An HOG-LBP Human Detector with Partial Occlusion Handling

Xiaoyu Wang* Tony X. Han* Shuicheng Yan†
*Electrical and Computer Engineering Dept.
University of Missouri
Columbia, MO 65211
xw9x9@mizzou.edu hantx@missouri.edu
†Electrical and Computer Engineering Dept.
National University of Singapore
Singapore 117576
eleyans@nus.edu.sg

Abstract

By combining Histograms of Oriented Gradients (HOG) and Local Binary Pattern (LBP) as the feature set, we propose a novel human detection approach capable of handling partial occlusion. Two kinds of detectors, i.e.*, global detector for whole scanning windows and part detectors for local regions, are learned from the training data using linear SVM. For each ambiguous scanning window, we construct an occlusion likelihood map by using the response of each block of the HOG feature to the global detector. The occlusion likelihood map is then segmented by Mean-shift approach. The segmented portion of the window with a majority of negative response is inferred as an occluded region. If partial occlusion is indicated with high likelihood in a certain scanning window, part detectors are applied on the unoccluded regions to achieve the final classification on the current scanning window. With the help of the augmented HOG-LBP feature and the global-part occlusion handling method, we achieve a detection rate of* 91.3% *with FPPW=* 10^{-6}*,* 94.7% *with FPPW=* 10^{-5}*, and* 97.9% *with FPPW=* 10^{-4} *on the INRIA dataset, which, to our best knowledge, is the best human detection performance on the INRIA dataset. The global-part occlusion handling method is further validated using synthesized occlusion data constructed from the INRIA and Pascal dataset.*

1. Introduction

Human detection has very important applications in video surveillance, content-based image/video retrieval, video annotation, and assisted living. However, detecting humans in images/videos is a challenging task owing to their variable appearance and the wide range of poses that they can adopt.

The results of The Pascal Challenge from 2005 to 2008 [12] and the recent research [8, 13, 15, 28, 18, 21] indicate that sliding window classifiers are presently the predominant method being used in object detection, or more specifically, human detection, due to their good performance.

Figure 1. The first row shows ambiguous images in the scanning windows. The second row shows the corresponding segmented occlusion likelihood images. For each segmented region, the negative overall score, *i.e.* the sum of the HOG block responses to the global detector, indicates possible partial occlusion. The first four columns are from the INRIA testing data. The last two columns are samples of our synthesized data with partial occlusion.

For the sliding window detection approach, each image is densely scanned from the top left to the bottom right with rectangular sliding windows (as shown in Figure 1) in different scales. For each sliding window, certain features such as edges, image patches, and wavelet coefficients are extracted and fed to a classifier, which is trained offline using labeled training data. The classifier will classify the sliding windows, which bound a person, as positive samples, and the others as negative samples. Currently, the Support Vector Machine (SVM) and variants of boosted decision trees are two leading classifiers for their good performance and efficiency.

Although preferred for its performance in general, compared to other detectors such as part-based detectors [1, 14, 16, 19, 32], the sliding window approach handles partial occlusions poorly. Because the features inside the scanning window are densely selected, if a portion of the scanning window is occluded, the features corresponding to the occluded area are inherently noisy and will deteriorate the classification result of the whole window. On the other side,

part based detectors [16, 19, 32] can alleviate the occlusion problem to some extent by relying on the unoccluded part to determine the human position.

In order to integrate the advantage of part-based detectors in occlusion handling to the sliding-window detectors, we need to find the occluded regions inside the sliding window when partial occlusion appears. Therefore, we have to answer *two key* questions: 1)*How to decide whether the partial occlusion occurs in a scanning window?* 2)*If there is partial occlusion in the sliding window, how to estimate its location?*

To infer the occluded regions when partial occlusions happen, we propose an approach based on segmenting the "locally distributed" scores of the global classification score inside each sliding window.

Through the study of the classification scores of the linear SVM on the INRIA dataset [8, 9], we found an interesting phenomenon: If a portion of the pedestrian is occluded, the densely extracted blocks of Histograms of Oriented Gradients (HOG) feature [8] in that area uniformly respond to the linear SVM classifier with negative inner products.

This interesting phenomenon leads us to study the cause behind it. The HOG feature of each scanning window is constituted by 105 gradient histograms extracted from $7 \times 15 = 105$ blocks (image patches of 16×16 pixels). By noticing the linearity of the scalar product, the linear SVM score of each scanning window is actually an inner product between the HOG feature (*i.e*. the concatenation of the 105 orientation histograms) and a vector $\mathbf{w}$, which is the weighted sum of all the support vectors learned. (The procedure of distributing the constant bias β to each block is discussed in section 3.3.)

Therefore, the linear SVM score is a sum of 105 linear products between the HOG blocks and the corresponding $\mathbf{w}_i$, $i = 1, \ldots, 105$. In our framework, these 105 linear products are called responses of the HOG blocks. For an ambiguous scanning window, we construct a binary occlusion likelihood image with a resolution of 7×15. The intensity of each pixel in the occlusion likelihood image is the sign of the corresponding block response.

For each sliding window with ambiguous classification score, we can segment out the possible occlusion regions by running image segmentation algorithms on the binary occlusion likelihood image. The mean shift algorithm [4, 5] is applied to segment the binary image for each window. The real-valued response of each block is used as the weighting density of each pixel in the mean shift framework. The segmented regions with a negative overall response are inferred as an occluded region for scanning window. Some examples of the segmented occlusion likelihood image are shown in Figure 1. The negative regions are possible occluded regions.

