PAPER

Histogram of Template for Pedestrian Detection

Shaopeng Tang,† *Non Member,* Satoshi Goto† *Fellow*

Summary
In this paper, we propose a novel feature named histogram of template (HOT) for human detection in still images. For every pixel of an image, various templates are defined, each of which contains the pixel itself and two of its neighboring pixels. If the texture and gradient values of the three pixels satisfy a pre-defined formula, the central pixel is regarded to meet the corresponding template for this formula. Histograms of pixels meeting various templates are calculated for a set of formulas, and combined to be the feature for detection. Compared to the other features, the proposed feature takes texture as well as the gradient information into consideration. Besides, it reflects the relationship between 3 pixels, instead of focusing on only one. Experiments for human detection are performed on INRIA dataset, which shows the proposed HOT feature is more discriminative than histogram of orientated gradient (HOG) feature, under the same training method.

Key words:
Human detection, Histogram of template, SVM

1. Introduction

Human detection technique is widely used in many applications ranging from image analysis, smart cars, and visual surveillance to behavioral analysis. In recent years, lots of research work has been focused on this field. But human detection is still a challenging task because of many difficulties. Most natural humans have large variations, such as the appearance, the pose and so on. Difference in clothes brings further challenge because some features such as skin color in the face detection can't be used in this application. Besides, complex backgrounds, illumination, occlusions and different scales must be considered in the detection. A robust detector must be independent for all these variations.

The gradient information is efficient for the object detection. A lot of human descriptors contain the gradient information more or less [1-6]. Histogram of orientated gradient (HOG) [1] and covariance matrix [3] are excellent descriptors using the gradient information. HOG is a gray-level image feature formed by a set of normalized gradient histogram. In [1-2], HOG feature is compared with many other features, such as Harr-like feature, Wavelet feature and so on, and it gets the best performance. Covariance matrix integrates coordinates and intensity derivatives into a matrix. They represent the gradient information well, and can get a good result on INRIA human dataset. But only using the gradient information may be not enough to detect humans from complex backgrounds or images in low resolution.

The texture information also has some discriminative abilities in the human detection. Some research work has been done on the feature of local binary pattern (LBP) [7] and Gabor filter [8]. Gabor filter and LBP are widely used in texture classification and face recognition. They represent the intensity information well. But only using these features is not enough to get the good result. The original definition of LBP is not suitable for the human detection. It must be combined with other features such as Laplacian EigenMap (LEM) in [8]. In [7], two variants of LBP: Semantic-LBP and Fourier-LBP are proposed. The modified definition of LBP makes it suitable for the human detection.

Besides the feature extraction, the training method is also very important for the human detection. They are two key components for the pattern classification problem. The features extracted from a large number of training samples are used to train a classifier. Support vector machine (SVM) [9] and various boosting methods [10] are efficient to train a classifier in practical applications. The SVM has some advantages. It is easy to train and the global optimum is guaranteed. The variance caused by the suboptimal training is avoided for the fair comparison. The boosting method combined with the cascade strategy is widely using in real-time applications. The boosting method aims at producing an accurate combined classifier from a sequence of weak classifiers, which are fitted to iteratively reweighted versions of the data. The cascade strategy saves detection time and makes it possible to detect object real time.

So there are two research directions for the human detection: finding more discriminative local features [1, 3], and developing more efficient training methods [11].

The main contribution of this paper is along the first direction. It focuses on building a more powerful local feature for the human detection. A new feature, histogram of template, is proposed. It extracts the texture information as well as the gradient information, and makes the two different types of information homologous. Compared with features using the gradient information, such as the HOG feature, the proposed feature shows more discriminative ability. Besides, this feature can encode the relationship of three pixels in one template. Compared with features that deal with each pixel independently, HOT feature can get higher detection rate. Last, the HOT feature has some properties of local binary pattern, such as

Manuscript received January xx, 20xx.
Manuscript revised March xx, 20xx.
† The author is with NTT , Musashino-shi, 180-8585,
†† The author is with IEICE Office, Minato-ku, Tokyo, 105-0011 Japan.

illumination-invariance. So the normalization is not so important for the detection result. This property can be used to reduce computation complexity in some circumstance.

2. Related Work

Human detection algorithms now can be separated into three groups.

