A new pedestrian detection method based on combined HOG and LSS features

Shihong Yao a, Shaoming Pan a, Tao Wang a, Chunhou Zheng b, Weiming Shen a, Yanwen Chong a,*

a State Key Laboratory for Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, 129 Luoyu Road, Wuhan 430079, China
b College of Electrical Engineering and Automation, Anhui University, Hefei 230039, Anhui, China

1. Introduction

Pedestrian detection is a critical issue in computer vision, with several feature descriptors can be adopted. Since the ability of various kinds of feature descriptor is different in pedestrian detection and there is no basis in feature selection, we analyze the commonly used features in theory and compare them in experiments. It is desired to find a new feature with the strongest description ability from their pair-wise combinations. In experiments, INRIA database and Daimler database are adopted as the training and testing set. By theoretic analysis, we find the HOG-LSS combined feature have more comprehensive description ability. At first, Adaboost is regarded as classifier and the experimental results show that the description ability of the new combination features is improved on the basis of the single feature and HOG-LSS combined feature has the strongest description ability. For further verifying this conclusion, SVM classifier is used in the experiment. The detection performance is evaluated by miss rate, the false positives per window, and the false positives per image. The results of these indicators further prove that description ability of HOG-LSS feature is better than other combination of these features.

© 2014 Elsevier B.V. All rights reserved.
one. Including co-occurrence with various positional offsets, the feature descriptors can express complex shapes of objects with local and global distributions of gradient orientations. CoHOG can express more various shapes than HOG, which uses single gradient orientation. Because the rectangle detection window cannot handle rotation transformation and the pedestrian must be in upright pose due to the limitation in geometrical variation, Panchit et al. [16] suggest to use square-shaped window as the detection window, which can contain more variations of pedestrian. Original LBP descriptor does not suit the pedestrian detecting problem well due to its high complexity and lacking of semantic consistency, so Mu et al. [17] propose two variants of LBP: Semantic-LBP and Fourier-LBP. Liu et al. [18] propose two new texture features called local self-similarities (LSS, C) and fast local self-similarities (FLSS, C) based on Cartesian location grid, which achieve more robust geometric translations invariance and less computing time. These researches show that a single feature has some limitation and deficiency, and seeking feature combination has become a research focus in pedestrian detection. Wang et al. [19] combine the trilinear interpolated HOG with LBP as the feature set. This novel pedestrian detection approach is capable of handling partial occlusion and outperforms other state-of-the-art detectors on the INRIA dataset. Yuan et al. [20] propose a combined method on the basis of Haar and HOG features and also gain a good performance. HOG, shape context and Haar wavelet-based features are combined as a new feature and also obtained a better performance than any of them on pedestrian detection [21]. Walk [22] combined the local color self-similarity and motion features into a whole.

The detection performance of combining features is much better than the single feature’s, but combining these features is not arbitrary. The paper [19,20] are lack of a reliable basis for the feature combination and not comparing with other features from the experiment and theory.

In order to provide the theoretical basis for the selection of feature combination, we suggest to combine any two among these features (SIFT, SURF, Haar, HOG, LBP, LSS) which are most frequently used, then compare and research these combining features in detail. SVM and Adaboost are selected as dominant classifiers in experiments. It’s expected to find a combination feature with the highest detecting precision.

2. Image feature analysis

The selected features (SIFT, SURF, Haar, HOG, LBP, LSS) all have good performance in pedestrian representation. We will extract these six kinds of features from a typical image with size 128 × 64 pixel in experiments.

2.1. SIFT feature

In 1999, Lowe et al. proposed SIFT feature and improve and optimize it in 2004. SIFT feature is a local feature descriptor and use a difference-of-Gaussian function to identify potential interest points that are invariant to scale and orientation. Based on measures of their stability and contrast, the potential interest points are selected again and transformed relative to the assigned orientation, scale, and location for each feature. The local image gradients are measured at the selected scale in the region around each keypoint.

2.2. SURF feature

SURF feature is deduced from SIFT feature. It approximates or even outperforms SIFT feature with respect to repeatability, distinctiveness, and robustness, yet can be computed and compared much faster. The scale changes are achieved by altering the size of the box filter instead of altering the image size. The important speed gain is due to the use of integral images, which drastically reduce the number of operations for simple box convolutions, independent of the chosen scale. They take only three additions and four memory access to calculate the sum of intensities inside a rectangular region of any size.