Once the occluded regions are detected, we minimize the occlusion effects by resorting to a part-based detector on the unoccluded area. (See details in Section 3.3).

The contribution of this paper is three-fold: 1) Through occlusion inference on sliding window classification results, we propose an approach to integrate the advantage of part-based detectors in occlusion handling to the sliding-window detectors; 2) An augmented feature, HOG-LBP, which combines HOG with cell-structured Local Binary Pattern (LBP) [3], is proposed as the feature, based on which the HOG-LBP human detector achieves better performance than all of known state-of-the-art human detectors [8, 28, 18, 34, 25, 27, 20] on INRIA dataset (refer to section 3.1 and section 4 for details). 3) We simplify the trilinear interpolation procedure as a 2D convolution so that it can be integrated to the integral histogram approach, which is essential to the efficiency of sliding window detectors.

2. Related Work

Wu and Nevatia [32, 33] use Bayesian combination to combine the part detectors to get a robust detection in the situation of partial occlusion. They assume the humans walk on a ground plane and the image is captured by a camera looking down to the ground. Stein [26] takes advantage of occlusion boundaries to help high-level reasoning and improve object segmentation. Lin and Tang [6] presents a framework to automatically detect and recover the occluded facial region. Fu *et al*. [23] proposed a detection algorithm based on the occlusion reasoning and partial division block template matching for tracking task.

Mu *et al*. [20] state that traditional LBP operator in [2] does not suit the human detection problem well. We proposed a different cell-structured LBP. The scanning window are divided into non-overlapping cells with the size 16×16. The LBPs extracted from cells are concatenate into a cell-structured LBP, similar to the cell-block structure in [8]. As shown in Figure 6(a) in the experiments section, the detection results based on our cell-structured LBP are much better than [20].

3. Approach

The human detection procedure based on the HOG-LBP feature is shown in Figure 2. Our occlusion handling idea is based on global and part detectors trained using the HOG-LBP feature.

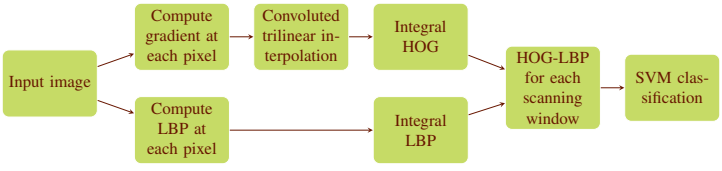


Figure 2. The framework of HOG-LBP detector (without occlusion handling).

3.1. Human Detection using Integrated HOG-LBP

As a dense version of the dominating SIFT [17] feature, HOG [8] has shown great success in object detection and recognition [8, 9, 13, 25, 34]. HOG has been widely accepted as one of the best features to capture the edge or local shape information.

While the LBP operator [22] is an exceptional texture descriptors. It has been widely used in various applications and has achieved very good results in face recognition [3]. The LBP is highly discriminative and its key advantages, namely its invariance to monotonic gray level changes and computational efficiency, make it suitable for demanding image analysis tasks such as human detection.

We propose an augmented feature vector, which combines the HOG feature with the cell-structured LBP feature. HOG performs poorly when the background is cluttered with noisy edges. Local Binary Pattern is complementary in this aspect. It can filter out noises using the concept of uniform pattern [22]. We believe that the appearance of a human can be better captured if we combine both the edge/local shape information and the texture information. As shown in Figure 7 in the experiments section, our conjecture is verified by our experiments on the INRIA dataset.

We follow the procedure in [8] to extract the HOG feature. For the construction of the cell-structured LBP, we directly build pattern histograms in cells. The histograms of the LBP patterns from different cells are then concatenated to describe the texture of the current scanning window. We use the notation $LBP^u_{n,r}$ to denote LBP feature that takes n sample points with radius r, and the number of *0-1* transitions is no more than u. The pattern that satisfies this constraint is called uniform patterns in [22]. For example, the pattern 0010010 is a nonuniform pattern for LBP^2, and is a uniform pattern for LBP^4 because LBP^4 allows four *0-1* transitions. In our approach, different uniform patterns are counted into different bins and all of the nonuniform patterns are voted into one bin.

Using the l_∞ distance to measure the distance to the center pixel, (*i.e.*$d_\infty((x_1, y_1), (x_2, y_2)) = max(|x_1 - x_2|, |y_1 - y_2|)$), we illustrate the $LBP_{8,1}$ feature extraction process in Figure 3.

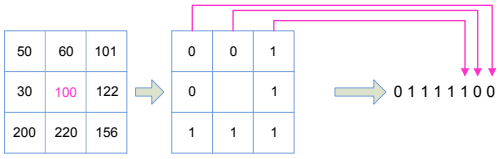


Figure 3. The $LBP_{8,1}$ feature extraction using l_∞ distance.

In our implementation, we use Euclidean distance to measure the distance to achieve better performance. Bilinear interpolation is needed in order to extract the circular local binary patterns from a rectangular lattice. The performance comparison of the cell-structured LBP features with different parameters is shown in Figure 6(a) in the experiments section.