The first group of methods is based on local features [1-4, 6-7, 11-13]. They extract some features from sub regions of images in the training dataset, to train a classifier by support vector machine (SVM) or boosting methods, such as Adaboost or Logitboost. For a new image, they extract the same features and send them to the classifier which will give a classification result. In [14], a local receptive fields (LRF) feature is extracted using multilayer perceptrons by means of their hidden layer. In [13], Haar wavelet is used as human descriptor. SVM in [9] is used to train the classifier. [1] uses the HOG feature as descriptor for the human detection, and [2] is developed from this one. It integrates the cascade-of-rejecter approach, and uses the Adaboost method in [11] to choose best sub window in each stage. In [11] Haar-like feature is used to detect humans. It uses the integral image to speed up the detection process. Cascade rejection method is proposed to make real-time human detection possible. In [3] the covariance matrix feature is used as human descriptor, to represent the coordinates, and the gradient information of humans. Covariance matrix can be formulated as connect Riemannian manifold. Each matrix can be treated as a point in Riemannian manifold, and can be mapped into a vector space. An edge let descriptor is used in [12] for the human detection. Different from just combining the orientations in horizontal and vertical direction in [4], it combines the orientations in edge let defined direction, which makes it more efficient for the human detection. This group of methods has a good performance, and if enlarge the training dataset, the detection rate can be improved.

The second group of methods is based on local appearance, and [15-18] are based on this. They detect the interesting points in the training images and use the patches around the interest points to construct a codebook. When given a new image, they first find the similar patches in codebook and all patches vote for the positions of humans.

The third group is based on chamfer matching. They use human templates to find the most marching regions in the edge map of an input image. [19-20] are based on this method. In [19] a direct template matching approach for the global shape-based human detection is proposed, and [20] is developed from this but uses some hierarchical templates to reduce the detection time and solve the occlusion problem to some extent. These methods may not give a good result when there are too many edge clusters in the edge map.

Our method belongs to the first group. It uses HOT feature to extract the texture information and the gradient information for the human detection. Two types of information are made homologous to increase the discriminative abilities of the proposed feature. The covariance matrix feature in [3] gets higher detection rate than HOG feature in [1]. But the training method is different. [3] uses the logitboost method and the size of sub windows is variable, but the SVM training method and the fixed sub window strategy are used in [1]. So it is hard to say that whether the covariance matrix feature is more discriminative or the training method is better. The HOT feature is compared with the HOG feature using the same training method, for the fair comparison.

3. Feature Extraction

3.1 Previous Feature

HOG is developed from the SIFT algorithm [5]. For calculating the HOG feature, the image is divided into blocks. The blocks overlap with each other. Each block contains four cells. Cell is the basic unit for the feature calculation. For each pixel $I(x, y)$, the orientation $\theta(x, y)$ and the magnitude $m(x, y)$ of the gradient are calculated by

$$dx = \mathrm{I}(x+1, y) - \mathrm{I}(x-1, y) \tag{1}$$

$$dy = \mathrm{I}(x, y+1) - \mathrm{I}(x, y-1) \tag{2}$$

$$m(x, y) = \sqrt{dx^2 + dy^2} \tag{3}$$

$$\theta(x, y) = \tan^{-1}(dy / dx) \tag{4}$$

A histogram is calculated for each cell, and the length of each bin is the sum of magnitude of the pixels whose orientations are in the corresponding interval. In [4], each block contains 2×2 cells, so a block can be represented by a 36-dimensional vector.

COV calculates a vector for each pixel in a sub window:

$$[x, y, |I_x|, |I_y|, \sqrt{I_x^2 + I_y^2}, |I_{xx}|, |I_{yy}|, \arctan\frac{|I_x|}{|I_y|}]^T \tag{5}$$

Where x, y are pixel locations, and I_x , I_{xx} , I_y , I_{yy} are intensity derivatives. The last term is the edge orientation. So for each sub region, we calculate a set of 8-dimensional vectors, and a covariance matrix can be obtained from these vectors:

$$C_R = \frac{1}{S-1}\sum_{i=1}^{S}(z_i-\mu)(z_i-\mu)^T \qquad (6)$$

Where μ is the mean, S is the number of these vectors. Due to the symmetry of covariance matrix, only the upper triangular part is stored as the feature for the detection. A descriptor of a sub region is a 36-dimensional vector.

3.2 Limitation

The HOG and the COV feature are mainly depended on the gradient information. There are some disadvantages of gradient-based features.

Sometimes, the gradient information is ambiguous. The same gradient may correspond to the different curves. See Fig.1 for example. Point P is the intersection of curve A and curve B . Only using the gradient information of P is not enough to discriminate A and B . But if the template feature is used, because the smooth degrees are different, P on A is more likely to meet the second template and the P on B is more likely to meet the first template.

