Discriminative Local Binary Patterns for Human Detection in Personal Album

Yadong Mu[1], Shuicheng Yan[2], Yi Liu[1], Thomas Huang[3], Bingfeng Zhou[1]
[1]Peking University, [2]National University of Singapore, [3]University of Illinois at Urbana-Champaign

Abstract

In recent years, local pattern based object detection and recognition have attracted increasing interest in computer vision research community. However, to our best knowledge no previous work has focused on utilizing local patterns for the task of human detection. In this paper we develop a novel human detection system in personal albums based on LBP (local binary pattern) descriptor. Firstly we review the existing gradient based local features widely used in human detection, analyze their limitations and argue that LBP is more discriminative. Secondly, original LBP descriptor does not suit the human detecting problem well due to its high complexity and lack of semantic consistency, thus we propose two variants of LBP: Semantic-LBP and Fourier-LBP. Carefully designed experiments demonstrate the superiority of LBP over other traditional features for human detection. Especially we adopt a random ensemble algorithm for better comparison between different descriptors. All experiments are conducted on INRIA human database.

1. Introduction

After face detection techniques has become practical, human detection in still images and videos is becoming the focus of the computer vision research community. A robust and fast human detector is especially useful for various vision applications. However, due to high variations of clothing, pose, lighting conditions, and cluttered backgrounds in common personal albums, the task of human detection is rather challenging.

Popular approaches for human detection can be divided into two categories: part based methods or sub-window based methods. The former relies on a human model with geometrical constraints, a more comprehensive introduction for which is referred to [14]. The latter category exhaustively searches every subwindow within an target image, performing a 0/1 testing, i.e. with or without human in it. The most representative work can be found in [5], where overlapped and dense local descriptors based on oriented per-pixel gradients (HOG) are extracted and trained via Support Vector Machine (SVM). Consequent work [20] uses the cascade strategy [15] for acceleration.

Our method belongs to the second category. Roughly speaking, there are two complemental directions for the sub-window based methods: building more discriminative and compact local features ([5], [14]), or developing more powerful learning algorithms beyond traditional AdaBoost ([16]), both of which are still open issues for human detection. In [14], covariance tensor feature (in this paper we will use *COV* for it) based on the correlation between pairwise sub-features was proposed and obtain better results than [5] and [20]. However, it is important to notice that the adopted learning strategies significantly affect the final detecting accuracy as well as the feature type. Since the authors in [14] use logitBoost, another variant of the boosting algorithm, it is difficult to judge whether the proposed COV feature or logitBoost makes more contribution. In this paper we will avoid this ambiguity via a random learning method named *RandomEnsemble* for better comparison between different types of features.

The main contribution of our work is along the first direction, i.e. we focus on building more powerful features for human detection. In recent years, increasing interest is paid on investigating image's local patterns for better detection and recognition. Especially, local patterns that are binarized with adaptive threshold provide state-of-the-art results on various topics, especially on face recognition and face detection ([10],[18]). Here we propose several variants of the original LBP (local binary pattern) feature; these new features can work in perceptually color space and prove more suitable for the human detection task. Based on both theoretic analysis and various experiments, we argue that local color configuration can provide more information than solely intensity gradient. Compared to the gradient based features (Wavelet, HOG and COV), LBP is much more accurate, sparse and easy to be computed.

The paper is organized as follow. In Section 2, we provide an introduction to the LBP-based region descriptor and other related features for human detection. In Section 3, for completeness we briefly review the human detection methodology. Experiments are provided in Section 4.

