is referred to as GMM and is expressed as follows:

$$p(x|\theta) = \sum_{j=1}^{K} \pi_j N\left(x|\mu_j, \Sigma_j\right), \tag{9}$$

$$d_M^2\left(x; \mu_j, \Sigma\right) = (x-\mu)^T \Sigma^{-1} (x-\mu), \text{ and}$$

$$N(x|\mu, \Sigma) = \frac{1}{(2\pi^{\frac{n}{2}})\sqrt{|\Sigma|}} \exp\left\{-\frac{1}{2} d_M^2(x; \mu, \Sigma)\right\}. \tag{10}$$

In Eqs. (9) and (10), x and μ are n-dimensional vectors, Σ represents an $n \times n$ variance-covariance matrix, and π represents the weight of each normal distribution. Thus, the GMM is expressed as the sum of all weighted normal distributions.

3.1.1. Training the GMM

Note that no specific procedure has been established to identify all parameters in a single round of calculations when training a GMM. We use the expectation-maximization (EM) algorithm, a maximum likelihood expectation method. An expectation is made for each parameter value, and the calculation is repeated until it approaches the optimal value. The latent variable $z \in 1, \ldots K$ is defined and employed as follows:

$$p(x|z, \theta) = N(x|\mu_z, \Sigma_z), \tag{11}$$

$$N(x|\mu,\Sigma) = \frac{1}{(2\pi^{\frac{n}{2}})\sqrt{|\Sigma|}} \exp\left\{-\frac{1}{2} d_M^2(x;\mu,\Sigma)\right\}. \tag{12}$$

The combined Gaussian function $p(x|z,\theta)$ is modified using Eqs. (11) and (12) as follows:

$$\begin{aligned} p(x|\theta) &= \sum_{j=1}^{K} p(x, z=j|\theta) \\ &= \sum_{j=1}^{K} p(x|z=j,\theta)p(z=j|\theta) \\ &= \sum_{j=1}^{K} \pi_j N\left(x\middle|\mu_j,\Sigma_j\right). \end{aligned} \tag{13}$$

From Eq. (13), the x distribution $p(x|\theta)$ can be expressed as a simultaneous distribution with z, i.e., a marginalized distribution with respect to z. Accordingly, using Bayes' theorem, this equation can be divided by the x distribution and prior distribution for z. Then, using Eqs. (11) and (12) and introducing the equation with latent variable z, we find the GMM for Eq. (9) as follows:

$$\hat{\theta} = \text{argmax} \sum_{j=1}^{K} \log\left(\int p(x_i, z_i|\theta) dz_i\right) \text{ and} \tag{14}$$

$$B[\{q_i(z_i)\},\theta] = \sum_{i=1}^{I} \int q_i(z_i) \log\left(\frac{p(x_i, z_i|\theta)}{q_i(z_i)}\right) dz_i \leq \sum_{i=1}^{I} \log\left(\int p(x_i, z_i|\theta) dz_i\right). \tag{15}$$

The logarithmic likelihood equation incorporating latent variable z is given by Eq. (14). From Eq. (15) and considering Jensen's inequality, we define the lower limit of the likelihood as $B[\{q_i(z_i)\},\theta]$, which is always lower than the maximum sought likelihood. The optimal parameters are estimated, and the lower limits of the maximum value and θ, i.e., the parameter that maximizes the lower limit, are subsequently found via recursive calculations.

3.2. GMM-based Procedure for Expressing Features

The Fisher vector-based algorithm [8][9] is a GMM-based procedure for expressing features that provides more accurate recognition with large image distributions than procedures that use image feature values, such as SIFT [10]. However, the Fisher vector treats all gradients of the GMM parameters as feature quantities; therefore, it requires many calculations and the feature quantities have high dimensionality. Therefore, we use the proposed GMM-MRCoHOG because it provides feature quantities of a lower dimensionality than the Fisher vector.

3.3. Training Feature Spaces

In our proposal method, we estimate an area of the GMM, where the similarity between positive image(in this paper, pedestrian) and negative image(an image of the other class) occurrence probability is low. To overcome above, at the first, MRCoHOG feature values in the probability distributions of a positive and of a negative are calculated to allocate basis functions using Eqs. (11) and (12) as follows:

$$f_p(x_p) = N\left(x_p\middle|\,\mu_z,\Sigma_z\right), \tag{16}$$

$$f_n(x_n) = N\left(x_n\middle|\,\mu_z,\Sigma_z\right). \tag{17}$$

Second, the obtained basis functions are used to generate samples and train new GMM parameters. This new GMM demonstrates a lower number of combinations than the obtained Gaussian mixture. Then, the second GMM is retrained.