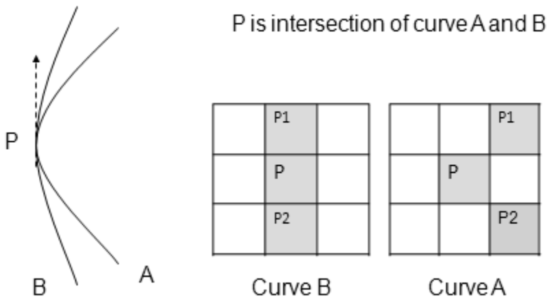

Fig.1 Disadvantage of Gradient based feature. It may be ambiguous in some circumstance, if only gradient information is used.

Gradient based features almost only use the gradient information for the detection, and drop the texture information in the original image, although three channels of color image are used in gradient calculation. The texture information also shows discriminative abilities in LBP based features [7-8] and local appearance based features [15-18]. So if texture information can be used with the gradient information, more accurate detection result can be obtained.

3.3 Histogram of Template Feature

Histogram of Template feature is proposed here. Some templates are given to define the spacial relationship of three pixels. See Fig.2 for example.

In Fig.2, 12 connected templates are given. In our experiment, the templates (1) to (8) are used for the feature calculating. 12 templates can be used for more accurate result.

These templates are used in some formulas. The texture information and the gradient information are also used in these formulas, to give a concrete definition of this feature. The formulas are designed to capture the shape of the human body, and have reasonable computation complexity.

For texture information, two formulas are given as following. First is:

$$I(P) > I(P1)\ \&\&\ I(P) > I(P2) \qquad (7)$$

For each template, if the intensity value of P is greater than the other two, it is regarded that the pixel P meets this template. It can capture the pixels that have the greatest value in one template, and the histogram of pixels that satisfy each template in a sub window can reflect the properties of local part of human body well.

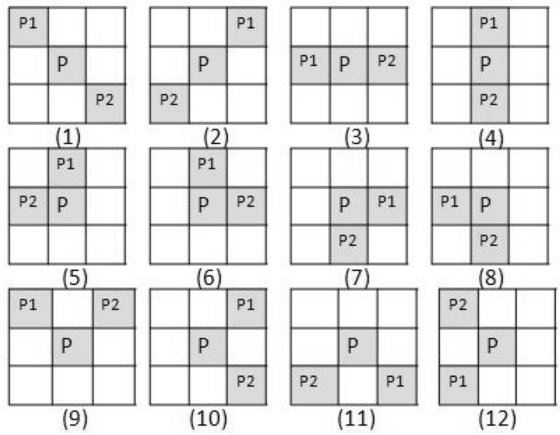

Fig.2 There is 12 templates here. They are three pixels' combination.

For each sub window, the number of pixels meeting each template is calculated to get a histogram. See Fig. 3. For example, eight templates are used to extract the feature.

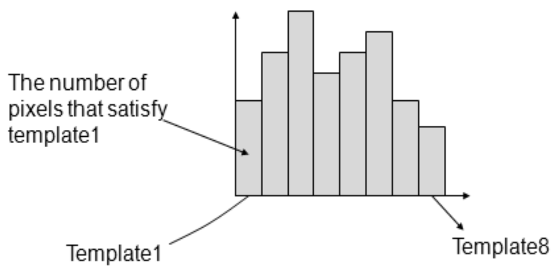

Fig.3 Example of histogram of template for one formula; 8 templates are used, and they correspond to 8 bins. The value of each bin is the number of pixels that meeting corresponding template.

The histogram has eight bins and each bin corresponds to one template. The value of each bin is the amount of pixels which meet the requirement of this template in this sub region.

The second formula is:

$$k == \arg\max_i \{I(P_i) + I(P1_i) + I(P2_i)\} \quad (8)$$

The sum of intensity values of three pixels in template k is greater than the values of other templates; it is can be regarded that P meets template k. A histogram can be calculated by using formula (8). By using this formula, we could find the template that has the greatest sum. They can be regarded as the basic unit of human body shape and the shape of human body can be represented well.

For the gradient magnitude information, there exist similar formulas:

$$Mag(P) > Mag(P1) \,\&\&\, Mag(P) > Mag(P2) \quad (9)$$

$$k == \arg\max_i \{Mag(P_i) + Mag(P1_i) + Mag(P2_i)\} \quad (10)$$

Eight templates are usually used to extract the feature, so for each formula, an eight-dimensional vector can be obtained. These vectors are combined together as the final feature. See Fig.4.

