PEDESTRIAN DETECTION BASED ON BIDIRECTIONAL LOCAL TEMPLATE PATTERNS

Jiu Xu, Ning Jiang and Satoshi Goto

Graduate School of Information Production, and System LSI, Waseda University, Japan
xujiu@ruri.waseda.jp; jiangning@ruri.waseda.jp; goto@waseda.jp

ABSTRACT

In this paper, a novel feature named bidirectional local template patterns (B-LTP) is proposed to achieve pedestrian detection in still image. This feature is a combined and modified version of histogram of template (HOT) [1] and center-symmetric local binary patterns (CS-LBP) [2]. For each pixel, four templates are defined, each of which contains the pixel itself and two of its neighboring center-symmetric pixels. For each template, not only the relationships between three pixels according to the template, but also information of two directions are calculated in our feature, which makes it more discriminative. Moreover, the feature length of B-LTP is very short, which costs less computational workload and memory consumption. Experimental results on INRIA dataset show that both the speed and detection rate of our proposed B-LTP feature outperform other features such as histogram of orientated gradient (HOG) [3], HOT and Covariance Matrix (COV) [4] .

Index Terms— Pedestrian detection, bidirectional local template patterns, support vector machine

1. INTRODUCTION

Pedestrian detection is one of the hotspots researches in computer vision in recent years and real time pedestrian detection is the most critical step in intelligent video surveillance systems. The main target of pedestrian detection is to locate the human body out of the background. It is a challenging task to find a robust feature which makes human body be discriminated even in complicated background.

In the past decade, many researchers have been focusing on people detection from still image [3, 4, 5, 6]. One of the most promising methods for this problem is sub-window based approach. For the given image, firstly, this method scans all possible sub windows by searching the input image in different scales and positions exhaustively. On the second stage, for each sliding window, features are extracted and fed to the classifier which has been trained in advance using the training data of the same type of features. Finally, the classifier will divide each sliding window into two types that contains a pedestrian or not.

In order to improve the performance of pedestrian detection, one of the most direct ways is the feature extraction. Good features should be discriminative and robust. It should also have the ability of anti-noise and invariance.

Up to now, a lot of different features have been proposed. For example, some gradient based features are proposed in [3, 4, 7, 8]. Histogram of orientated gradient (HOG) [3] and covariance matrix feature (COV) [4] achieve excellent performance using gradient information. Some texture based features is also compact in human detection. In [5], Semantic-LBP and Fourier-LBP are proposed to make conventional LBP suitable for human detection. In [9], Center-Symmetric Local Binary/Trinary Patterns (CS-LBP/LTP) are utilized in considering the relation of the center-symmetric pairs of pixels. In [1], a Histogram of Template (HOT) is proposed which using eight templates and four formulas to extract the local templates based binary features.

Some other features use a combination of multiple complementary features, such as in [10], where HOG and LBP [13] are combined as the feature set. After adding an occlusion handling process, it can achieve the best performance by solving the occlusion problem.

The choice of classifiers such as support vector machines (SVMs) [11] and Boosting [12] is another important method for pedestrian detection. SVM method is easy trainable and the global optimum is guaranteed. Moreover, SVM detection can be easily accelerated by GPU and Multi-core in feature extraction part. Boosting methods are often used in cascade detector. It has a good performance in combining a strong classifier with lots of weak classifiers. At the same time, these detectors can save huge detecting time by discarding the background windows rapidly.

The key contribution of this paper focuses on building a more powerful and efficient feature for human detection, which can be summarized as follows.

- A two-directional local template patterns for human detection is proposed with an appropriated novel feature.
- Better than CS-LBP feature and HOT feature, both the gradient information and the texture information of the four templates selected are considered entirely in B-

- LTP, because not only each center pixel is compared with its center-symmetric neighbors but also the center-symmetric neighbors themselves are compared with each other.
- Regarding the B-LTP feature, calculations carried on for two directions of each template, which could be more discriminative than unidirectional one. It is the first time to apply the directional information of template to human detection.
- The feature dimensions of B-LTP are much less than HOG or HOT. For instance, the length of the HOT feature for a 64×128 image is 3360 and HOG is 3780 while our proposed B-LTP feature only cost 1008 dimensions, which is beneficial for extremely fast treatment. However, the performance of B-LTP is the best, not to mention that if we increase the length of our feature, the performance could be further improved.

The rest of this paper is organized as follows. In Section 2, proposed feature is described in detail. Experimental results are presented in Section 3, and finally, Section 4 provides summary and conclusions.