978-1-4244-2243-2/08/$25.00 ©2008 IEEE

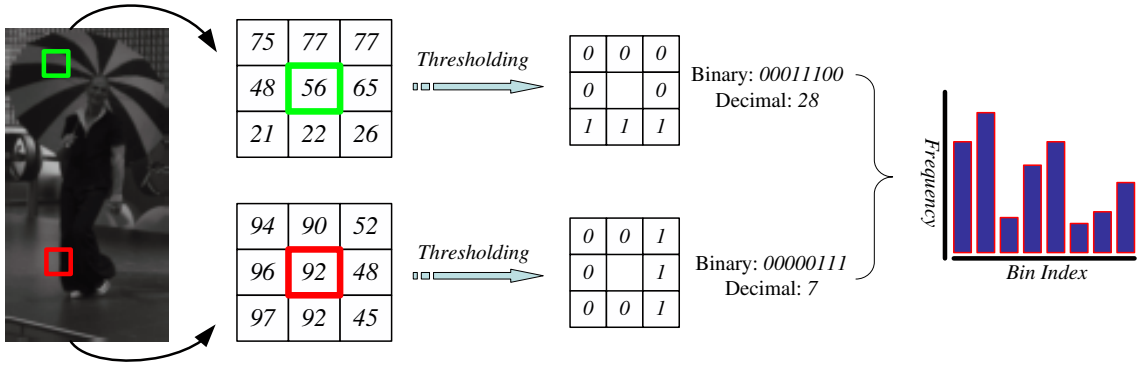


Figure 1. Illustration of LBP. Typically the binary codes obtained by local thresholding are transformed into decimal codes. Note that in this example we use a threshold of 30, which is slightly different from the original LBP. See text for more details.

2. Region Description with LBP

2.1. Basic idea of LBP

The idea of LBP (local binary pattern) is originally proposed by Ojala et al. in [11] for the aim of texture classification, and then extended for various fields, including face recognition ([2]), face detection ([10]), facial expression recognition [19] etc. The most attractive advantages of LBP are its invariance to monotonic gray-scale changes, low computational complexity and convenient multi-scale extension. The philosophy behind LBP is simple and elegant: unify statistical and traditional structural methods.

In Figure 1, we give an illustration for how LBP serves as local descriptor. Each neighbor pixel is compared with the center pixel, and the ones whose intensities exceed the center pixel's are marked as "1", otherwise as "0". In this way we get a simple circular point features consisting of only binary bits. Typically the feature ring is unfolded as a row vector; and then with a binomial weight assigned to each bit, the row vector is transformed into decimal code for further use. For clarity, we adopt the same notation $LBP_{P,R}$ as in [1], where R is the radius of the circle to be sampled (see Figure 2), and P is the number of sampling points. Examples for various choices of these two parameters can be found in Figure 2. It is obvious to see that LBP can be effortlessly extended to the multi-scale case.

Denote the ring feature for image pixel (x, y) as $B(x, y) =< b_{P-1}, \ldots, b_1, b_0 >$, where $b_i \in \{0, 1\}$. It is common to transform $B(x, y)$ into decimal code via binomial weighting:

$$LBP_{P,R}(x, y) = \sum_{i=0}^{P-1} b_i 2^i, \qquad (1)$$

which characterizes image textures over neighborhood of (x, y). And a 1D histogram for an target image region can be built by counting the frequencies of each value of LBP codes, which is finally normalized with L1-norm or L2-norm as image region representation.

An important special case of LBP is the *uniform* LBP. A LBP descriptor is called *uniform* if and only if at most two bitwise transition between 0 and 1 over the circulated binary feature. For example, 00000000 (0 transition), 11100011 (2 transitions) are uniform, while 01010000 (4 transitions), 01110101 (6 transitions) are non-uniform ones. An important observation was made by Ojala et al. [11] that in texture images, majority of LBP features can be categorized to be uniform. In practice, all non-uniform LBP are labeled with a single label, while each uniform LBP is cast into a unique histogram bin according to its decimal value.

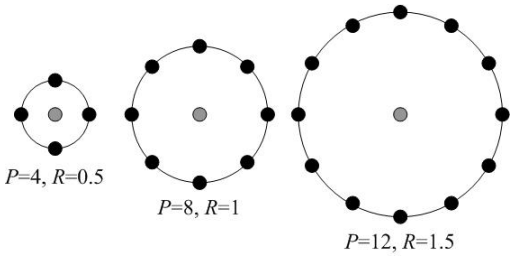


Figure 2. Multi-scale LBP. R: radius of sampling circle. P: number of sampling pixels.

2.2. Related works

Several region descriptors based on simple local features have been extensively used for human detection in still image, we will briefly review them first (we ignore the Harr feature for the consideration of space; a detailed discussion about it can be found in [5]):

HOG (Histogram of Gradient) ([5], [6], [20]) can be regarded as a simplified version of SIFT. It computes intensity gradients from pixel to pixel. For each pixel, it selects cor-

responding histogram bin according to gradient direction, and determines voting strength as proportional to gradient magnitude.

