Evaluation of Local Feature Descriptors and Their Combination for Pedestrian Representation

Jixiang Liang$^1$ Qixiang Ye$^1$
$^1$Graduate University of Chinese Academy of Sciences, Beijing 100049, China.
jiaojb@gucas.ac.cn

Jie Chen$^2$ Jianbin Jiao$^1$
$^2$Center for Machine Vision Research
Department of Electrical and Information Engineering, University of Oulu Finland

Abstract

Pedestrian detection problem has been a touchstone of various image feature descriptors. In this paper, we evaluate four kinds of representative local descriptors (HOG, Haar-like, SURF and LBP) for pedestrian representation. Our goal is to find out the best combination of feature descriptors by analyzing and evaluating the complementarities of them. With the cross validation method, we first find out the best descriptor, which is then combined with other descriptors one by one for evaluation. In addition to direct descriptor combination, we propose a new descriptor strategy, called structural combination. Experiments on two public pedestrian datasets show that the performance evaluation can support the complementarily analysis and the complementarities is relevant to combination strategies.

Index Terms—Pedestrian detection, feature representation, feature complementarities

1. Introduction

Object detection in images is a key problem in computer vision. Among various object detection researches, pedestrian detection has been most investigated in recent years because of its potential applications. But it is still challenged in the situations of complex backgrounds and large view/posture variation [1].

In a pedestrian detection system, feature representation is the primary problem being investigated. The computational approach to biological vision proposed by Maar claims that the primitives of visual-information representation are simple components of forms and their local properties. Therefore local features are most often investigated for pedestrian detection. These features include Haar-like features [2], Histogram of Oriented Gradient (HOG) [3], Gabor filter based cortex features[4], Covariance features[5], local binary pattern (LBP) features [6], the HOG-LBP features [7], Edgelet features [8], Shapelet features [9], local receptive field (LRF) features [10], Multi-scale orientation (MSO) features [11], Speed Up Robust Feature (SURF) [12] etc.

Although there are many feature descriptors, their complementarities are not fully investigated. In this paper, we evaluate four kinds of representative local image descriptors (HOG, Haar-like, SURF and LBP) for pedestrian representation. Other feature descriptors are not selected for the reason that either they are similar to any one of the four descriptors or they are classifier-oriented. The selected four feature descriptors are evaluated one by one with linear SVM using cross validation on pedestrian samples. The descriptor of the highest classification accuracy will be selected firstly. Other three descriptors are combined with the best descriptor one by one to find out the best descriptor couple which performs the highest classification accuracy. When implementing feature combination, two strategies are employed. One is direct combination strategy, by which two descriptors will be concatenated to form a feature vector. The other is structural combination, where the dominant features will be concatenated with the SVM scores of the complementary features. And the concatenated feature vector will be used as the input of a linear SVM for classification. The proposed approach is expressed by the following flowchart.

Fig.1.Flowchart of image descriptor evaluation.
The remainder of this paper is organized as follows. In section 2 the feature extraction procedures are described. In section 3 the structural combination strategy is presented. Experiments are presented in section 4, and we conclude the paper in section 5.

\[ \text{Fig. 2. Illustration of feature extraction. (a) is a pedestrian example, (b) denotes the cell grid, (c) is an illustration of HOG feature extraction of a block, (d) Haar-like feature extraction, (e) SURF extraction and (f) LBP extraction.} \]

**2. Feature Extraction and Analysis**

As shown in Fig. 2, when we extract the features for each descriptor, a 64x128 sample is divided into cells with a size of 8x8 pixels and each group of 2x2 cells are integrated into a block in a sliding fashion, and blocks overlap with each other.

**2.1. Extraction of the features**

**HOG features:** We follow the work of Dalal and Trigg to extract the HOG features [3]. In each cell we calculate a 9-dimensional HOG features by calculating the 9-bin histogram of gradient orientations of all pixels in this cell. Each block contains 4 cells, on which 36-dimensional features are extracted. Each sample is represented by 105 blocks, on which 3780-dimensional HOG features are extracted.