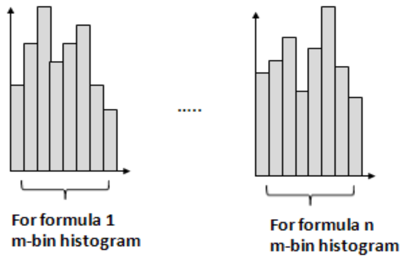

Fig.4. Final HOT feature for a sub window. It is a $m \times n$ dimensional vector. In our experiment, m =8 and n =4.

The integral image can be used for feature extraction. For example, if 4 formulas and 8 templates are used, the histogram has 32 bins. 32 additional images are used. One image corresponds to one bin. If the pixel in the original image satisfies one template for one formula, the value of the pixel in the corresponding additional image is 1; otherwise it is 0. Then, by constructing the integral images of the additional image, we could get the 32-bin histogram for each sub window quickly.

Compared with HOG feature, HOT feature has three advantages. First is that it not only uses gradient information, but also uses texture information. Although HOG feature also uses three channels of color image for gradient calculation, the texture information is ignored and it is not treated as a cue for detection. The second is that HOT feature is more macrostructures. HOG is actually an orientation voting, and after the gradient is computed, the feature is calculated from pixel level. The HOT feature is specific pattern voting and the feature is extracted from middle level, which contains several three pixels' combination. So HOT feature is more discriminative, and experiment confirms this point. The third is that HOT feature is illumination-invariant, so the normalization is not necessary as HOG feature.

4. Training Method

The training method is also very important for the detection result. A reasonable training method improves the result efficiently. So for the fair comparison of different features, the effect of training method should be considered. Support vector machine and many boosting methods, such as Adaboost, Logitboost and Gentleboost, are widely used in many tasks. In our experiment, SVM is used for comparison.

SVM in [1] is effective for learning with small sampling in high-dimensional spaces. The objective of SVM is to find a decision plane that maximizes the inter-class margin. The feature vectors are projected into a higher dimensional space by kernel function. The kernel function makes it possible to solve the linear non-separable problems and the mapping function is not necessarily known explicitly. So the decision rule is give by the following formula.

$$f(x) = \sum_{i=1}^{N_s} \beta_i K(x_i, x) + b \quad (11)$$

Where x_i are support vectors, N_s is the number of support vectors. $K(x, y)$ is the kernel function. So the training process of SVM is to find the proper parameters of (11).

Compared with boosting methods, SVM needs more computational resources and it is difficult for real-time application. The size of sub windows should be fixed. It is hard to take the variable sub-window size strategy due to the computation problem, although the variable window strategy can improve the performance efficiently. But for the comparison purpose, SVM is suitable. The training time is less and the optimization is guaranteed. The difference of the performance caused by the optimization can be ignored. The parameters of SVM are controllable. The suitable parameters can be selected avoiding the difference caused by the parameter difference. In our experiment, LibSVM [21] is used. RBF and linear kernel functions are used in our experiment respectively.

5. Experiment

5.1 Dataset

The experiment is performed on INRIA dataset [22]. It is widely used for the human detection in still images. The database contains 1774 human annotations and 1671 person free image. This dataset is made up of a training dataset and a testing dataset. 1208 human annotations and 1218 non-human images are used for the training stage, and the left images for testing. For positive samples, left-right reflections are also used. So, 2416 positive samples are used for training. More detail can be seen in [22]. There are varieties of variations in human pose, clothing, lighting, clutters and occlusions, so it is difficult for the human detection and it is suitable as a benchmark for comparison between different methods. Some examples can be seen in Fig. 5.

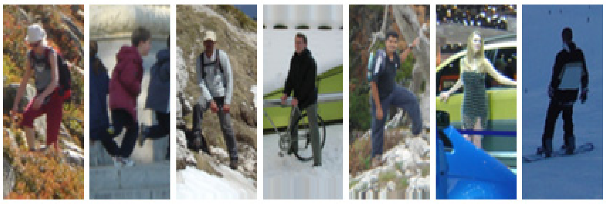

Fig. 5 Selected positive samples in INRIA dataset

5.2 Comparison with other features

In order to show the advantage of the proposed feature, we design the following three experiments. In the first experiment, we compare our feature with HOG feature and COV feature. We use the same strategy with [1]. The re-sample strategy and normalization strategy are also used in our experiment. The size of sub window and the stride between sub windows are provided by [1]. In the second experiment, a random ensemble strategy is used. So we don't have to consider the size of sub window and the stride. The comparison is fairer. In the third experiment, we evaluate the length of the proposed feature. Only 8 templates are used in the first experiment, and we show that if more templates are used, the performance would be further improved. In the forth experiment, we show the results with respect of the change of parameters and training strategies. We want to show that the result of our feature in different configurations.