COV (Covariance Tensor Feature) ([14]) assumes the computed tensor descriptors lie on a Riemannian manifold. For each pixel p, a d-dimensional feature vector z_p can be calculated as:

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

where (x, y) are pixel coordinates in the image plan. I_x, I_y, I_{xx} and I_{yy} denote first and second partial intensity derivatives and ϵ is a small constant for numeric consideration. The last term is introduced to retain the orientation information. Denote index set for target region R as $\mathcal{I}_R$, a $d \times d$ covariance matrix can then be computed as:

$$C_R = \frac{1}{|\mathcal{I}_R| - 1} \sum_{i=1}^{|\mathcal{I}_R|} (z_i - \mu)(z_i - \mu)^T, \tag{3}$$

where μ is the statistical mean of z_i. Note that due to the symmetry of C_R, only the upper triangle part need to be stored, thus possible to be unfolded as a $\frac{d \times (d+1)}{2}$ dimensional vector.

2.3. Comments

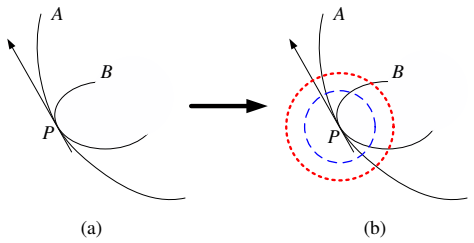


Figure 3. An illustration to highlight the disadvantage of traditional gradient-based methods. At the intersection point P of curve A and B. Solely gradient information can hardly discriminate the two (see (a)). However, an investigation about the local structure around P makes it possible (see (b)). Behaviors along the multi-scale circles (red or blue circles corresponds to different scales) for A and B are distinct.

The most important observation for HOG and COV is that they both work based on first/second image derivatives (including direction and magnitude), which well capture local shape characteristics for the targeted object and can be fast computed via the *Integral Image* tricks [15]. For human detection, the most discriminative human parts - limbs, torsos, maintain a roughly upright pose, while head+neck form a shape similar to "Ω". Against a relative "clean" background, those shapes will dominate the gradient histograms or tensor features.

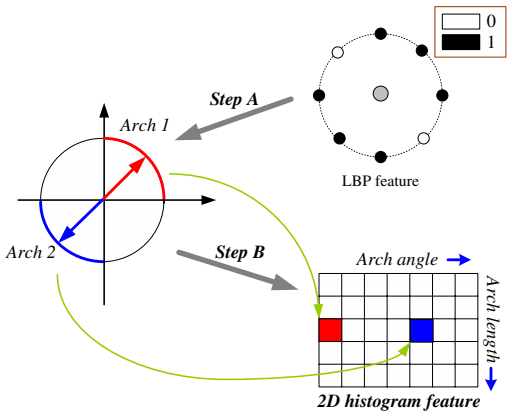
However, the disadvantages of the gradient-based features lie in three-folds. First, gradient sketches the intensity distribution around current pixel in a rather rough style. Two same gradients may correspond to rather different local structures, thus ambiguous. We illustrate this point in Figure 3 by making an analogy between image gradient and curve tangent. As in Figure 3, two curves A and B sharing the same tangent direction may have distinct local distortion. Similarly, neither HOG or COV is able to discriminate two pixels with similar gradients, although in many cases they have totally different local textures. On the contrary, in many contexts such as face recognition, LBP outperforms gradient method because it provides an implicit and approximate representation for *curvatures* on image manifold, which makes it more discriminative.

Secondly, the target object typically appears in a cluttered environment, and the unexpected noises will drastically degrade the performance. Only gradient information is insufficient to judge useful points and outliers. On the contrary, the concept of "uniform LBP" provides the possibility to effectively remove outliers. In fact, previous work by Ojala discovers that over 90% local structures in textured image are uniform in a $LBP_{8,1}$ neighborhood system, while for $LBP_{16,2}$ the percentile is relatively lower yet still more than 70%. Also we can intuitively imagine that a non-uniform LBP descriptor indicates a noisy and less informative local configuration for classifying purpose. For example, highly textured image region produces strong response to gradient filters yet provides little useful information for the detection task, which is labeled as non-uniform and can be safely removed with little information loss.

Thirdly, gradient based methods typically drop color information contained in original images, for it is difficult to define a metric for colors similar to intensity gradient. Fortunately, LBP can intrinsically avoid this issue. As an extension to naive LBP discussed in Section 2.1, we can utilize the well-defined metric in any perceptually color space to measure the dissimilarity of two adjacent pixels, rather than calculating intensity difference, and perform binarization according to locally adaptive thresholds (as in Figure 1). We propose two variants of original LBP descriptors to overcome all above-mentioned problems.