2. PROPOSED BIDIRECTIONAL LOCAL TEMPLATE PATTERNS

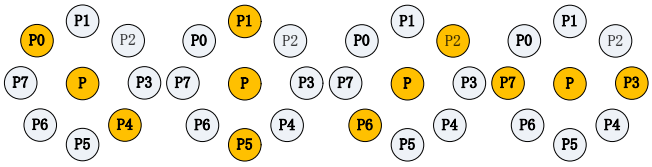

Fig 1 – Four templates in Bidirectional Local Template Patterns

Motivated from HOT and CS-LBP, the proposed B-LTP feature is provided. First of all, we select four most important templates (see Fig 1) as our basic templates in B-LTP. The reason why these four templates are the most important ones is that it considers gradient or texture value differences among central pixel and pairs of opposite pixels in a neighborhood, which is closely related to some gradient operator, such as HOG, and some texture operator, take LBP for example. Moreover, in CS-LBP, the four opposite pairs are used to represent 16 orientation bins which are evenly spaced over 0 to 360. It could also capture the edges and salient textures.

Based on these four templates, similar with HOT, we define several formulas to calculate the gradient and texture information separately.

For the texture part, three formulas are designed to generate histogram. This first one is modified from CS-LBP and is given as follows (I (P) means the intensity value of pixel P):

$$BLTP_{formula\ 1} = S\big(I(P_i), I(P), I(P_{i+4})\big)2^i \quad (1)$$

$$S(I(P_i), I(P), I(P_{i+4})) = \begin{cases} 1 & \big(I(P_i) \geq I(P) \cap I(P) \geq I(P_{i+4})\big) \cup \big(I(P_{i+4}) \geq I(P) \cap I(P) \geq I(P_i)\big) \\ 0 & \text{otherwise} \end{cases}$$

Fig 2 shows the difference of structure between CS-LBP and the formula adopted in B-LTP. The gradient information is not considered entirely because of the ignorance of the central pixel in CS-LBP. It is also hard to choose an adaptable threshold. Our formula in B-LTP could overcome these drawbacks by associating the relation according to the predefined templates. Furthermore, B-LTP calculates two directions instead of single direction in CS-LBP, which help reflect better local properties of different types of human body.

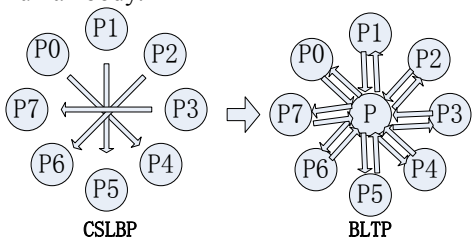

Fig 2 – Structure difference between CS-LBP and the formula one in B-LTP

The second and the third formulas utilize the abilities of HOT and they are given as:

$$BLTP_{formula\ 2} = I(P) \geq I(P_i) \cap I(P) \geq I(P_{i+4}) \quad (2)$$

$$BLTP_{formula\ 3} = \arg\max_k\{I(P)_k + I(P_i)_k + I(P_{i+4})_k\} \quad (3)$$

For each template, if the intensity value of the central pixel is greater than the other two members, it is regarded that pixel P meets this template. Meanwhile, we calculate the sum of intensity values of each template members and choose the greatest one. According to formula (3), the value of $BLTP_{fomula\ 3}$ is from 1 to 4 and is the number of the template wherein it is the greatest sum.

In this way, for the texture information of a certain pixel, we could get a 24-bin length histogram. For the first 16 bins, we could get one value calculated by $BLTP_{fomula\ 1}$; for the following 4 bins, if any one out of four templates is met according to $BLTP_{fomula\ 2}$, the value of the corresponding bin are set as 1; for the last 4 bins, we select the template who has greatest sum and set the bin value as 1 according to the template number. So the texture information for each sub window could be summarized into a 24-bin histogram, as can be seen in Fig 3.