(1) In the first experiment, we compare the HOT feature with the HOG and COV feature. We use the same strategy with the HOG feature. The Re-sample and normalization are used in our experiment, just like the HOG in [1]. The size of sub window and the stride are also provided by [1]. In the re-sample stage, 2416 positive samples and 12800 negative samples random selected are used as the initial training dataset. And there are 39000 hard negative samples in our experiment. The initial training dataset and the hard negative samples are used to get the final classifier. The Linear kernel and the RBF kernel are used for training respectively. C value and g value of RBF kernel are selected by cross-validation method, which is a common method in SVM. LibSVM provides this tool. C and g are obtained by only using the training dataset. The C value is 128 and g value is 0.00048828125. We also use the default C and g to evaluate the performance of our feature. From Fig.6, it can be seen that the result is nearly the same. The size of block is set as 16×16 and the stride between two blocks is 8. They are the same with the HOG feature. The feature length of a block is 32, since the first eight templates in Fig.2. So the length of the feature for a 64×128 image is 3360. It is shorter than HOG and COV, which means that the computation complexity and the memory consumption is less. The comparison result can be seen in Fig.6. The data of HOG and COV are copied from [3] for comparison. The configuration for each feature can be seen in Table.1.

Table.1 the experiment configurations of three features

	HOG	COV	HOT
Feature Dimension	36	36	32
Re-Sample	Y	Y	Y
Sub window size	16×16	Variable	16×16
Stride	8	N	8
Training method	SVM	Logitboost	SVM
Normalization	Y	Y	Y
Unbalance data	N	N	N

From figure 6, it can be seen that the HOT feature gets higher detection rate than HOG and COV feature at 10^{-4} FPPW. Usually, we compare the miss rate at 10^{-4} FPPW [1]. Take into account that COV feature uses variable sub window. It can improve the detection rate a lot compared with using fixed sub window [2-3, 7]. We could say that our feature outperforms HOG and COV.

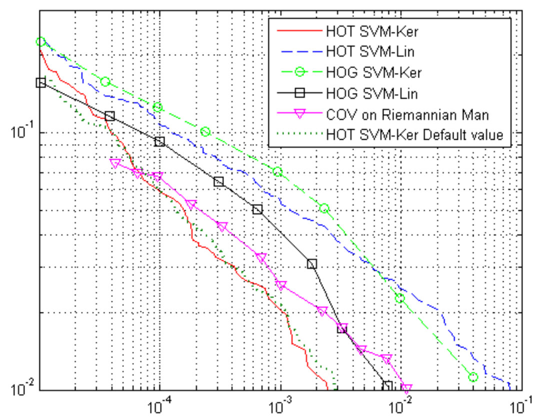

Fig.6. Comparison with methods of HOG [1] and COV [3]. The curves of HOG and COV are copied from [3].

(2) In the second experiment, we try to reduce the influence of block size and stride. A random ensemble strategy in [7] is used.

From an image (64×128) in our experiment, lots of sub windows in different sizes and on different positions can be extracted. The minimum sub window size is set as $K \times K$. This size is incremented in a step of K horizontally and vertically, or both. Finally we can get all possible sub windows: $W_{subwin} = \{r_i\}$. In our experiment, K =8. So the cardinality of W_{subwin} is 4896. Smaller K will give more sub windows.

Random ensemble means that n_w sub windows are random selected from W_{subwin} . In our experiment, n_w =150. A set of sub windows can be obtained: $\{r_j, j = 1, 2, ...n_w\}$. For each sub window, a feature is calculated. The features for all random selected sub windows can be obtained as: $\{f_j, j = 1, 2, ...n_w\}$. So the final feature for this detection window can be represented as $F = \{f_1, f_2 ... f_{n_w}\}$. If the lengthen for each sub window is d , the dimension of the final feature is $d \times n_w$.

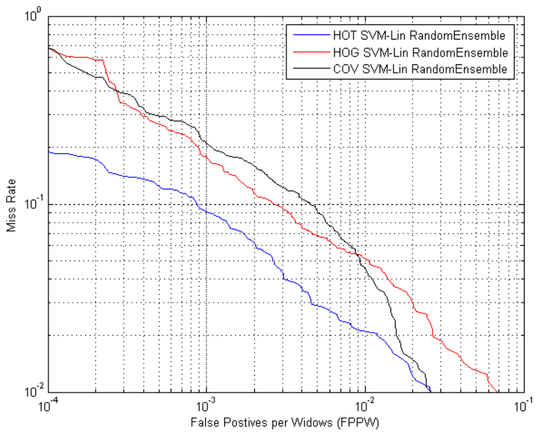

Fig. 7 Comparison of three features Random ensemble strategy is used here. The influence of parameters can be ignored.