2.3. Haar feature

Viola et al. proposed a simple rectangular feature which is similar to Haar wavelet. The values of Haar feature are equal to difference between the sum of pixels which lie within the white rectangles and the sum of pixels in the grey rectangles. The template library of Haar feature includes edge template, linear template, center template, and diagonal template etc. The feature template can be set arbitrarily with any size in sub-window. As shown in Fig. 1, we select five Haar templates and each image block contains 5-dimensional Haar features. After the shapes of template are identified, the number of features depends on the size of image and templates.

2.4. HOG feature

Local object appearance and shape can be characterized pretty well by the distribution of local intensity gradients or edge directions, even without precise knowledge of the corresponding gradient or edge positions. HOG feature descriptors used to compute local intensity gradients or edge directions are similar to Histograms of Edge Orientation features and SIFT features.

In practice, the feature extraction process conclude three stage: the first stage applies an optional global image normalisation equalisation that is designed to reduce the effects of external illumination variations and local shadow, then computes first order image gradients in x and y directions and accumulates weighted votes for gradient orientation over spatial blocks. At last use overlapping local contrast normalizations for improved performance and collect HOG descriptors for all blocks over detection window. In our experiments, we use the multi-level HOG feature with six levels, block sizes of 64 × 64, 32 × 32, 32 × 64, 64 × 32, 32 × 32, 16 × 16 and 8 × 8 at levels 1–6, respectively. The features at a level l are weighted by a factor Wl which is equal to the ratio of block size at levels l and block size at level 6. That can obtain a 1575 dimensional vector.

As shown in Fig. 2(a), the image window is divided into small spatial regions (“blocks”), and for each block, a local 9-D histogram of gradient directions or edge orientations over the pixels of the block is accumulated. Thereupon each block has 9 orientation bins and the combined histogram entries form the representation in experiments.

2.5. LBP feature

LBP feature is a general texture description operator for measuring and extracting the local texture information of image. The most attractive advantages of LBP are its invariance to monotonic grayscale changes, low computational complexity and convenient multi-scale extension. At each pixel, LBP can be defined as an ordered set of binary comparisons of pixel intensities between the center pixel and its surrounding pixels.

LBP operator can be expressed as:

$$s(I_c - I_l) = \begin{cases} 1 & I_c - I_l > 0 \\ 0 & I_c - I_l \leq 0 \end{cases}$$

Fig. 1. Template of Haar.
where \( I_i \) and \( I_c \) are the grey value of image and \( P \) is the total number of involved neighbors. LBP operators have three patterns: uniform pattern, rotation invariant pattern, rotation invariant uniform pattern.

Uniform pattern requires that the hopping times between the corresponding circulation binary numbers 0 and 1 are no more than two, thus the number of code in basic LBP operator decrease to \( P(P-1)/2 \) instead of \( 2^P \).

Rotation invariant pattern extract the minimum value from a series of initial definition of LBP values in constantly rotating circular neighborhood.

Rotation invariant uniform pattern is deduced by uniform pattern and rotation invariant pattern. The number of code decrease to \( P+2 \) instead of \( 2^P \).

### 2.6. LSS feature

Determining similarity between object data is necessary in object detection and tracking. LSS features are used in video and image matching. They presented a local self-similarity descriptor which captures internal geometric layouts of local self-similarities within images, while accounting for small local affine deformations. LSS features capture self-similarity of color, edges, repetitive patterns and complex textures in a single unified way. LSS can and should be used globally rather than locally to capture long-range similarities and their spatial arrangements.

As shown in Fig. 2(b), the selected image block is compared with its neighboring blocks in a larger surrounding image region. The resulting distance surface is normalized and projected into the space intervals partitioned by the number of angle intervals and radial intervals. The maximum value in an interval space would be considered as the value of feature. \( \text{varnoise} \) is a constant that corresponds to acceptable photometric variations (in color, illumination or noise), and \( \text{varauto}(q) \) takes into account the patch contrast and its pattern structure, such that sharp edges are more tolerable to pattern variations than smooth patches.