2.4. Our proposed LBP descriptors

As argued above, LBP has advantages over other features for the applications like human detection. However, previous LBP operator in [1] does not suit the human detection problem well, thus we propose two variants of LBP, named S-LBP and F-LBP respectively.



Step A: Calculate principle directions and lengths for each arch.
Step B: Vote for corresponding histogram bins.

Figure 4. Computing S-LBP. Note that the ring feature has two segments of arches, thus a non-uniform one will be abandoned in practice. See text for more details. This figure is best viewed in color.

2.4.1 S-LBP (*Semantic* LBP)

The uniform LBP descriptor defined in Section 2.1 has a space complexity on order of $\mathcal{O}(P^2)$, i.e. the histogram for $LBP_{P,R}$ has about $P(1-P)$ bins, which increases rapidly with P. For example, $LBP_{8,1}$ needs 56 bins, $LBP_{12,1.5}$ needs 132 bins and $LBP_{16,2}$ needs 240 bins. Although longer feature vector is more sparse and discriminative, it has huge storage requirement on the other hand. For human detection, typically in each iteration there are $> 10K$ training samples and approximately 200 candidate subwindows; even a 64 dimensional feature representation in double floating precision will takes roughly $1.2G$ memory. A novel LBP representation that can be intuitively understood and flexibly controlled will be more favorable.

Moreover, for the histogram built according to binary ring feature's decimal codes, there is no guarantee that semantically similar features must fall into spatially nearby histogram bins. For example, 10000111 and 00001111 are similar since the latter actually differs from the former by $45°$ (remember LBP feature has a topology of ring). However, their decimal codes have a large distance (135 v.s. 15). Also, in many scenarios, low-pass filtering for histogram is needed to mitigate image aliasing, while original LBP fails for it.

To attack above issues, we propose the Semantic-LBP (S-LBP). Instead of decimal coding, we redefine LBP based on the following geometrical interpretation: several continuous "l" bits form an arch on the sampling circle, which can be compactly represented with its principle direction and arch length. See Figure 4 and 5 for illustrations. The new representation has a space complexity of $|\alpha| \times |l|$ (i.e. the product of quantized arch angle/length bin numbers), which is unrelated to P and can be much easily controlled.

In practice, first we perform the binarizing on color space such as CIE-LAB. Neighbors whose distances to the central pixel exceed local threshold are marked as "1", else "0". After that we count the number of arches, and non-uniform ones (i.e. having more than one arches) are abandoned. 2D histogram descriptor for any image region can be obtained by collecting information from all its inner pixels. Finally we perform "matrix-to-vector alignment", concatenate each column of the 2D histogram to get a 1D vector.

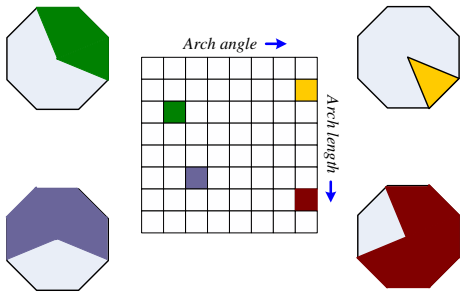


Figure 5. Illustration for the semantic meaning of S-LBP. We draw the corresponding local shapes for four selected histogram bins. This figure is best viewed in color.

2.4.2 F-LBP (*Fourier* LBP)

Sometimes "soft" LBP is useful, i.e. skipping the binariztion step when calculating LBP. Merits of this kind of representation lie in two-folds: first and most importantly, it avoids the potential errors caused by improper local thresholding. Secondly, controllable compression is possible.

We design a soft version of LBP via similar idea to the Fourier boundary descriptor [9]. Let $\mathcal{S} = \{s(k), k = 0 \dots P-1\}$ denote raw feature vector, where $s(k)$ is real-valued color distance between the k-th samples and central pixel. From a signal processing point of view, this soft ring feature can be regarded as an infinite periodic signal. We'd like to transform $\mathcal{S}$ into frequency domain, denoting it as $\mathcal{A} = \{a(u), u = 0 \dots P-1\}$. Coefficients for low frequencies are more useful since they capture salient local structures around current pixel, and lossy compression can be obtained via dropping some highest frequency coefficients, which are supposed to make less contribution for detection and recognition task. For multi-scale LBP with large sampling number P, such compression is quite meaningful. In implementation, we apply one-dimensional DFT (discrete Fourier transform) to $\mathcal{S}$ via the following formula:

$$a(u) = \frac{1}{P} \sum_{k=0}^{P-1} s(k) e^{-j2\pi uk/P} \tag{4}$$

Figure 6. Selected pedestrian images in INRIA human database.