By using the random ensemble strategy, the influence of block size and stride can be ignored, because all sub windows are random selected from W_{subwin} . HOT, HOG and COV are evaluated by using this strategy. The initial training dataset in the first experiment is used for training. Lin-Ker SVM is used, so there is no C value and g value. In this evaluation strategy, there is no any parameter for the feature extraction. The comparison of three features can be seen in Fig.7. Our feature outperforms HOG and COV in this experiment, which shows the discriminative ability of our feature.

(3) In the third experiment, lengthen of HOT feature is evaluated. The HOT feature is computed by using formulas and templates. The first eight templates in Fig.2 are used for the 32-dimensional feature for a sub window in the above experiments. If more templates are used, the detection result will be improved. The performance is evaluated in this experiment.

The template is actually the three pixels' combination. There are 9×8×7 possible templates in all in a 3×3 region. We only consider the connected one. Some connected templates containing central pixel can be seen in Fig.2 and Fig.8. Other connected templates can be obtained by shifting these 20 templates.

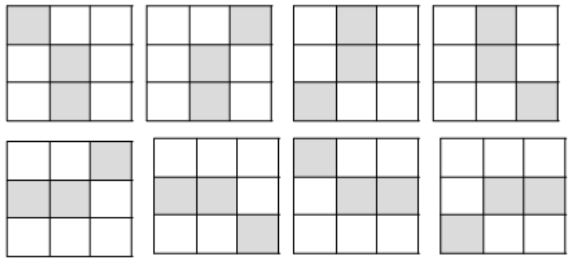

Fig.8 Another 8 connected templates containing central pixel

The detection result of 12 templates and 20 templates can be seen in Fig.9.

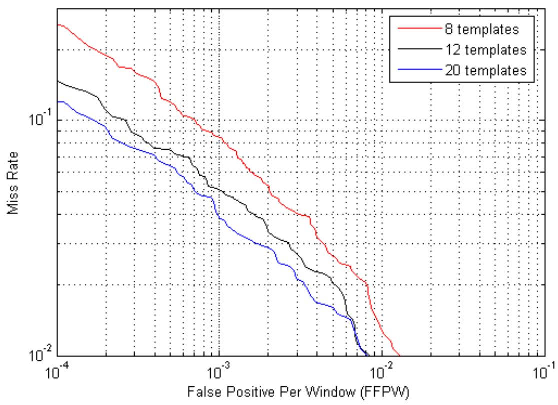

Fig.9 Detection result when more templates are used.

In Fig.9, the initial training dataset and SVM of RBF kernel are used for training. For 8 templates case, the first 8 templates in Fig.2 are used. For 12 templates case, all templates in Fig.2 are used. For 20 templates case, all corrected templates containing central pixel in Fig.2 and Fig.8 are used. From Fig.9, it can be seen that when increase the number of templates, the detection result is improved. But the improvement is limited when the number of template is increased from 12 to 20. When use more templates, the length of the feature will increase. It means that more time is needed for the classification. So it may be not necessary to use all the three-pixel combination. 8 templates or 12 templates are suitable for the human detection. It is a tradeoff between the detection rate and the computation complexity. Mention that in the first and second experiment, only the first 8 templates are used.

(4) In the fourth experiment, the performances of HOT feature in different parameter configurations and training strategies are evaluated. We want to show the performance of proposed feature in the different configuration. For the training strategy, we consider the normalization strategy, unbalance data strategy; for the parameter, we evaluate the size of sub window and the stride between two neighboring sub windows. The initial training dataset in the first experiment is used for training.

Normalization schemes: In experiment, Non-norm, L1-norm and L2-norm strategy are used for comparison. Let v be un-normalized feature vector. The schemes are:

(a) Non-norm $v->v$;

(b) L1-norm $v->v/(\|v\|_1+\xi)$;

(c) L2-norm $v->v/\sqrt{\|v\|_2^2+\xi^2}$.

See Fig.10 for performance comparison. L2-norm outperforms Non-norm and L1-norm schemes, but the difference is not too much. So this step is not necessary for HOT feature. HOT feature has illumination-invariant property itself. This operation can't be ignored for HOG feature because of illumination. It means that we could accelerate the feature extraction by abandoning the normalization. So only the integer calculation is needed in the feature extraction. It is easy for hardware based acceleration. In the first experiment, L2 is used for fair comparison because HOG uses L2.