**Haar-like features:** Haar-like features are derived from Haar basis functions. More specifically, we use two rectangles, showing in Fig.2d, to extracted features. Each cell is divided into left-right and up-down region and then the Haar-like features (\(D_x, D_y\)) of a cell are extracted as follows.

\[
D_x = \sum_{X\text{left subcell}} I(X) - \sum_{X\text{right subcell}} I(X)
\]

\[
D_y = \sum_{X\text{up subcell}} I(X) - \sum_{X\text{down subcell}} I(X)
\]

where \(I(X)\) is the color value at pixel \(X\). For color image, \((D_x, D_y)\) will be calculated on R, G and B color components respectively and the component of largest absolute value will be used. In each block, the Haar-like features are normalized with L_2-norm. The dimension of Haar-like features is 840.

**SURF:** For the SURF descriptor, we follow the idea of Dalal and Trigg [7]. The SURF descriptor is extracted on the grid cells. As shown in Fig.2e, each cell is divided into four 2x2 sub-cells. Then we calculate Haar-like features \(D_x\) and \(D_y\) in the sub-cells. We then obtain 4 dimensional features \((\sum D_x, \sum D_y, \sum |D_x|, \sum |D_y|)\) for each cell. 16 dimensional features for a block and 1680 dimensional features for a sample.

**LBP features:** Local binary pattern (LBP) is an effective texture descriptor and is widely used in texture classification and object recognition. For the LBP, we follow the idea of Dalal and Trigg [7] to directly extract features in cells, whose sizes are the same with the HOG blocks, as shown in Fig.2f. We denote a LBP feature as \(LBP^{u,r}\), which takes \(n\) sample points with a neighbor region of radius \(r\), and the number of 0-1 transitions is no more than \(u\). As shown in Fig.2f, \(n\) is set to 8 with \(r=1\) pixel since these parameters have the best performance reported by [7]. The pattern that satisfies this constraint is called uniform patterns. For example, the pattern 0010010 is a non-uniform pattern for \(LBP_{u,r}\), and it is a uniform pattern for \(LBP_{u}\) because \(LBP_{u}\) allows four 0-1 transitions. Different uniform patterns are counted into different bins and all of the non-uniform patterns are voted into the last bin. A sample of 105 overlapped cells has a LBP descriptor of 6195 dimension.

**2.2. Analysis of the features**

In Fig.3, four kinds of patterns are used to illustrate the discriminative capability of the four descriptors. The “smooth”, “texture” and “gradual” patterns correspond to the different background of an image. The “sharp” pattern corresponds to the pedestrian-background boundary. It can be seen from Fig.3 that four descriptors have different discriminative capability among the patterns. HOG descriptor can discriminate “smooth” pattern from others but cannot discriminate...
“texture”, “gradual” and “sharp” patterns, showing its disadvantages when used to represent pedestrian objects. Haar-like descriptor cannot discriminate “smooth” pattern from “texture” pattern. It also cannot discriminate “sharp” from “gradual” patterns. SURF cannot discriminate “sharp” from “gradual” pattern, although it is discriminative to other patterns. LBP is discriminative to all patterns, but its dimensionality is the highest.

Fig. 3 Feature values on four patterns. (a) Patterns ("Smooth", “Texture”, “Sharp” and “Gradual”), (b) HOG features value corresponding to patterns of (a), (c) Haar-like feature values, (d) SURF values and (e) LBP feature values.

This shows that only use one kind of descriptor, we cannot represent and discriminate the pedestrians from non-pedestrians well. Therefore, we need to explore the complementarities of the descriptors.

3. Structure Combination of Descriptors

Given two feature descriptors denoted by vector $\mathbf{X}$ and $\mathbf{X}'$, when a direct feature combination strategy is used, two vectors will be concatenated to form a new feature vector $\begin{bmatrix} \mathbf{X} \\ \mathbf{X}' \end{bmatrix}$. When a structural combination strategy is used, we put the first feature vector together with the block score vector of the second feature. Details are as follows.