3. Human Detection in Personal Album

We perform all of the experiments on the INRIA human database [5], which is one of the most widely used database for human detection in still images, consisting of thousands of cropped human images in urban scenes. This database contains 2416 human annotations and 1218 non-human images for the training stage, and similar number of samples for testing. Moreover, there are a variety of variations in human pose, clothing, lighting, clutters and occlusions, thus challenging and suitable as a benchmark for comparison between different algorithms and features. Selected images are shown in Figure 6.

Given an image window R (typically of size 128×64 pixels) to be classified, we can extract a large number of subwindows with varying size and position. Some early work of human detection [5] only used subwindows with fixed small size (8×8), while later works ([20],[14] further demonstrate that ensemble of variable-size subwindows can greatly promote detection efficiency. In our experiments, we adopt the variable-size strategy, and sampling subwindows in a similar way to [14]: the minimum subwindows units are sampled with $1/K$ (typically we set $K = 10$) of the width and height of its parent detecting window, and this size is incremented in a step of $1/K$ either horizontally or vertically, or both. Finally we get the set of all valid subwindows $\mathcal{W}_{subwin} = \{r_i\}$. For $K = 10$, the cardinality of $\mathcal{W}_{subwin}$ is 3025. Larger K gives more subwindows while complicates the training of the detector.

4. Evaluations

In this section we present three experiments to evaluate the effectiveness and efficiency of out proposed features, and provide a comprehensive comparison with other popular features. It should be noted that implementing human detecting algorithm is somewhat tricky. Even within the same learning framework such as AdaBoost, small changes for some parameters may bring about twice better or worse results (see [5] for some examples). Thus experiments for comparing different features should avoid as many ad hoc optimizing tricks as possible. To rule out various unrelated factors and highlight the distinctiveness of feature itself, we carefully design the last two experiments. Especially, the learning algorithms there are chosen according to following criterions:

- it should be state-of-the-art.
- suitable for a variety of feature types.
- fewer parameters to be tuned are preferred.

In the first experiment, we compare S-LBP with two other local descriptors: COV [14] and HOG [20], both of which are known to be state-of-the-art and fast to compute. We do not include Shapelet [12] and tree-based approach [16], since these two mainly focus on the learning part and impractical due to their low speed. In practice we adopt cascade-of-rejectgors+logitBoost [8] as our basic learning algorithm and use a L2 normalized 8×8 histogram descriptor. Detection error tradeoff (DET) for all descriptors are plotted in Figure 7. The horizontal coordinate corresponds to FPPW (false positive per window), and vertical coordinate denotes miss rate. Obviously low miss rate together with low FPPW is favorable, in practice we usually seek a good tradeoff between the two. According to our experiments, S-LBP have higher accuracy yet lower speed compared to HOG, while higher speed and comparable accuracy compared to the vectorized COV descriptor. Moreover, for S-LBP, the best sub-window in the first cascade stage is able to reject more than 74% negative samples while keeping 0% miss rate, which is the best reported result in related literatures. The first three cascade levels reject about 90% negative samples, requiring evaluation of about 1.9 LBP descriptors on average. This number for COV is about 2.1, while HOG takes exactly 4.

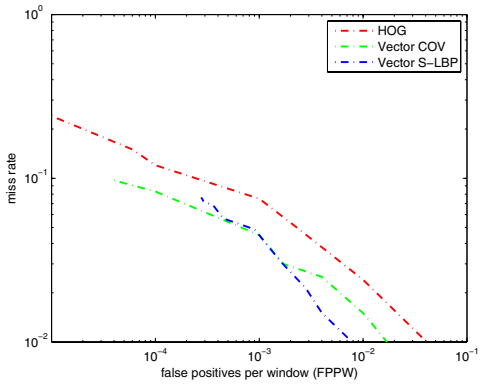


Figure 7. Detection error tradeoff (DET) curves for different descriptors on a log-log scale.