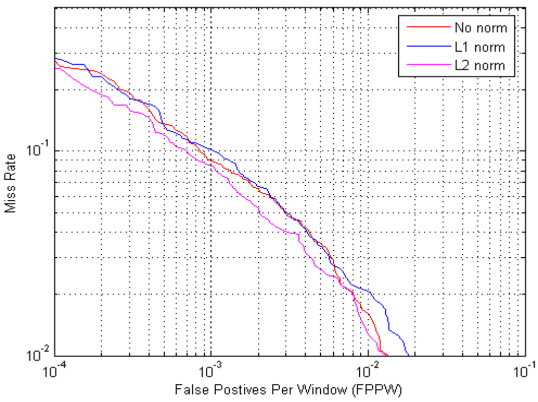

Fig.10 Comparison of different normalization schemes

Unbalance data: Since the number of negative samples is larger than that of positive samples, more negative images are used for training than positive images. It is reasonable that different penalty value for different classes may increase the detection rate. C of negative samples: C of positive samples is set as 3:1, 2:1, 1:1, 1:2 and 1:3 for comparison. See Fig.11. It can be seen that if the penalty value of positive samples is greater than that of negative samples, the performance can be improved. But the difference is not too much. In the first experiment, we don't consider it, and the same penalty is used for positive samples and negative samples for fair comparison.

Sub window size: For an inputting detection image (64×128), it is first divided into many sub windows (block). Sub windows can overlap with each others. Since the size of sub window is fixed, suitable size should be decided for training and detection. In experiment, 12×12, 16×16 and 20×20 are used for comparison. See Fig.12. 20×20 has the best performance. The result of 16×16 and 12×12 are nearly the same as 20×20. For fair comparison, 16×16 is used in first experiment.

Stride between sub windows: the distance between two neighboring blocks. The area of overlap region of two sub windows is decided by this value. The less this value is, the longer the final feature is. In experiment, 4, 8 and 16 are evaluated. See Fig.13. 8 is used in the first experiment, just like the HOG feature.

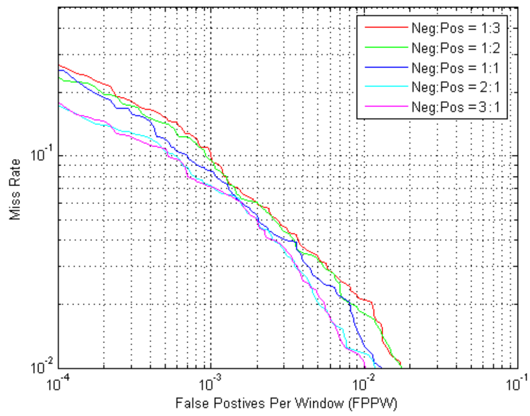

Fig.11 Different penalty values for the different classes

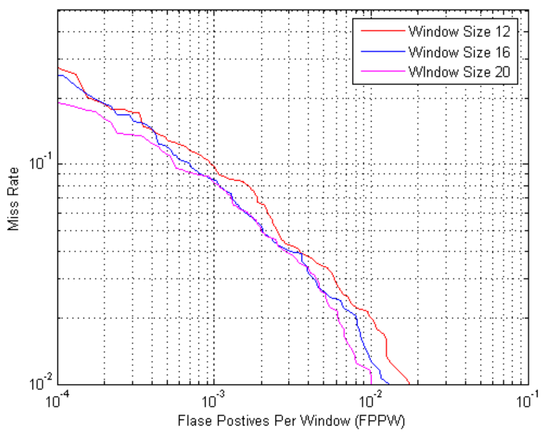

Fig.12 The size of the sub window

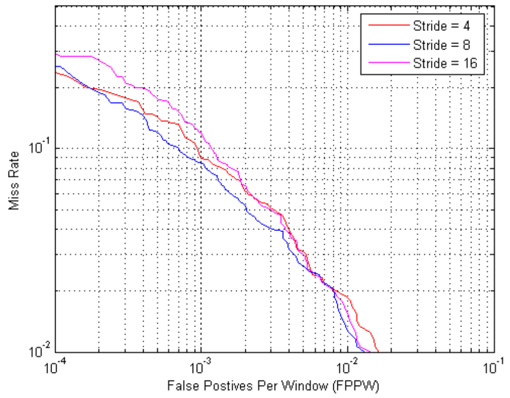

Fig.13 Stride between two sub windows

Finally, some detection results from natural images by using the classifier obtained in the first experiment can be seen in Fig.14.