3.2. Integral Histogram Construction with Convoluted Trilinear Interpolation

In spite of its good performance, the approach of sliding window classification is often criticized as being too resource and computationally expensive. The integral image/histogram [29, 24, 34], the efficient subwindow search [15], and the increasingly powerful parallel computing hardware (e.g. GPU and multicore CPU) help to alleviate the speed problem. Within the framework of the integral image/histogram [29, 24, 34], the extraction of the features for scanning windows has a constant complexity $O(c)$ (two vector addition and two vector subtraction). Many state-of-the-art detectors [28, 15, 34, 30, 25] based on sliding window classifiers use the integral image method to increase the running speeds by several folds.

Trilinear interpolation and Gaussian weighting are two important sub-procedures in HOG construction [8]. The naive distribution scheme of the orientation magnitude would cause aliasing effects, both in orientation bin and spatial dimensions. Such aliasing effects can cause sudden changes in the final features which make them not stable enough. For example, if a strong edge pixel is at the boundary of a cell in one image and, due to certain slight changes, it falls into the neighboring cell in another image, the naive voting scheme assigns the pixel's weight to different histogram bins in the two cases. To avoid this problem, we should distribute the effect of the gradient of each pixel to its neighborhood. In our experiments on the INRIA dataset, when FA=10^{-4}, we found that the HOG-LBP detector without the trilinear interpolation has a detection rate 3% lower. The performance of our HOG-LBP detector is not affected by the Gaussian weighting procedure.

It was believed that the trilinear interpolation didn't fit well into integral image approach [34]. While the integrated HOG feature without trilinear interpolation is fast to compute, it is inferior to the original HOG, as mentioned in [34].

In order to take the advantage of the integral image without impairing the performance, we propose an approach, named as *Convoluted Trilinear Interpolation* (CTI), to do the trilinear interpolation [7]. For HOG, the direction of the gradient at each pixel is discretized into 9 bins. So at each pixel, the gradient is a 2D vector with a real-valued magnitude and a discretized direction (9 possible directions uniformly distributed in $[0, \pi)$). During the construction of the integral image of HOG, if we treat the feature value at each pixel as a 2D vector, we won't be able to do the trilinear interpolation between pixels. To conquer this difficulty, we treat the feature value at each pixel as a 9D vector, of which the value at each dimension is the interpolated mag-

nitude value at the corresponding direction. The trilinear interpolation can be done by convolution before constructing the integral image as shown in Figure 4.

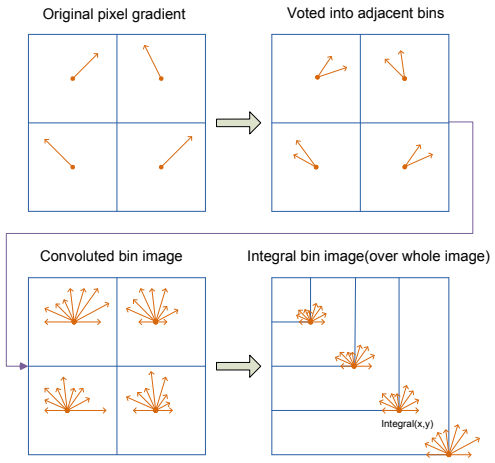


Figure 4. The illustration of the trilinear interpolation in the framework of integral image.

We designed a 7 by 7 convolution kernel to implement the fast trilinear interpolation. The weights are distributed to the neighborhood linearly according to the distances.

$$Conv(k)_{7\times 7} = \frac{1}{256} \times \begin{vmatrix} 1 & 2 & 3 & 4 & 3 & 2 & 1 \\ 2 & 4 & 6 & 8 & 6 & 4 & 2 \\ 3 & 6 & 9 & 12 & 9 & 6 & 3 \\ 4 & 8 & 12 & 16 & 12 & 8 & 4 \\ 3 & 6 & 9 & 12 & 9 & 6 & 3 \\ 2 & 4 & 6 & 8 & 6 & 4 & 2 \\ 1 & 2 & 3 & 4 & 3 & 2 & 1 \end{vmatrix} \quad (1)$$

First, we need to vote the gradient with a real-valued direction between 0 and π into the 9 discrete bins according to its direction and magnitude. Using bilinear interpolation, we distribute the magnitude of the gradient into two adjacent bins(as shown in the top-right subplot of Figure 4).

Then, the kernel in Equation (1) is used to convolve over the orientation bin image to achieve the trilinear interpolation. The intermediate results are the trilinearly interpolated gradient image (bottom-left subplot of Figure 4), ready for integral image construction.

We want to emphasize that the CTI approach doesn't increase the space complexity of the integral image approach. The intermediate trilinear interpolated results can be stored using the space allocated for the integral image. The trilinear interpolated gradient histogram image is of the same size as the integral image. The extra computation time is slim. For each image, it is only a convolution with a 7×7 kernel, which can be further accelerated by Fast Fourier Transform (FFT).