In the second experiment, we inspect the classification score for each individual subwindow for HOG, COV and our LBP-based descriptors. This process can be described as follows: we sample thousands of subwindows $r_i, i = 1...M$ from the whole set $\mathcal{W}_{subwin}$. For each r_i, descriptive vectors f_i^t can be calculated (the index t denotes different feature type, and we assume f^t have a length d_t). Then discriminant analysis is performed in Euclidean space

$\mathcal{R}^{d_t}$ to find the optimal projective direction. An important fact should be pointed out that herein positive and negative samples aren't in same status. Detecting humans in still images is fundamentally a rare-event detection problem. Each positive image contains only one human patch, while from a non-human image we can extract thousands of negative small patches. On the one hand, sample numbers of human/non-human patches are tremendously unbalanced. On the other hand, penalty for misclassifying a positive sample is supposed to be different from that for a negative one. In a word, special techniques should be developed to address this intrinsic asymmetry lying in the data. In practice we adopted the recently proposed LAC (linear asymmetric classifier) [17] to learn a linear separating hyperplane.

Denote data as $X_i = \{x_k^i, k \in \mathcal{I}_i\}$ where $i = 1$ for positive samples, $i = 2$ for the negative; $\mathcal{I}_1, \mathcal{I}_2$ are index sets. Let n_i be the total sample number for class i. Then sample means for each class can be calculated as below:

$$m_i = \frac{1}{n_i} \sum_{k \in \mathcal{I}_i} x_k \tag{5}$$

And the covariance matrices are computed as:

$$\Sigma_i = \frac{1}{n_i - 1} \sum_{k \in \mathcal{I}_i} (x_k - m_i)(x_k - m_i)^T \tag{6}$$

The LAC method extracts optimal discriminant direction w^* by maximizing the following objective:

$$w^* = \arg_w \max J(w) = \frac{w^T (m_1 - m_2)(m_1 - m_2)^T w}{w^T \Sigma_1 w} \tag{7}$$

This objective functional $J(w)$ is similar yet subtly differs from Fisher discriminant analysis (FDA) [7], and actually in a form of generalized Rayleigh quotient, whose optimal solution can be elegantly obtained via eigenvalue decomposition [7], i.e. the optimal direction $w^* = \Sigma_1^{-1}(m_1 - m_2)$. In our experiments we find LAC is more suitable for the asymmetric detection problem compared to other traditional linear discriminant approaches including FDA. More details about LAC can refer to [17].

After finding w^*, all feature vectors are then projected onto it, i.e. $x'_k = (x_k)^T w^*$. Then we can obtain probability distribution over the range of $[\min(x'_k), \max(x'_k)]$. In practice, we build histograms with $N_h = 50$ bins for both classes, and use h_+^j, h_-^j to represent the value of bin j for positive and negative distributions respectively.

Now we can measure the distinctiveness of each feature type based on these histograms. It is known that the classification power of a subwindow varies with its size and position. For each subwindow r_i with feature type t, we adopt Z value [13] to reflect the separability of the projected data:

$$Z^t(r_i) = 2 \sum_{j=1}^{N_h} \sqrt{h_+^j h_-^j} \tag{8}$$

In fact, $Z \in [0, 1]$ is the Bhattacharyya distance between the postive/negative distributions. In ensemble learning algorithms such as AdaBoost [8], Z value is widely used to estimate how discriminative a weak classifier is. A small Z value indicates "good" data distribution for the classification task. We randomly sample thousands of subwindows from $\mathcal{W}_{subwin}$, record their Z values. This process runs for several times and finally we get averaged probabilistic distribution of Z values for each kind of feature. See Figure 9 for the results. It is shown that S-LBP has its peak most near to 0, which demonstrates its superiority over other features. Also, we draw the subwindow with lowest Z value selected by each kind of feature in Figure 8.

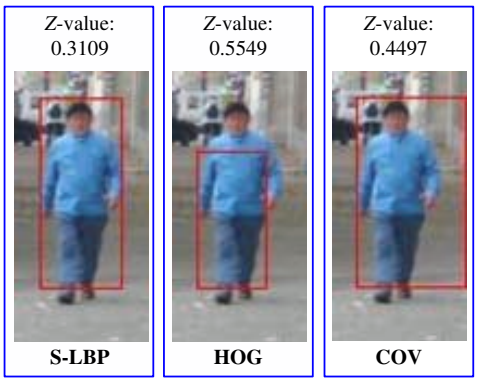


Figure 8. Best subwindow with smallest Z value.