Fig.14 Detection result of natural images

6. Conclusion

A new feature for human detection is proposed in this paper. A histogram of pixels meeting different templates is used as a feature for the human detection. It integrates the texture information and the gradient information together, and shows more discriminative ability than only gradient based feature, even if the length of feature is shorter. Other advantage is that HOT feature is illumination-invariant, so the normalization is not the necessary step for detection, which is very useful for some hardware accelerators that only support integer calculation. In our experiment, the size of the sub window is fixed. It is expected that the variable window size and the boosting method will further improve the performance of the HOT feature. The computation of the HOT feature is parallel, so it is easy for hardware acceleration. Besides, integral image can also be used for computation. These factors make it possible for real-time application.

Acknowledgments

This research was supported by “Ambient SOC Global COE Program of Waseda University” of the Ministry of Education, Culture, Sports, Science and Technology, Japan, and CREST Program.

References

[1] N.Dalal and B.Triggs, "Histograms of oriented gradients for human detection," in *Conference on Computer Vision and Pattern Recognition*, 2005.

[2] Q.Zhu, *et al.*, "Fast human detection using a cascade of histograms of oriented gradients," in *IEEE Conf. on Computer Vision and Pattern Recognition*, New York, 2006, pp. 1491-1498.

[3] T.Oncel and P.Fatih, "Pedestrian detection via classification on riemannian manifolds," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 30, Oct. 2008.

[4] K.Mikolajczyk, *et al.*, "Human detection based on a probabilistic assembly of robust part detector," in *ECCV*, 2004, pp. 69-82.

[5] D.Lowe, "Distinctive image features from scale-invariant keypoints," *IJCV,* vol. 60, pp. 91-110, 2004.

[6] Z.Lin and S.Larry, "A Pose-Invariant Descriptor for Human Detection and Segmentation," in *ECCV*, 2008.

[7] M.Yadong and Y.Shuicheng, "Discriminative local binary patterns for human detection in personal album," in *CVPR*, 2008.

[8] L.Nanni and A.Lumini, "Ensemble of multiple pedestrain reprssentations," *IEEE Transactions on Intelligent Transportation Systems,* vol. 9, June 2008.

[9] B. Scholkopf and A. Smola. (2002) Learning with kernels support vector machines, regularization, optimization and beyond.

[10] J.Friedman, *et al.*, "Additive logistic regression: a statistical view of boosting," *Ann. Stat.,* vol. 28, pp. 337-407, 2000.

[11] P.Viola and M.Jones, "Rapid object detection using a boosted cascade of simple features," in *Coference on Computer Vision and Pattern Recognition*, 2001.

[12] B.Wu and R.Nevatia, "Detection and tracking of Multiple, partially occluded humans by bayesian combination of edgelet based part detectors," *IJCV,* 2007.

[13] P. Papageorious and T. Poggio, "A trainable system for object detection," *IJCV,* vol. 38, pp. 15-33, 2000.

[14] S.Munder and D.M.Gavrila, "An experimental study on pedestrian classification," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 28, November 2006.

[15] B.Leibe, *et al.*, "Pedestrian detection in crowded scenes," in *IEEE Conf. on Computer Vision and Pattern Recognition*, San Diego, 2005, pp. 878-885.

[16] B.Leibe, *et al.*, "Combined object categorization and segmentation with an implicit shape model," in *ECCV*, 2004, pp. 17-32.

[17] S.Agarwal and D.Roth, "Learning a sparse representation for object detection," in *ECCV*, 2002.

[18] E.Seemann, *et al.*, "Towards robust pedestrian detection in crowded image sequences," in *CVPR*, 2007.

[19] D.M.Gavrila and V.Philomin, "Real-time object detection for smart vehicles," 1999, pp. 87-93.

[20] Z.Lin, *et al.*, "Hierarchical part-template matching for human detection and segmentation," in *IEEE International Conference on Computer Vision*, Rio de Janeiro, Brazil, 2007.

[21] LibSVM [Online]. Available: http://www.csie.ntu.edu.tw/~cjlin/libsvm/

[22] INRIA Dataset [Online]. Available: http://lear.inrialpes.fr/data

Manuscript received January xx, 20xx.
Manuscript revised March xx, 20xx.
† The author is with NTT , Musashino-shi, 180-8585,
†† The author is with IEICE Office, Minato-ku, Tokyo, 105-0011 Japan.