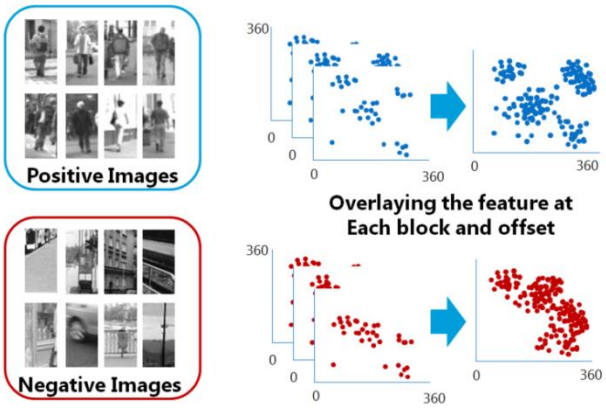


Fig. 3: Distribution of positive and negative images.

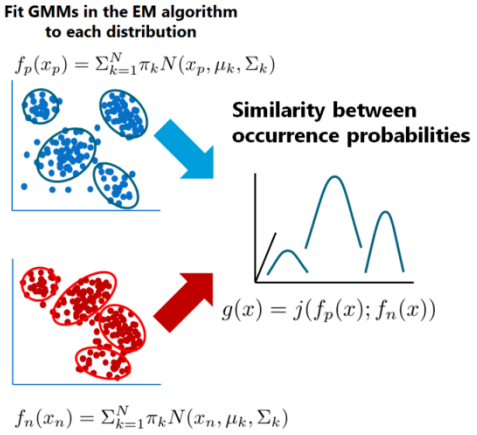


Fig. 4: Combinations of feature spaces.

Note that this process can reduce the number of parameters. Figure 3 shows distributions obtained from positive and negative images after the MRCoHOG feature values are calculated for those images and the feature planes of the block offsets are combined. Then, the EM algorithm is used to train the GMM on the obtained feature planes (Fig. 4). Note that we use the Jensen–Shannon (JS) divergence between positive image and negative image occurrence probability (Fig. 4).

3.3.1. Jensen–Shannon Divergence

The Kullback–Leibler (KL) divergence (Eq. (18)) measures the similarity between probabilities. As the certainty degree distribution $p(x)$ and $q(x)$ becomes increasingly similar, the KL divergence decreases, and when this value is equal, it becomes 0. However, since KL divergence does not satisfy symmetry as in Eq. (19) and does not satisfy triangular inequality, it does not satisfy the definition of general distance.

$$f_k(p(x); q(x)) = -\int p(x) \log \frac{p(x)}{q(x)} dx \geq 0, \tag{18}$$

$$\begin{aligned} f_k\big(p(x); q(x)\big) &= 0 \quad \big(p(x) = q(x)\big), \text{and} \\ f_k\big(p(x); q(x)\big) &\neq f_k \quad \big(q(x); p(x)\big). \end{aligned} \tag{19}$$

Using the KL divergence, the JS divergence is defined to satisfy symmetry and is defined as follows:

$$f_j\big(p(x);q(x)\big)=\tfrac{1}{2}f_k\big(p(x);q(x)\big)+\tfrac{1}{2}f_k\big(q(x);p(x)\big) \text{ and}$$
$$a(x)=\frac{1}{2}p(x)+\frac{1}{2}q(x). \tag{20}$$

From the KL divergence, Eq. (21) is used to define the JS divergence (Eq. (22)) by interpreting the similarity between the probabilities of p and q in a certain random variable x As follows:

$$k\big(p(x);q(x)\big)=p(x)\left(\log\frac{p(x)}{q(x)}\right), \tag{21}$$

$$j\big(p(x);q(x)\big)=\frac{1}{2}k\big(p(x);a(x)\big)+\frac{1}{2}k\big(q(x);a(x)\big), \tag{22}$$

$$a(x)=\tfrac{1}{2}p(x)+\tfrac{1}{2}q(x), \text{ and}$$
$$g(x)=j\big(f_p(x);f_n(x)\big). \tag{23}$$