In the third experiment, we aim to compare the discriminative ability for ensemble of several distinct subwindows, which is a next step for experiment two. In previous human detection systems, typically the final classifier is made up of a large number of small weak classifiers. For example, in [20] and [14], the trained classifier consists of hundreds of boosted local classifiers. While in the shapelet method [12], the authors use more than 20,000 weak decision stumps. For comparison between several descriptors, it is a serious issue to optimally select the learning parameters such as the minimum detection rate in each cascade stage, which usually depend on feature type and need be empirically determined (see [3]).

As a results, we adopt a scheme which we called "*RandomEnsemble*" to learn a naive detector, which works as follows: given a target detecting window R, first we randomly sample as many as n_w (set to be 150 in practice) subwindows $\{r_j, j = 1 \dots n_w\}$ from $\mathcal{W}_{subwin}$, compute feature vectors over individual subwindow to get $\{f_j^t, j = 1 \dots n_w\}$ with dimension d_t for feature type t, and con-

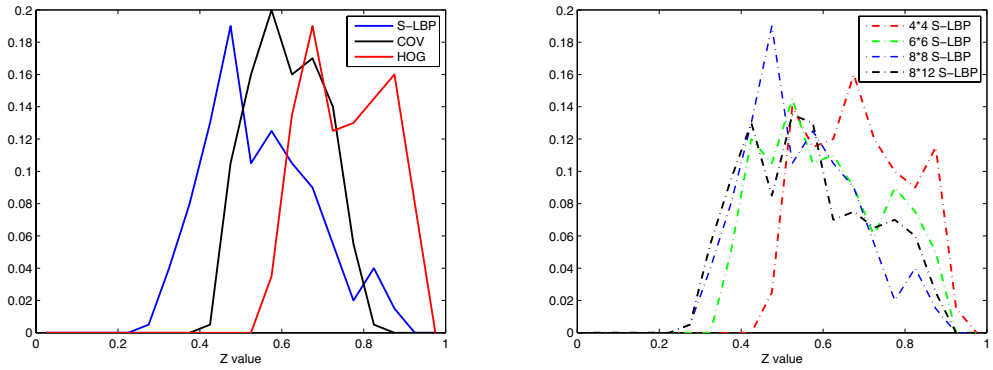


Figure 9. Left: Z value distributions for various descriptors. Right: Z value distribution for different histogram binning of S-LBP.

catenate all these linear vectors into a $d_t \times n_w$ dimensional "supervector" according to fixed order, i.e. $F^t =< f_1^t, f_2^t, \ldots, f_{n_w}^t >$. After that, standard SVM is utilized to learn a separating vector w and bias b as linear classifier. The label y for the testing image window R is determined by:

$$y = \mathrm{Sgn}(F^t(R)^T \cdot w + b), \tag{9}$$

where $\mathrm{Sgn}(x) = 1$ if $x >= 0$, otherwise -1. In implementation we adopt SVM-Torch [4] as the base learner. Moreover, it is possible to incorporate the kernel trick by defining a kernel within two "supervector". For example, the RBF kernel can be defined as:

$$K(F^t(R_1), F^t(R_2)) = \exp(-\frac{\| F^t(R_1) - F^t(R_2) \|^2}{2\sigma^2}) \tag{10}$$

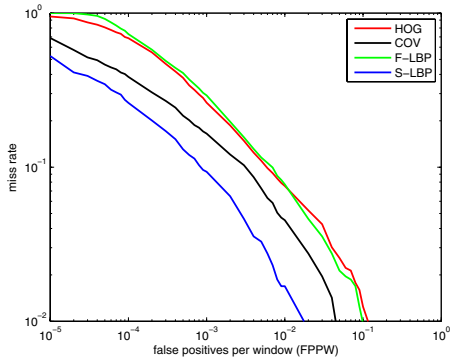


Figure 10. DET curves for human detectors trained through *RandomEnsemble*. Here the S-LBP use L2-norm 8×8 histogram feature, while F-LBP only keeps 62.5% low frequency coefficients. This figure is best viewed in color.