To derive the LSS features \( D_q \) associate with an image pixel \( q \), the surrounding image patches (typically patch size: \( 3 \times 3, 5 \times 5, 7 \times 7 \) and \( 9 \times 9 \)) are compared with a larger surrounding image window centered at \( q \) (typically radius values are 10, 20, 30, 40), using simple sum of square differences (SSD) between patch properties such as pixel intensity and color. The resulting distance surface is normalized and transformed into a "correlation surface". The correlation surface is then transformed into log-polar coordinates centered at \( q \), and partitioned into 80 bins (if we set 20 angles and 4 radial intervals). That can obtain 1920 dimension multi-scale LSS features.

LSS have four primary parameters: the size of image, the radius of window, the interval radius of image patches and angle interval. These parameters are closely associated with each other. When these parameters are fixed, the maximum value can be found directly from the marked circular projection region.
3. Feature combination

Varied feature descriptors have different description ability. Therefore, it is highly recommended to analyze the six features based on their own characteristic. From the theoretical analysis of the six features, optimal descriptor for pedestrian representation is discovered. As shown in Table 1, with the exception of HOG, the other five descriptors require to calculate the difference operator obtained by subtraction between the pixels of image blocks. SIFT, SURF and HOG descriptors are built on the basis of gradient values, and they need to convolve with filters (here, Gaussian function and Box filter are regarded as filter) to achieve the pyramid scale space. Except for Haar descriptor, all the other five descriptors require results projection or histogram calculation. SIFT, SURF and HOG descriptors calculate histograms through the angle interval of pixels, while the LBP and LSS descriptors need to project the results into their own templates.

SURF and SIFT contain all the four operators and have comprehensive image information, but they are highly complex. HOG descriptors contain three operators except difference operator. When HOG combines with Haar, LBP or LSS descriptors, the new combined features include all the four operators. It is believed that human appearance can be better captured if we combine the edge, local shape information and the texture information.

Haar features are called rectangular features. The sums of the pixels in the black rectangles and the white rectangles are calculated separately, and then the subtraction of the sums is calculated. When the smooth patterns correspond to the different background of an image, the Haar features are almost equal to zero and do not have any discrimination. So Haar descriptors are applied only in the images with obvious edge changes. HOG performs poorly when the background is cluttered with noisy edges. LBP and LSS are complementary with HOG in this aspect. As described above, LBP descriptors are obtained by the pixel comparison and LSS descriptors are obtained by the image block comparison. In terms of difference operator in LBP and LSS descriptors, pixel comparison is too fine grained to exactly describe the texture characteristics of image. LSS descriptors are measured within a surrounding image region and HOG is a descriptor which represents local object appearance. The combined feature of HOG and LSS descriptors contains the local information and entire information of images, so HOG combined with LSS descriptors is the focus of our research in this paper.

To verify our analysis result, any two among six features are combined to some new features. The two feature descriptors are denoted as vector A and B and concatenated to a new feature vector \( \{A, B\}\). The six features are normalized with L2-norm respectively before they concatenated. Since SURF features are deduced from SIFT features, we do not concatenate them together in experiments. The proposed approach is described by Fig. 3.

4. Experimental results

We demonstrate our suggestion in this section. The training and testing procedures are conducted on INRIA database whose resolution is 128 × 64 and Daimler database whose resolution is 96 × 48. Images of Daimler dataset are taken from a mobile recording setup. INRIA contains images of high resolution pedestrians collected mostly from holiday photos. INRIA and Daimler dataset are more comprehensive and complex than the older or more limited datasets, such as MIT, NICTA and CVC datasets. Hence we adopt these two databases as experiment data. We randomly select 2300 positive samples of frontal and other views in the two databases, respectively, and randomly select 5000 negatives in INRIA database. INRIAN dataset is downloaded from the website “http://pascal.inrialpes.fr/data/human/”. The negatives patches are cropped to 128 × 64 pixels from the original negatives images in the folders called ‘train_64 × 128_H96’ and “test_64 × 128_H96”. For consistent with INRIA database, we resize the image in Daimler from 96 × 48 to 128 × 64. The adopted datasets are equally split between training set and testing set randomly and we take 10-fold cross-validation in experiments. There are some popular classifiers including neural networks [23–25], SVM and Adaboost. Due to the SVM [26] and Adaboost [27] have theoretical basis, good extensibility and the outstanding performance, we select them as the classifiers and use different criteria: the miss rate (miss rate = 1-detection rate), the false positives per window (FPFW) and the false positives per image (FPPI) to evaluate the description ability of features. In the feature evaluation, Gentle AdaBoost with weak classifier based on decision-tree which contains two branches at most and a SVM with gamma kernel function \(c=2, g=0.0078125\) are used as classifiers.