3.3. Combined Global/Part-based Detector for Occlusion Handling

Through the study of the classification scores of the linear SVM classifiers, we found that if a portion of the pedestrian is occluded, the densely extracted blocks of features in that area uniformly respond to the linear SVM classifier with negative inner products. Taking advantage of this henomenon, we propose to use the classification score of each block to infer whether the occlusion occurs and where it occurs. When the occlusion occurs, the part-based detector is triggered to examine the unoccluded portion, as shown in Figure 5. The HOG feature of each scanning window is a 3780 dimensional feature. This 3780 dimensional feature is constituted by the sub-HOG of 105 blocks. The sub-HOG at each block is a 36 dimensional vector denoted as $\mathbf{B}$. The 3780 dimensional HOG feature of each sliding window is: $\mathbf{x} = \begin{pmatrix} \mathbf{B}_1 \\ \vdots \\ \mathbf{B}_{105} \end{pmatrix}$. With its canonical form, the decision function for SVM classifier is:

$$f(\mathbf{x}) = \beta + \sum_{k=1}^{l} \alpha_k \langle \mathbf{x}, \mathbf{x}_k \rangle, \quad (2)$$

where $\mathbf{x}_k$: $k \in \{1, 2, \ldots, l\}$ are the support vectors. If the linear kernel SVM is used here, the inner product $\langle ., . \rangle$ is computed as the scalar product of two vectors in $\mathbb{R}^n$. Taking into account the linearity of the scalar product, we can rewrite the decision function as:

$$f(\mathbf{x}) = \beta + \mathbf{x}^T \cdot \sum_{k=1}^{l} \alpha_k \mathbf{x}_k = \beta + \mathbf{w}^T \cdot \mathbf{x}, \quad (3)$$

where $\mathbf{w}$ is the weighting vector of the linear SVM, *i.e.*, the weighted sum of all the support vectors learned:

$$\mathbf{w} = \sum_{k=1}^{l} \alpha_k \mathbf{x}_k = \begin{pmatrix} \tilde{\mathbf{w}}_1 \\ \vdots \\ \tilde{\mathbf{w}}_{105} \end{pmatrix}. \quad (4)$$

We distribute the constant bias β to each block $\mathbf{B}_i$. Then the real contribution of a block could be got by subtracting the corresponding bias from the summation of feature inner production over this block. That is, to find a set of β_i such that $\beta = \sum_{i=1}^{105} \beta_i$ for the following equation:

$$f(\mathbf{x}) = \beta + \mathbf{w}^T \cdot \mathbf{x} = \sum_{i=1}^{105} \beta_i + \tilde{\mathbf{w}}_i^T \cdot \mathbf{B}_i = \sum_{i=1}^{105} f_i(\mathbf{B}_i). \quad (5)$$

We learn the β_i, *i.e.* the constant bias from the training part of the INRIA dataset by collecting the relative ratio of the bias constant in each block to the total bias constant. Denote the set of HOG features of positive training samples as: $\{\mathbf{x}_p^+\}$ for $p = 1, \ldots, N^+$ (N^+ is the number of positive samples). The set of HOG features of negative samples is: $\{\mathbf{x}_q^-\}$ for $q = 1, \ldots, N^-$ (N^- is the number of negative samples). The ith blocks of $\mathbf{x}_p^+$ and $\mathbf{x}_q^-$ are denoted as $\mathbf{B}_{p;i}^+$

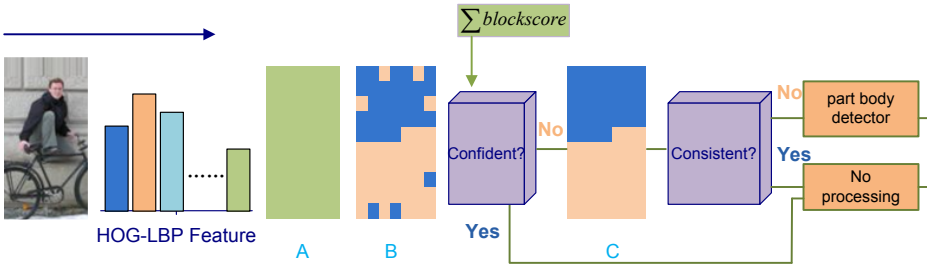


Figure 5. Occlusion reasoning/handling framework. A: block score before distributing bias(summation of SVM classification scores in blocks); B: block score after distributing the bias; C: Segmented region after the mean shift merging.

and $\mathbf{B}^-_{q;i}$, respectively. By summing all the positive and negative classification scores, we have:

$$\sum_{p=1}^{N^+} f(\mathbf{x}^+_p) = S^+ = N^+\beta + \sum_{p=1}^{N^+}\sum_{i=1}^{105} \tilde{\mathbf{w}}_i^T \cdot \mathbf{B}^+_{p;i} \quad (6)$$

$$\sum_{q=1}^{N^-} f(\mathbf{x}^-_q) = S^- = N^-\beta + \sum_{q=1}^{N^-}\sum_{i=1}^{105} \tilde{\mathbf{w}}_i^T \cdot \mathbf{B}^-_{q;i}. \quad (7)$$

Denote $A = -\frac{S^-}{S^+}$. By adding the equations (6) and (7), we have:

$$0 = A{\cdot}N^+\beta + N^-\beta + \sum_{i=1}^{105} \tilde{\mathbf{w}}_i^T \cdot \left(A\sum_{p=1}^{N^+} \mathbf{B}^+_{p;i} + \sum_{q=1}^{N^-} \mathbf{B}^-_{q;i} \right), \quad (8)$$

i.e.,

$$\beta = B \cdot \sum_{i=1}^{105} \tilde{\mathbf{w}}_i^T \cdot \left(A\sum_{p=1}^{N^+} \mathbf{B}^+_{p;i} + \sum_{q=1}^{N^-} \mathbf{B}^-_{q;i} \right), \quad (9)$$

where $B = -\frac{1}{A{\cdot}N^+ + N^-}$. We have:

$$\beta_i = B \cdot \tilde{\mathbf{w}}_i^T \cdot \left(A\sum_{p=1}^{N^+} \mathbf{B}^+_{p;i} + \sum_{q=1}^{N^-} \mathbf{B}^-_{q;i} \right). \quad (10)$$

By Equation (10), we distribute the constant bias β to each block $\mathbf{B}_i$ which translates the decision function of the whole linear SVM to a summation of classification results of each block. This approach of distributing keeps the relative bias ratio across the whole training dataset.