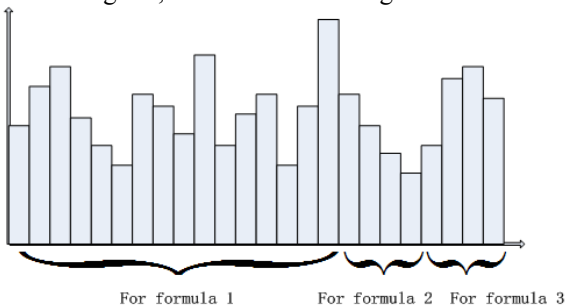

Fig 3 – Example of texture histogram for one sub window

At the same time, for the gradient magnitude information, another three formulas could be designed as:

$$BLTP_{formula\ 4} = S\big(Mag(P_i), Mag(P), Mag(P_{i+4})\big)2^i \quad (4)$$

$$BLTP_{formula\ 5} = Mag(P) \geq Mag(P_i) \cap Mag(P) \geq Mag(P_{i+4}) \quad (5)$$

$$BLTP_{formula\ 6} = \arg\max_k\{Mag(P)_k + Mag(P_i)_k + Mag(P_{i+4})_k\} \quad (6)$$

Similarly, another 24-bin histogram is applied to the gradient magnitude information. Together with the previous texture information, a 48-bin histogram is used as the final feature.

Compared with HOT feature, our feature takes full scaled use of the most important four templates so that not only the relationships between three pixels according to the template, but also information of two directions are calculated in our feature. Meanwhile, the formula one considers the entire orientation information which makes it more discriminative. The B-LTP feature perfectly combines the CS-LBP with HOT and inherits their advantages. Better than HOG, HOT or LBP, not only the gradient and texture information, but also the bidirectional information is taken into account, thus the detection performance is further improved. Moreover, by using larger block as basic unit, the feature dimension for a 64x128 image could be reduced a lot, this means that our feature could be very fast and experimentally confirms this fact.

3. EXPERIMENTAL RESULTS

3.1. Training Method

As we mentioned in section 1, the choice of classifiers is very important and will significantly affect the detection result. Here we choose SVM method to train the detector. SVM is effective for learning with small sampling in high-dimensional space. The decision rule is given by the following formula:

$$f(x) = \sum_{i=1}^{N_s} \beta_i K(x_i, x) + b \quad (7)$$

Where x_i are supported by support vector, N_s is the number of support vectors, $K(x, y)$ is the kernel function.

In our experiments, LibSVM [14] is used, all the training parameters are set as the default values and RBF and linear kernel functions are used in our experiment respectively.

3.2. Dataset

Fig 4 – Samples from INRIA dataset

We utilize the training and testing dataset from INRIA dataset [15] to evaluate our result. Since this dataset is widely used for human detection, we could get some fair comparisons with other methods. This dataset contains 1774 human annotations and 1671 person free images. Meanwhile, 1208 human annotations and 1218 non-human images are used for training the detectors and the rest image are used for testing. For positive images, left-right reflections are also used, so 2416 human images are applied to training. This dataset include various kinds of samples from different point of view, colorful clothing, diverse pose, partial occlusion and sundry illumination. (See Fig 4), so it is suitable as a benchmark database for pedestrian detection.

3.3. Evaluations

In our work, several experiments are adopted in order to evaluate our results and show the promising point of our proposal.

3.3.1. Comparisons with other features

In the first experiment, we compare our B-LTP feature with HOT, HOG, COV and CS-LBP/LTP. Both the feature extraction framework and training dataset are same so that the comparison is fair. Similar with HOG and HOT, we select 16568 negative samples in the resample stage and normalization method is also applied. Mention that in this experiment, the length of our feature for a 64x128 image is 1008, while HOG is 3780, HOT is 3360 and CS-LBP/LTP is 2720. It is also shorter than COV since COV uses variable block size. Short length represents less workload and memory cost. This is because that the size of block in B-LTP is 32x32 and the stride between two blocks is 16. The details of different feature are shown in Table 1.

Method	HOG	COV	HOT	CSLBP/LTP	B-LTP
Block size	16x16	Variable	16x16	Pyramid	32x32
Stride	8	N	8	Pyramid	16
Dimension	3780	Variable	3320	2720	1008
Classifier	SVM	Logitboost	SVM	HK-SVM	SVM
Normalization	Y	Y	Y	Y	Y

Table 1 –Configurations of different features.