Let $\mathbf{X}$ denotes the HOG feature vector of a pedestrian sample. This 3780 dimensional vector is constituted by the sub-HOG of 105 blocks as $\begin{bmatrix} x_1 & x_2 & \ldots & x_{105} \end{bmatrix}$. The classification function of the SVM classifier for $\mathbf{X}$ is

$$f(\mathbf{X}) = \mathbf{W}^T \mathbf{X} + b,$$

where $\mathbf{W}$ is the weight vector of the linear SVM and is expressed as $\mathbf{W}^T = (w_1^T, w_2^T, \ldots, w_{105}^T)$. We distribute the constant bias $b$ to each block as [7] does. Then we obtain the responses of 105 blocks, and the responses form a vector denoted by $\mathbf{V}$ as follows:

$$\mathbf{V} = \begin{bmatrix} w_1 & 0 & \cdots & 0 \\ 0 & w_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & w_{105} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_{105} \end{bmatrix} + \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_{105} \end{bmatrix} = \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_{105} \end{bmatrix}. \quad (3)$$

When a structural combination strategy is used, we put $\mathbf{X}$ and $\mathbf{V}$ together to form the new representation $\begin{bmatrix} \mathbf{X} \\ \mathbf{V} \end{bmatrix}$, called local structural feature. Here, $\mathbf{X}$ is a visual feature vector, while $\mathbf{V}$ is the SVM score vector. $\mathbf{X}$ captures the visual contour. $\mathbf{V}$ corresponds to the spatial distributions of block scores, which corresponds to the variation of view or some occlusions, as shown in Fig. 4.

Fig. 4. Feature weights on patterns. (a) and (e) are two samples, (b) and (f) are learned HOG weights corresponding to (a) and (e), (c) and (g) are block score weights of (a) and (e).

The combined vector $\begin{bmatrix} \mathbf{X} \\ \mathbf{V} \end{bmatrix}$ is trained by a linear SVM for classification as

$$g(\tilde{\mathbf{X}}) = \beta^T \begin{bmatrix} \mathbf{X} \\ \mathbf{V} \end{bmatrix} + t$$

where $\beta^T$ is the structural weight vector and $t$ is a threshold.

4. Experiments

Two different pedestrian datasets are used for experiments. One is the INRIA pedestrian dataset [3], which has 2400 training positives of 64×128 pixels. The other is the SDL pedestrian dataset [11], which has 4300 positives of standing humans with frontal and side view. We prepare 4949 negative samples. Some samples are shown as Fig. 5.

Fig. 5. Pedestrian and negative samples. The first row is from INRIA dataset. The second row is from SDL dataset.

In the feature evaluation, a linear SVM with soft margin (C=0.01) is used as classifier. For each experiment, we do five-fold cross validation and the results are shown in Fig. 6, which shows that the HOG descriptor performs the best on both datasets under
single descriptor condition. It also shows that the best descriptor couple is HOG+SURF, when a direct feature combination strategy is employed. Compared with single descriptor the HOG+SURF has a 5% performance improvement on INRIA dataset. As it is more and more difficult to improve the cross validation accuracy when it is close to 100%, 5.0% accuracy improvement is significant. This shows that when we use a direct combination strategy, SURF is significantly complementary to HOG.

When the structural combination strategy is employed, the HOG+LBP_S descriptor couple performance the best on both datasets. Compared with single descriptor, HOG+LBP_S has a 3.5% performance improvement on INRIA dataset, which is also significant. This shows that when we use a structural combination strategy, LBP is significantly complementary to HOG.

Dimensionality of combined features should also be considered. When using a direct combination strategy the dimensions of HOG+Haar-like, HOG+SURF and HOG+LBP are 4620, 5460 and 9975 respectively. When using a structural combination strategy the dimensions of HOG+Haar-like_S, HOG+SURF_S and HOG+LBP_S are all 3885.

![Fig.6. Cross valuation performance. (a) Descriptors, (b) direct combination of descriptors and (c) structural combination of descriptors.](image)

### 5. Conclusions

We evaluate four kinds of most frequently used feature descriptors for object detection. Cross validations on two pedestrian datasets shows that HOG descriptor is most appreciate for pedestrian representation. Two kinds of feature combination strategies are tested. When we use a direct feature combination, HOG+SURF report the best performance. When we use the structural feature combination, HOG+LBP outperform other combinations. This shows that SURF and LBP descriptors are complementary to HOG descriptors, which gives a direction of object representation and detection. In the future, more descriptors and other kinds of objects will be tested.

### 6. References