The negative blocks (*i.e.* $f_i(\mathbf{B}_i) < 0$) is, denoted as $\mathbf{B}^-_i$. Similarly we denote positive blocks as $\mathbf{B}^+_i$. If the geometric locations of some negative blocks $\mathbf{B}^-_i$s are close to each other, while other high-confident $\mathbf{B}^+_i$s fall into other neighboring areas of the scanning window, we tend to conclude that this scanning window contains a human, who is partially occluded in the location, where $\mathbf{B}^-_i$s dominate.

We construct the binary occlusion likelihood image according to the response of each block of the HOG feature to the trained linear SVM. The intensity of the occlusion likelihood image is the sign of $f_i(\mathbf{B}_i)$.

For each sliding window with ambiguous classification score (i.e.the score falls in the SVM classification margin [-1, 1]), we can segment out the possible occlusion regions by running image segmentation algorithms on the binary occlusion likelihood image. Each block is treated as a pixel in the binary likelihood image. Positive blocks have the intensity 1 and negative blocks have the intensity -1. The mean shift algorithm [4, 5] is applied to segment this binary image for each sliding window. The absolute value of the real-valued response of each block (*i.e.* $|f_i(\mathbf{B}_i)|$) is used as the weight ω_i in [5]. The binary likelihood image can be then segmented to different regions. A segmented region of the window with an overall negative response is inferred as an occluded region. But if all the segmented regions are consistently negative, we tends to treat the image as a negative image. Some examples of the segmented occlusion likelihood image are shown in Figure 1.

Our experiments on the INRIA dataset show that the approach can detect the occluded region accurately. Based on the localization of the occluded portion, the part detector running on the positive regions will be activated to make more confident decision. The whole framework is shown in Figure 5. In our approach, we train the upper body and lower body detector as part detectors to handle occlusion, combining with the global detecor.

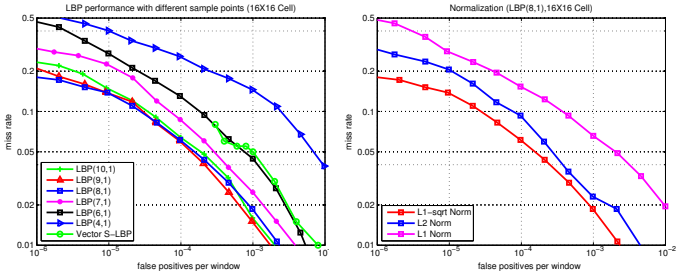
4. Experimental Results

Three groups of experiments are carried to validate our assumptions. We first study the factors affecting the performance of the cell-structured LBP. Comparing to state-of-the art human detectors, the second group of experiments shows the exceptional performance of the convolutional-trilinear-interpolated HOG-LBP feature. Finally, we compare the detection results between the algorithms with and without occlusion handling on both the original INRIA data and the synthesized occlusion data constructed from the INRIA and Pascal dataset.

4.1. Cell-structured LBP detector

We study the effects of different choices of sample points $\{4, 6, 7, 8, 9, 10\}$ and radius $\{1, 2\}$ to the cell structured LBP. Linear SVM is used to train and classify on the INRIA

human dataset. We also compared our cell-structured LBP with S-LBP in [20]. As shown in Figure 6(a), $LBP^2_{8,1}$ performs best. Using $\{4, 6\}$ sample points or radius $\{2\}$ would decrease the performance very much. We also tried LBP features with cell size 8×8, 16×16, 32×32 and find that 16×16 cell works best. This is because the $LBP^2_{8,1}$ patterns of a 8×8 cell are too few to be discriminative and a 32×32 cell introduces too much smoothing over the histogram bins.



(a) (b)

Figure 6. **(a)** The performance comparison of LBP features with different parameters on the INRIA dataset. The LBP with proper parameter setting outperforms vector S-LBP proposed in [20]. The performance of F-LBP in [20] is not available in the normal INRIA training-testing setup. **(b)** The performance comparison for different normalization schemes for LBP feature using $LBP_{8,1}$ with a cell size 16×16.

Choosing a good normalization method is essential for the performance of cell-structured LBP. As shown in Figure 6(b), the L1-sqrt normalization gives the best performance. The L2 normalization decreases the performance by 4% while using the L1 normalization would decrease the performance by 9.5% with a false alarm of 10^{-4}. According to Figure 6(b) and Figure 7, the cell-structured LBP detector has outperformed the traditional HOG detector on INRIA data.

4.2. Detection Results with HOG-LBP Feature

We use augmented HOG-LBP as the feature vector and linear SVM as the classifier for the human detection on the INRIA dataset. We use two different criteria: 1) The detection rate vs. False Positive Per Window (FPPW); and 2) The detection rate vs False Positive Per Image (FPPI). Evaluated using both criteria, our HOG-LBP detector (with/without occlusion handling) out perform all known state-of-the-art detectors [8, 13, 28, 10, 31, 25, 34, 18] on the INRIA dataset. Results are shown in Figure 7[1] and Figure 8. The detector with occlusion handling algorithm is slightly better than the HOG-LBP detector without occlusion handling. The performances of the other algorithms are compared [11].