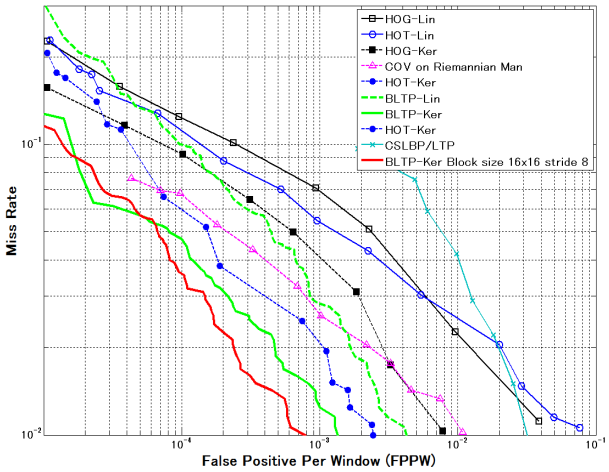

Fig 5 – Performance comparisons

We compare the miss rate at 10^{-4} FPPW which is a representative region that other papers always use [3]. From Fig 5, it can be seen that our proposed B-LTP achieve the highest detection rate compared with HOG [3], COV [4], HOT [1] and CSLBP/LTP [9]. Here Lin means linear kernel functions and Ker means RBF kernel function as other paper [1] [3], so the results are under same SVM training methods. We also use the block size 16x16 and stride size 8 which are the same as HOG and HOT to evaluate out feature. The curve shows that the result is better, but not too much. Thereby, since the dimension length will increase a lot due to the smaller block, it is not necessary to choose the smaller size.

3.3.2. Comparisons between different normalizations

In this second experiment, two different normalization methods together with no normalization are applied for comparisons. The models are defined as:

$$\text{L1: } v = v/(\|v\|_1 + \xi) \quad (8)$$

$$\text{L2: } v/\sqrt{\|v\|_2^2 + \xi^2} \quad (9)$$

In Fig 6, the initial training dataset and SVM of RBF kernel are used for training. We can find that L2 normalization is a little bit better, but the performance is not very much in difference. In this way, like HOT, the normalization part is not necessary for our B-LTP. Thus, the computation time and memory cost for feature extraction can be reduced. For fair comparison, we use L2 in the first experiment.

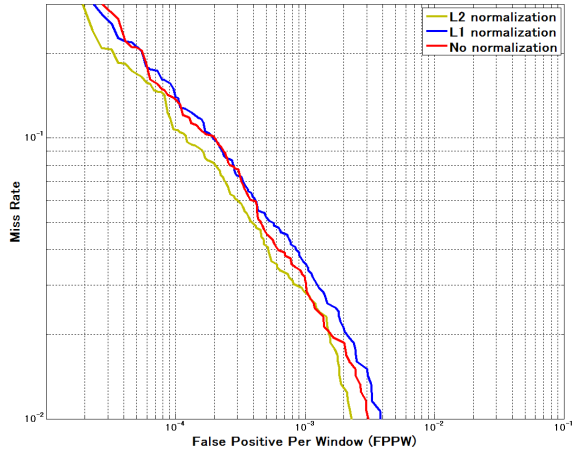

Fig 6 – Normalization method comparisons

3.3.3. Comparisons between different formulas

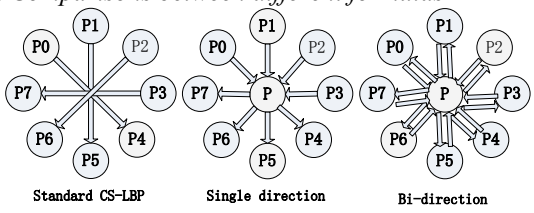

Fig 7 – Different formula structures

For the formula one and four (see (1) and (4)) in our proposed feature, a two-directional-pattern is designed. In this experiment, we would like to compare the result by using different formulas, standard CS-LBP, Single-directional LBP and Bidirectional LBP. (See Fig 7)

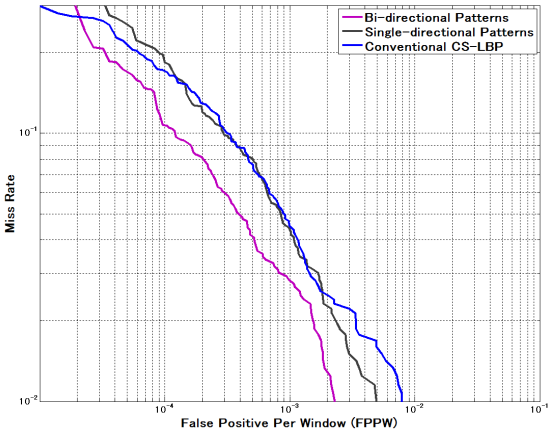

Fig 8 – Performance of different formula

From Fig 8, it can be seen that two directions are better than single and conventional ones. The proposed bidirectional formulas not only take the central pixel into account, but also have the ability of rotational and symmetrical invariance. Moreover, double directions make our feature more descriptive with edge and less sensitive to noise, thus become much more discriminative.