First, we select AdaBoost as the classifier and conduct the experiments on the two different databases. We plot miss rate versus false positives per window (lower curves indicate better performance) and use log-average miss rate as a common reference value for summarizing performance. Results on INRIA are shown in Fig. 4(a), HOG and LSS feature descriptors have the best performance in single feature experiments, and HOG is slightly better than the LSS. The curves tendency of HOG and LSS are roughly identical. SURF and SIFT are multi-scale features and have better performance in object matching, but only selecting features of keypoints no longer meet the pedestrian detection’s requirements. The constructions of Haar feature are simpler than the other features and its descriptive ability is inferior to HOG, LBP and LSS features.

As the results on Daimler shown in Fig. 4(b), the curve of the LSS features is the lowest among these six kinds of feature and the curve

![Fig. 3. The process of feature combination](image)

<table>
<thead>
<tr>
<th>Operator/Desktops</th>
<th>SIFT</th>
<th>SURF</th>
<th>Haar</th>
<th>HOG</th>
<th>LBP</th>
<th>LSS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Difference</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td></td>
</tr>
<tr>
<td>Convolution</td>
<td>√</td>
<td>√</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Projection (histogram)</td>
<td>√</td>
<td>√</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
of the HOG features is close to it. On the Daimler and INRIA databases, the curves tendency of the features performance is roughly identical and the curve of Daimler is almost under the INRIA because the pedestrians in the INRIA database are much more complex and varied.

On the Comprehensive consideration, we suggest to combine the HOG and LSS feature descriptors as a new descriptor. For certificating our suggestion and evaluating the performance of the combination features, any two among these six kinds of features are combined. Fig. 5(a) and (b) shows the performance comparison among the combination features on INRIA and Daimler database. The curves’ tendency of the detection rate is roughly identical. The performances of combining features are superior to any single feature. The description ability of HOG–LSS feature is the highest among all the feature combinations.

It is insufficient to develop the experiments on Adaboost classifier only, we select the SVM as the other classifier and the experimental results are shown as follows:

Fig. 6(a) and (b) shows the tendency of miss rate versus false positives per window. From these two figures, we can deem that HOG and LSS feature still have the best description ability in SVM classifier. LBP and HOG features are built on pixels and too fine grained. They can’t capture contour and rough information, but LSS features can do this. Global and local features have different effect and are complementary in biological cognitive processes. The conclusion was proved by Lin Chen, published on Visual Cognition and Science [28–30]. As the shown in Fig. 7(a) and (b), the HOG–LSS feature has the best performance and the HOG–LBP feature is inferior to it.

In experiments, we also consider the classification time of these combination features. Fig. 8 shows the classification time on the Adaboost and SVM classifiers. As shown in Fig. 8, the classification time on these two classifiers is almost the same. The classification time is composed of the time cost of patch features computing and the time cost of judging there is a pedestrian or not. Compared with the time cost of patch features computing, the time cost of pedestrian judging is too little and can be neglectable. The classification times prove that the HOG–LSS and HOG–LBP features have less elapsed time than most combinations in this part. The two combinations are more parsimony. Fig. 8 shows that the detection time of HOG–LBP is 58 ms and the HOG–LSS is 63 ms, respectively. They are considered to be the same level. All the experiments are running on Intel(R), Core (TM), dual-core i3 processor, 3.07 GHz, RAM 2.99 G, Matlab2009a.
All of our experiments are not optimized. If we use the GPU acceleration, the speed of the pedestrian detection can be further improved.