[1]It has been reported in [11] that the features extracted in [18] contains the boundary of the cropped positive examples, which implicitly encodes the label information.

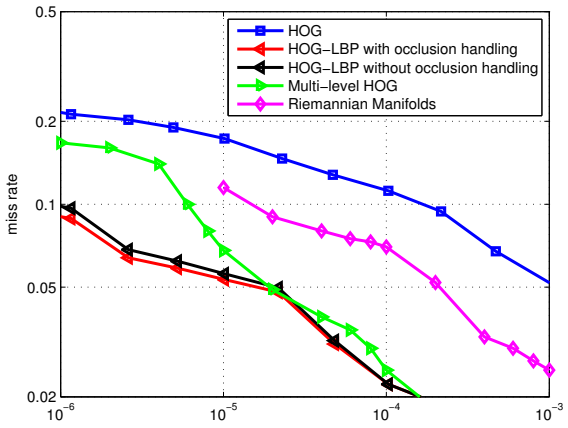


Figure 7. The performance comparison between the proposed human detectors and the state-of-the-art detectors on INRIA dataset using detection(=1−missing rate) VS FPPW. **HOG-LBP with occlusion handling**: The augmented HOG-LBP with Convoluted Trilinear Interpolation.**Multi-Level HOG**[1]: The detector [18] using Multilevel HOG and IKSVM. **Riemannian Manifolds**: The detector [28] based on covariance tensor feature. **Multi-Level HOG** and **Riemannian Manifolds** are the best curves in year 2008 and 2007, respectively.

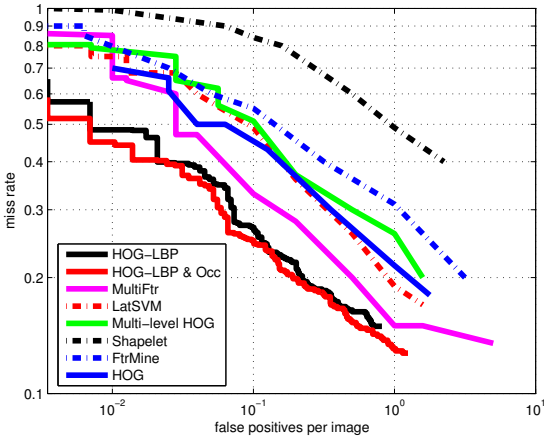


Figure 8. The performance comparison between the proposed human detectors and the state-of-the-art detectors on INRIA dataset using detection(=1−missing rate) VS FPPI. **MultiFtr**: The detector [31] using Shape Context and Haar wavelets feature. **LatSVM**: The detector [13] using deformable model. **Multi-Level HOG**:The detector [18] using Multilevel HOG and IKSVM. **Shapelet**: The detector [25] using shapelet features. **FtrMine**: The detector [10] using Haar features and feature mining algorithm. **HOG-LBP**: Our HOG-LBP detector without occlusion handling. **HOG-LBP & Occ**: Our HOG-LBP detector with occlusion handling.

We achieve a detection rate of 91.3% at 10^{-6} FPPW and 94.7% at 10^{-5} FPPW. The result closest to ours is from Maji *et al.* [18] using Multi-Level HOG and Intersection Kernel SVM (IKSVM). We improve the detection rate by 1.5% at FPPW=10^{-5} and by 8.0% at FPPW=10^{-6}. It is reported in [18] that the Multi-Level HOG can get only

50% detection rate using linear SVM, but it is improved by about 47% at 10^{-4} FPPW [18] by using IKSVM. So it's interesting to see what the detection performance will be by applying IKSVM as the classifier for our feature.

Since we achieved the desired performance on INRIA data (only 25 positive samples are missed out of 1126 testing positive image with the FPPW=10^{-4}), we test the HOG-LBP detector on a very challenging upper body dataset (with 6000 positive samples and 4000 negative images), which is made available to public for download [2] . Our detector gains more than 20% improvement at 10^{-4} compared to the HOG detector as shown in Figure 9.

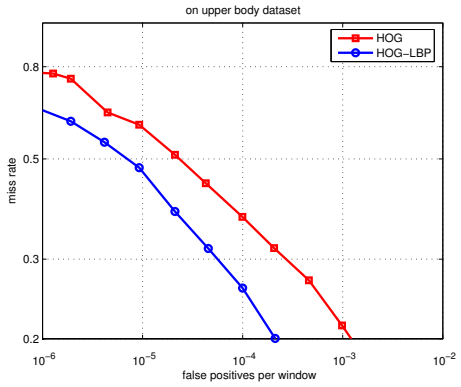


Figure 9. The performance comparison of HOG-LBP and HOG on the NUS upper body dataset.