3.3.4. Comparisons between different block size and overlapping rate

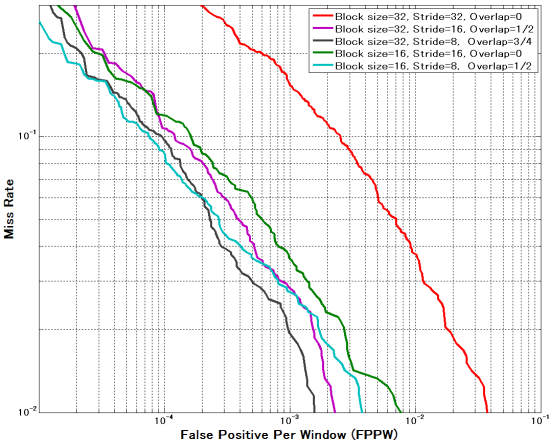

Fig 9 – Performance of different block size and overlapping rate

Take an input image for example, the size of image is first resized to 64x128. The sub-windows (block) are used for detection. Since the size of block is fixed, it is necessary to select suitable size and overlapping rate for training and detection. In the fourth experiment, several different block sizes and overlapping rate are used to evaluate the performance.

Fig 9 indicates that detection rate is increased when we choose smaller block size and higher overlapping rate.

However, compared with the memory consumption and detection workload, the improvement is limited. Hence, block size=32, Stride=16 (1008 dimensions) is relative better than block size=32, Stride=8 (3120 dimensions) and block size=16, Stride=8 (5040 dimensions). It should be noticed that even the length of 3120 dimensions is short than HOG (3780) and HOT (3320). Not to mention that block size=32, Stride=16 (1008 dimensions) is used in the first experiment for comparisons.

3.3.5. Example of detection result

In the last part, several detection results on natural images from INRIA dataset are illustrated in Fig 10 by using the B-LTP detector obtained in the first experiment.

Fig 10 – Examples of detection result

5. CONCLUSION

In this paper, a novel powerful local feature call Bidirectional Local Template Patterns (B-LTP) is proposed for pedestrian detection. Four basic templates are applied and three formulas are used to calculate the gradient and texture information separately. The B-LTP feature takes advantage of the properties of both the CS-LBP and HOT feature so as to inherit the desirable properties of both texture features and gradient based features. Moreover, the direction concept is firstly adopted for object detection, which makes this feature more effective. In addition, since the length of B-LTP is very short, it is computationally cheaper and easier to implement. This point makes it more suitable for real-time systems.

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 Core Research for Evolution Science and Technology (CREST) project, JST.

REFERENCES

[1] S.Tang and S.Goto, "Histogram of Template for Human Detection" In *International Conference on Acoustics, Speech and Signal Processing (ICASSP),* 2010.

[2] M. Heikkila5, M. Pietikainen, and C. Schmid, "Description of interest regions with local binary patterns", *Pattern Recognition*, 2009, 42(3):425-436.

[3] N. Dalal and B. Triggs. "Histograms of oriented gradients for human detection", In *International Conference on Computer Vision and Pattern Recognition* (CVPR), 2005.

[4] T.Oncel and P. Fatih, "Pedestrian detection via classification on Riemannian manifolds," *IEEE Trans. Pattern Anal*. Mach. Intell., vol.30, no.10, pp.1713-1727, Oct. 2008.

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

[6] Z. Lin and L. S. Davis, "A pose-invariant descriptor for human detection and segmentation," in *ECCV, 2008*.

[7] Q.Zhu, S.Avidan, M.C.Yeh, and K.T.Cheng, "Fast human detection using a cascade of histogram of oriented gradients," in CVPR, pp.1491-1498, New York, 2006

[8] D.Lowe, "Distinctive image features from scale-invariant key points," in IJCV, vol.60, pp.91-110, 2004.

[9] Y. Zheng, C. Shen, R.I. Hartley, and X. Huang, "Effective pedestrian detection using center-symmetric Local Binary/Trinary Patterns", in CoRR, 2010.

[10] X.Wang, T.X.Han and S.Yan, "An HOG-LBP Human Detector with Partial Occlusion Handling," in ICCV 2009. pp.32-39, 2009.

[11] B.Scholkopf and A.Smola, Learning with Kernels Support Vector Machines, Regularization, Optimization and Beyond, 2002.

[12] P.Viola and M.Jones, "Rapid object detection using a boosted cascade of simple features," in CVPR, 2001.

[13] T.Ojala, M. Pietikäinen and D.Harwood. "A comparative study of texture measures with classification based on featured distributions," in Pattern Recognition, 29(1):51-59, 1996

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

[15] INRIA Dataset [Online], Avilable: http://lear.inrialpes.fr/data