We use the inverse function method in Eq. (23) to define Eqs. (24) and (25) (Fig. 5). Then, $H^{-1}(x)$, i.e., the inverse of $H(x)$, is used to generate random numbers obeying $h(x)$ from a uniform distribution of random numbers. The, we generate a new sample. Next, we employ GMM again with the EM algorithm.

$$h(x)=\frac{|g(x)|}{\int|g(x)|}, \tag{24}$$

$$H(x)=\int_0^x h(t)dt. \tag{25}$$

Then, the feature values found while learning the feature spaces with $g(x)$ are calculated. The load ratios that can be calculated using Eq. (26) using the feature values represent the probability that the GMM comprising the various weighted normal distributions is generated correctly using data x.

$$\gamma(z_k)=p(z=k|x)=\frac{\pi_k N(x|\mu_k,\Sigma_k)}{\sum_{j=1}^{K}\pi_j N\big(x\big|\mu_j,\Sigma_j\big)}. \tag{26}$$

While calculating feature values, the number of feature spaces equals the number of histograms; thus, the load ratio is calculated using the GMM obtained for each feature space. Then, the load ratios are listed, toconstitute feature

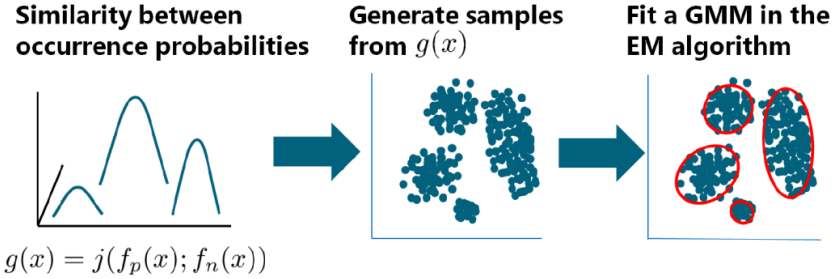


Fig. 5: Training the GMM using EM algorithm on generated samples.

quantities. Thus, the dimension of the proposed feature quantity, i.e., the number K of GMM combinations, is determined by the number of blocks and number of offsets.

3.3.2. Normalization of Feature Quantities

The procedure used for GMM feature calculations also tends to generate sparse features; therefore, to avoid overtraining, L2 normalization (Eq. (27)) is employed using the Fisher vector algorithm [8], followed by power normalization (Eq. (28)). The proposed procedure also employs feature quantities as GMM loads; thus, L2 and power normalizations are also performed.

$$\|x\|_2 = \sqrt{\Sigma x_i^2}, \tag{27}$$

$$x_i = \text{sign}(x_i)|x_i|^{\alpha}. \tag{28}$$

3.4. Experiment to assess GMM-MRCoHOG

The Daimler pedestrian dataset [11] was employed to train the proposed GMM-MRCoHOG, and the INRIA person dataset [5] was employed for evaluation. The subsequent numbers of feature dimensions are given in Table 1. The objects of comparison are an MRCoHOG (cmb3) and a CoHOG using pixels up to 4 pixels distant from the element employed for co-occurrence (mg4). Note that an SVM [11] was used for recognition. The results are discussed below.

Figure 6 shows that the proposed GMM-MRCoHOG provides higher accuracy than both MRCoHOG (cmb3) and CoHOG (mg4).

Table 1: Feature dimensions for feature qualities.

Method	Dimensions
CoHOG (mg4)	34,704
MRCoHOG (cmb3)	22,176
GMM-MRCoHOG (cmb3, K = 64)	21,504
GMM-MRCoHOG (cmb3, K = 32)	10,752
GMM-MRCoHOG (cmb3, K = 16)	5,376

4. Conclusion

This paper has proposed a procedure to describe features that employ GMM-based feature space training. The experimental results demonstrate that employing the GMM-MRCoHOG reduces the number of feature dimensions by ~75% while still providing high recognition accuracy. This suggests that the proposed method can be performed on less expensive hardware than conventional MRCoHOG. In future, we intend to establish an automatic procedure to determine mixture combinations and improve image analysis accuracy. In addition to the SVM employed in the current study, we intend to conduct experiments using algorithms such as AdaBoost and neural network. The obtained results will be compared to those of conventional methods, and the capabilities of the proposed GMM-MRCoHOG will be further examined.