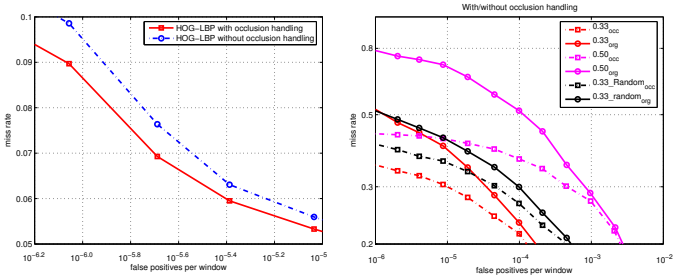
4.3. Experiment on Combined Global/Part-based Detection for Occlusion Handling

As shown in Figure 10(a), our occlusion handling approach improved the detection results. The improvement is less than 1% in detection rate. This is because the INRIA dataset contains very few occluded pedestrians. We save all the miss detection at 10^{-6} FPPW and find that only 28 positvie images are missclassified because of partial occlusion. Our detector picks up 10 of them. Figure 11 shows the samples.

In order to evaluate the proposed occlusion handling approach, we create synthesized data with partial occlusion by overlaying PASCAL segmented objects to the testing images in the INRIA dataset, as shown in Figure 1. First, we just add the objects to the lower part of the human. Then they are added to a random position of the human to simulate various occlusion cases. Objects are resized in order to generate different ratios of occlusion. Three detectors based on the INRIA training dataset are built: the global detector, the upper body detector and the lower body detector.

Following the procedure discussed in section 3.3, we first check the consistency of the segmented binary occlusion likelihood image. A part detector is activated over the positive region when inconsistency is detected. The final decision would be made based on the detector that has the higher confidence. If both detectors are not confidential enough (*i.e.* the classification score is smaller than a threshold, 1.5 for example), we combine global and part detectors by weighting the classification score. We give the score of the global detector a weight 0.7 and 0.3 for part detector in our experiments. The reason that we give part detector a smaller weight is that the global detector and the part detector have different classification margins. In order to keep the consistency of the confidence score, we make the weights proportional to the corresponding classification margins. As shown in Figure 10(b), our method improves the detection results a lot on the synthesized dataset.

[2] http://www.lv-nus.org/NUS-UBD.html



(a) (b)

Figure 10. **(a)** The performance comparison between with and without occlusion handling on the original INRIA dataset. **(b)** The occlusion handling comparison on the synthesized occlusion dataset. 0.33: the occlusion ratio is 0.33; Org: Detection without occlusion handling; OCC: Detection with occlusion handling; Random: testing images are randomly occluded.

Figure 11. Samples of corrected miss detection

5. Conclusion

We propose a human detection approach capable of handling partial occlusion and a feature set that combines the trilinear interpolated HOG with LBP in the framework of integral image. It has been shown in our experiments that the HOG-LBP feature outperforms other state-of-the-art detectors on the INRIA dataset. However, our detector cannot handle the articulated deformation of people, which is the next problem to be tackled.

Acknowledgments

The research was sponsored by the Leonard Wood Institute in cooperation with the U.S. Army Research Laboratory and was accomplished under Cooperative Agreement Number W911NF-07-2-0062. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Leonard

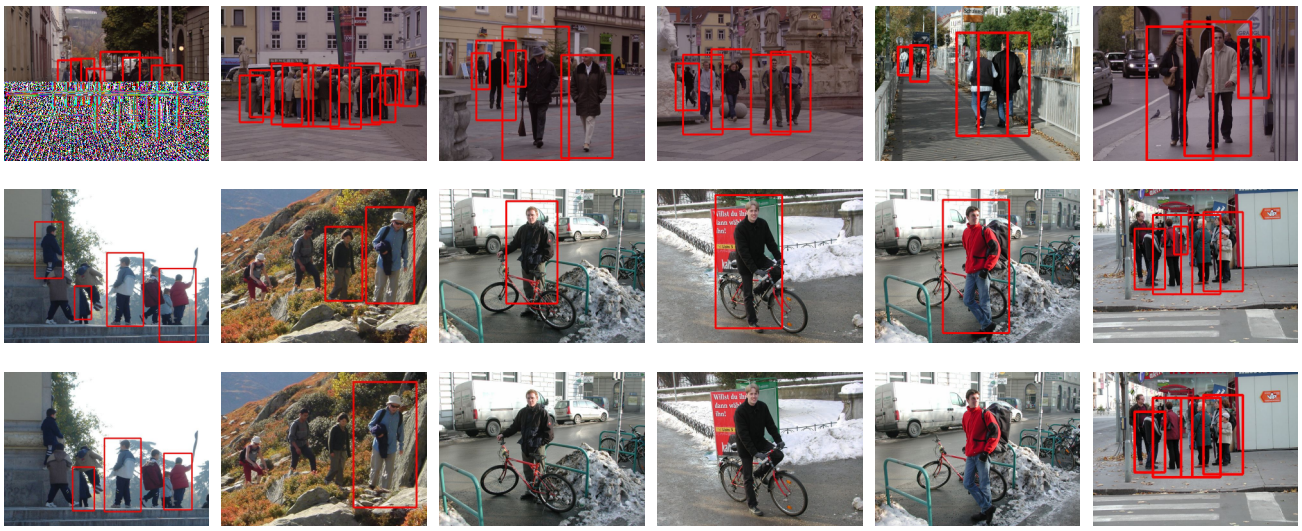


Figure 12. Sample detections on images densely scanned by the HOG-LBP detectors with/without occlusion handling. First row: detected by both. Second Row: detected by the HOG-LBP with occlusion handling. Third row: Missed detection by the HOG-LBP without occlusion handling.

Wood Institute, the Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation heron. Yan is partially supported by NRF/IDM grant NRF2008IDM-IDM004-029.