We choose five representative pedestrian detection methods, shapelet [10], HOGLbp [19], LatSVM-V2 [31], VJ [32], HOG [6], which are the state-of-the-art pedestrian detection methods, to compare with HOG–LSS feature with MutualCascade Adaboost [33]. The experimental processes refer to bibliography [1]. The experiments in Figs. 4–8 are worked with patches to verify that the HOG–LSS have better pedestrian description ability. In order to compare with other state-of-the-art pedestrian detection methods, we carried out the experiments with images, not patches. We adopt HOG–LSS feature and MutualCascade Adaboost classifier, which use the bootstrap algorithm that has been studied in the paper [33]. MutualCascade improves the classic method of cascade to remove irrelevant and redundant features. The mutual correlation coefficient is utilized as a criterion to determine whether a feature should be chosen or not. In this algorithm, a new updating mechanism for negative samples is utilized to ensure that the number of the negative samples in any level is the same as the original one during training. The updating mechanism is as follows: empty the negative samples, we add not only the negative samples which are false positives in this stage, but also negative samples from the optional negative samples library that are false positives after being detected through all former stages.

Fig. 9(a) and (b) show the performance comparison between HOG–LSS feature and other methods on INRIA database and Daimler database. We take 10/10000 as reference line. Of two databases, the description performance on INRIA is best, which contains high resolution pedestrians, with HOG–LSS feature achieving log-average miss rate of 16–17 percent (see Fig. 9(a)). Performance is also fairly competitive on Daimler database.
high with 20–22 percent log-average miss rate attained by LatSVM-V2. HOGLbp is more challenging, with log-average miss rate of about 43 percent. The log-average miss rate of HOG is 50–52 percent. Shapelet and VJ are not ideal, with log-average miss rates of 89–91 percent and 77 percent, respectively. As shown in Fig. 9(b), LatSVM-V2 has the best performance in these five representative pedestrian detection methods, with log-average miss rates of about 40 percent, and our method achieves log-average miss rates of 25–26 percent. These prove that our method has obvious advantages compared with other methods on two databases.

Fig. 10 presents several detection results. As shown in Fig. 10, it can be seen that most of pedestrian are correctly detected. The detection boxes with different colors in images represent the detection windows with different scales. In Fig. 10(f), the missing pedestrian is the far scale pedestrian with the height of at least 50 pixel. In Fig. 10(g), the missing positive is brought out by partial occlusions. As shown in Fig. 10, the performance of pedestrian detection is excellent in the near and medium scales but degrades seriously in the far scale. In this paper, we choose the detection in the medium scale as our research target, which meets the requirements of real systems.

According to the results of these combination features on different classifiers and overall consideration of the detection rate, the false alarm rate and the classification time, we consider the HOG–LSS feature is the best feature descriptor among these combining features. In theory HOG–LSS feature has a good performance, because it contains these four operators: gradient, difference, convolution, projection, which are complementary with each other.

5. Conclusion

Since the ability of variety of feature descriptor is different in pedestrian detection and a single feature cannot extract the comprehensive information of pedestrian, it is desired to find a new combination feature with the strongest description ability. In this paper, we analyze the six commonly used features: SIFT, SURF, Haar, HOG, LBP and LSS in theory, and consider the HOG–LSS would be a new pedestrian detection feature containing all the image descriptive operators. For further verifying that, we combine any two among the six features on the INRIA and Daimler databases respectively and training them by the different classifiers: SVM and Adaboost. The results of experiments in the performance indicators show that the HOG–LSS feature is superior to other feature combination and expected to replace the HOG–LBP feature commonly used in pedestrian detection research currently.

Acknowledgments

This paper was supported by the Foundation Research Funds for the central Universities (204201kf0242, 204201kf0263), National
Natural Science Foundation of China (41271398), Natural Science Foundation of Hubei Province of China (2012FFB04204), China Postdoctoral Science Foundation (20110491190). Thanks are due to associate professor Tao Wang for assistance with paper polishing.

References


Yanwen Chong received the B.S. degree from the Qufu Normal University, China, in 1995, M.S. and Ph.D. degrees from Wuhan University, China. He is a Associate Professor with the State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing, Wuhan University, Wuhan, China. His research interests include video processing, intelligent transportation system and pattern recognition.