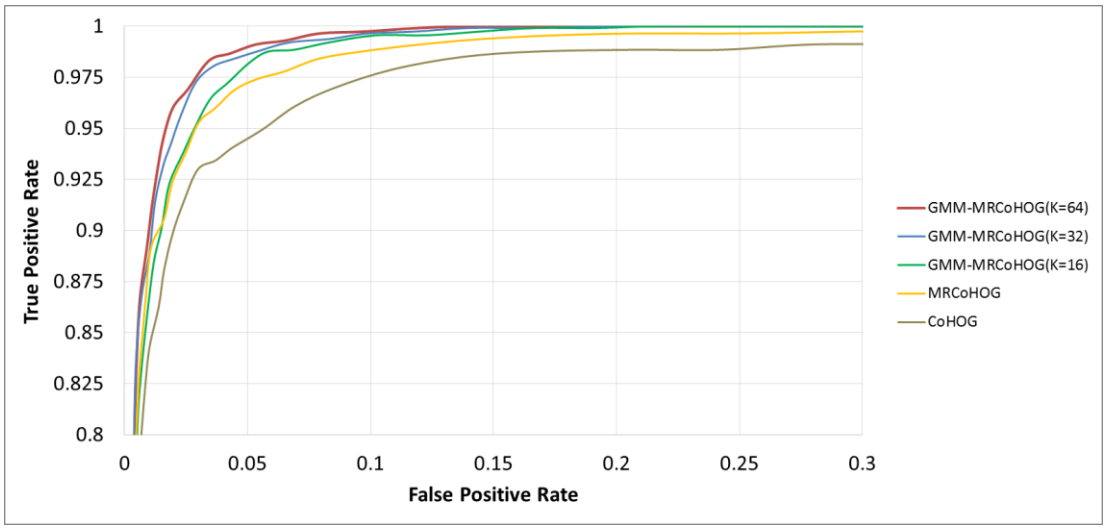


Fig. 6: Experimental result.

References

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, ed, *Advances in Neural Information Processing Systems* 25, pp. 1097–1105, Curran Associates, Inc., 2012.

[2] Y. Yamasaki, S. Ooe, A. Suzuki, K. Kuno, H. Yamada, S. Enokida, and H. Tamukoh, "Evaluation of hardware oriented MRCoHOG and digital circuit using logic simulation," *Proc. of the 2th International Joint Conference on Computer Vision Theory and Applications (VISAPP2017)*, vol. 6, pp. 341–345, 2017.

[3] M. Soga, S. Hiratsuka, H. Fukamachi, and Y. Ninomiya, "Pedestrian detection for a near infrared imaging system," *Proc. the 11th International IEEE Conference on ITSC*, pp.1167–1172, 2008.

[4] S. Iwata and S. Enokida, "Object detection based on multiresolution CoHOG," *Advances in Visual Computing Lecture Notes in Computer Science*, vol. 8888, no. 427–437, 2014.

[5] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," *Proc. IEEE Computer Vision and Pattern Recognition*, vol. 1, pp. 886–893, 2005.

[6] T. Watanabe, S. Ito, and K. Yokoi, "Co-occurrence histograms of oriented gradients for human detection," *Proc. Pasific-Rim Symposium on Image and Video Technology*, pp. 37–47, 2009.

[7] H. Cao, K. Yamaguchi, M. Ohta, T. Naito, and Y. Ninomiya, "Feature interaction descriptor for pedestrian detection," *Proc. IEICE Transactions on Information and Systems*, E93-D, no. 9, pp. 2656–2659, 2010.

[8] F. Perronnin, S. Jorge, and T. Mensink, "Improving the fisher kernel for large-scale image classification," In *European Conference on Computer Vision*, 2010.

[9] J. Sanchez, F. Perronnin, T. Mensink, and J. Verbeek, "Image classification with the fisher vector: Theory and practice," *International journal of computer vision*, no. 105, pp. 222–245, 2013.

[10] D. G. Lowe, "Object recognition from local scale-invariant features," In *Proc. of IEEE Intl. Conf. on Computer Vison*, pp. 1150–1157, 1999.

[11] M. Enzweiler and D. M. Gavrila, "Monocular pedestrian detection: Survey and experiments," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 31, no. 12, pp. 2179–2195, 2009.

[12] N. Cristianini and J. Shawe-Talor, "An Introduction to Support Vector Machines," Cambridge University Press, 2000.