References

[1] S. Agarwal, A. Awan, and D. Roth. Learning to detect objects in images via a sparse, part-based representation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 26(11):1475–1490, 2004. 1

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

[3] T. Ahonen, A. Hadid, and M. Pietikinen. Face description with local binary patterns: Application to face recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 28(12):2037–2041, 2006. 2, 3

[4] Y. Cheng. Mean shift, mode seeking, and clustering. *IEEE Trans. Pattern Anal. Mach. Intell.*, 17(8):790–799, 1995. 2, 5

[5] D. Comaniciu and P. Meer. Mean shift: A robust approach toward feature space analysis. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24(5):603–619, 2002. 2, 5

[6] L. Dahua and T. Xiaoou. Quality-driven face occlusion detection and recovery. In *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on*, pages 1–7, 2007. 2

[7] N. Dalal. *Finding People in Images and Videos*. PhD thesis, INRIA Rhne-Alpes, Grenoble, France, 2006. 3

[8] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR 2005*, volume 1, pages 886–893, 2005. 1, 2, 3, 6

[9] N. Dalal, B. Triggs, and C. Schmid. Human detection using oriented histograms of flow and appearance. In *ECCV (2)*, pages 428–441, 2006. 2, 3

[10] P. Dollar, Z. Tu, H. Tao, and S. Belongie. Feature mining for image classification. In *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on*, pages 1–8, June 2007. 6

[11] P. Dollár, C. Wojek, B. Schiele, and P. Perona. Pedestrian detection: A benchmark. In *CVPR*, June 2009. 6

[12] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2008 (VOC2008) Results. 1

[13] P. Felzenszwalb, D. McAllester, and D. Ramanan. A discriminatively trained, multiscale, deformable part model. In *CVPR*, 2008. 1, 3, 6

[14] R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. *CVPR*, 02:264, 2003. 1

[15] C. H. Lampert, M. B. Blaschko, and T. Hofmann. Beyond sliding windows: Object localization by efficient subwindow search. In *CVPR*, 2008. 1, 3

[16] B. Leibe, E. Seemann, and B. Schiele. Pedestrian detection in crowded scenes. In *CVPR*, pages 878–885, 2005. 1, 2

[17] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60:91–110, 2004. 3

[18] S. Maji, A. Berg, and J. Malik. Classification using intersection kernel support vector machines is efficient. In *CVPR*, June 2008. 1, 2, 6, 7

[19] K. Mikolajczyk, C. Schmid, and A. Zisserman. Human detection based on a probabilistic assembly of robust part detectors. In *ECCV*, 2004. 1, 2

[20] Y. Mu, S. Yan, Y. Liu, T. Huang, and B. Zhou. Discriminative local binary patterns for human detection in personal album. In *CVPR*, 2008. 2, 6

[21] S. Munder and D. Gavrila. An experimental study on pedestrian classification. *IEEE Trans. Pattern Anal. Mach. Intell.*, 28(11):1863–1868, Nov. 2006. 1

[22] T. Ojala, M. Pietikinen, and D. Harwood. A comparative study of texture measures with classification based on feature distributions. *Pattern Recognition*, 29(1):51–59, 1998. 3

[23] F. Ping, L. Weijia, X. Dingyu, X. Xinhe, and G. Daoxiang. Research on occlusion in the multiple vehicle detecting and tracking system. In *Intelligent Control and Automation, 2006. WCICA 2006. The Sixth World Congress on*, volume 2, pages 10430–10434, 2006. 2

[24] F. Porikli. Integral histogram: A fast way to extract histograms in cartesian spaces. *CVPR*, 2005. 3

[25] P. Sabzmeydani and G. Mori. Detecting pedestrians by learning shapelet features. In *CVPR*, pages 1–8, 2007. 2, 3, 6

[26] A. N. Stein. Occlusion boundaries: Low-level detection to high-level reasoning. *Thesis*, 2008. 2

[27] D. Tran and D. Forsyth. Configuration estimates improve pedestrian finding. In *Advances in Neural Information Processing Systems 20*, pages 1529–1536. MIT Press, Cambridge, MA, 2008. 2

[28] O. Tuzel, F. Porikli, and P. Meer. Human detection via classification on riemannian manifolds. In *CVPR*, pages 1–8, 2007. 1, 2, 3, 6

[29] P. Viola and M. Jones. Robust real-time object detection. *IJCV*, 2001. 3

[30] P. Viola, M. J. Jones, and D. Snow. Detecting pedestrians using patterns of motion and appearance. In *ICCV*, pages 734–741 vol.2, 2003. 3

[31] C. Wojek and B. Schiele. A performance evaluation of single and multi-feature people detection. In *Proceedings of the 30th DAGM symposium on Pattern Recognition*, pages 82–91, Berlin, Heidelberg, 2008. Springer-Verlag. 6

[32] B. Wu and R. Nevatia. Detection of multiple, partially occluded humans in a single image by bayesian combination of edgelet part detectors. In *ICCV 2005*, volume 1, pages 90–97, 2005. 1, 2

[33] B. Wu and R. Nevatia. Tracking of multiple, partially occluded humans based on static body part detection. In *CVPR*, pages 951–958, 2006. 2

[34] Q. Zhu, M.-C. Yeh, K.-T. Cheng, and S. Avidan. Fast human detection using a cascade of histograms of oriented gradients. In *CVPR*, pages 1491–1498, 2006. 2, 3, 6