In Figure 10 we plot the DET curves for several aforementioned descriptors with linear kernel. The performance is certainly inferior to boosting-based method, since subwindows here are randomly selected rather than optimally selected along the objective functional's gradient. Instead, weights for selected subwindows will be optimally adjusted by SVM. Also we do not re-train on the "hardest" samples (i.e. the ones misclassified by initial detector) to get a much stronger detector as in [5]. However, through "*RandomEnsemble*" and all other experimental settings, we carefully rule out unrelated factors while highlighting the impact of choosing different region descriptors, thus obtaining much more reasonable comparisons.

5. Conclusion and Future Work

We have attacked the human detection problem by utilizing LBP as region descriptors. Extensive experiments show that local patterns outperform other gradient-based features, owing to the fact that these binarized patterns capture more about the local structures in image manifold. Incorporating these local features with global statistics is a potential future direction.

Acknowlegement

This work was supported by China NSF grant no.60573149, Beijing NSF grant no.4072013, and partially supported by Singapore IDM grant R-705-000-018-279.

References

[1] T. Ahonen, A. Hadid, and M. Pietikäinen. Face recognition with local binary patterns. In *ECCV (1)*, pages 469–481, 2004. 2, 3

[2] T. Ahonen, A. Hadid, and M. Pietikäinen. Face description with local binary patterns: Application to face recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 28(12):2037–2041, 2006. 2

[3] L. D. Bourdev and J. Brandt. Robust object detection via soft cascade. In *CVPR (2)*, pages 236–243, 2005. 6

Figure 11. First two rows are the selected examples of false negative, and last two rows are false positive.

[4] R. Collobert and S. Bengio. Svmtorch: Support vector machines for large-scale regression problems. *Journal of Machine Learning Research*, 1:143–160, 2001. 7

[5] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR (1)*, pages 886–893, 2005. 1, 2, 5, 7

[6] N. Dalal, B. Triggs, and C. Schmid. Human detection using oriented histograms of flow and appearance. In *ECCV (2)*, pages 428–441, 2006. 2

[7] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. Wiley-Interscience Publication, 2000. 6

[8] J. Friedman, T. Hastie, and R. Tibshirani. Additive logistic regression: a statistical view of boosting. Technical report, Dept. of Statistics, Stanford University Technical Report, 1998. 5, 6

[9] R. C. Gonzalez and R. E. Woods. *Digital Image Processing*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 2001. 4

[10] A. Hadid, M. Pietikäinen, and T. Ahonen. A discriminative feature space for detecting and recognizing faces. In *CVPR (2)*, pages 797–804, 2004. 1, 2

[11] T. Ojala, M. Pietikäinen, and D. Harwood. A comparative study of texture measures with classification based on featured distributions. *Pattern Recognition*, 29(1):51–59, 1996. 2

[12] P. Sabzmeydani and G. Mori. Detecting pedestrians by learning shapelet features. In *CVPR*, 2007. 5, 6

[13] R. E. Schapire and Y. Singer. Improved boosting algorithms using confidence-rated predictions. *Machine Learning*, 37(3):297–336, 1999. 6

[14] O. Tuzel, F. Porikli, and P. Meer. Human detection via classification on riemannian manifolds. In *CVPR*, 2007. 1, 3, 5, 6

[15] P. A. Viola and M. J. Jones. Rapid object detection using a boosted cascade of simple features. In *CVPR (1)*, pages 511–518, 2001. 1, 3

[16] B. Wu and N. Ram. Cluster boosted tree classifier for multi-view, multi-pose object detection. In *ICCV*, 2007. 1, 5

[17] J. Wu, M. D. Mullin, and J. M. Rehg. Linear asymmetric classifier for cascade detectors. In *ICML*, pages 988–995, 2005. 6

[18] W. Zhang, S. Shan, W. Gao, X. Chen, and H. Zhang. Local gabor binary pattern histogram sequence (lgbphs): A novel non-statistical model for face representation and recognition. In *ICCV*, pages 786–791, 2005. 1

[19] G. Zhao and M. Pietikäinen. Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE Trans. Pattern Anal. Mach. Intell.*, 29(6):915–928, 2007. 2

[20] Q. Zhu, M.-C. Yeh, K.-T. Cheng, and S. Avidan. Fast human detection using a cascade of histograms of oriented gradients. In *CVPR (2)*, pages 1491–1498, 2006. 1, 2, 5, 6