Distance-Based Descriptors and Their Application in the Task of Object Detection

Radovan Fusek(✉) and Eduard Sojka

Department of Computer Science, Technical University of Ostrava, FEECS,
17. Listopadu 15, 708 33 Ostrava-Poruba, Czech Republic
{radovan.fusek,eduard.sojka}@vsb.cz

Abstract. In this paper, we propose an efficient and interesting way how to encode the shape of the objects. A lot of state-of-the art descriptors (e.g. HOG, Haar, LBP) are based on the fact that the shape of the objects can be described by brightness differences inside the image. It means that the descriptors encode the gradient or intensity differences inside the image (i.e. edges). In the cases that the edges are very thin, the edge information can be difficult to obtain and the dimensionally of feature vector (without the method for reduction) is typically large and contains redundant information. These ills are motivation for the proposed method in that the edges need not be hit directly; the input brightness function is transformed using the appropriate image distance function. After this transformation, the values of distance function inside objects and backgrounds are different and the values can be used for description of object appearance. We demonstrate the properties of the method for the case of solving the problem of face detection using the classical sliding window technique.

1 Introduction

The detectors that are based on the sliding window technique showed a great performance in the last decade. The main idea behind the sliding window detection technique is based on the fact that the input image is scanned by a rectangular window in different scales. Inside the sliding window the appropriate image descriptors are calculated and composed to the final feature vector. The feature vector is then used as an input for the trainable classifiers (e.g. support vector machine, neural network, random forest). After the classification process, each window is marked as the background or object of interest.

In this area, the three types of features and their modifications that can be used in the sliding window technique became dominant in recent years; HOG, Haar, and LBP features. In [22], the detection framework based on the Haar-like features was presented by Viola and Jones. The framework consists of the image representation called the integral image combined with the rectangular Haar-like features, and AdaBoost algorithm [10]. In [7], the authors proposed the method in that the histograms of oriented gradients (HOG) are used to encode the appearance of the object. Ojala et al. [19] proposed the Local Binary Patterns

X. Jiang et al. (Eds.): GCPR 2014, LNCS 8753, pp. 488–498, 2014.
DOI: 10.1007/978-3-319-11752-2_40

(LBP) in that the local image structures (e.g. lines, edges, spots, and flat areas) can be efficiently encoded by comparing every pixel with its neighboring pixels (more details can be found in Sect. 2).

All mentioned features are based on the fact that the appearance of the objects is described by the image edge information (intensity differences). In general, the features based on the edge information (e.g. length, magnitude, orientation, localization) require large training sets due to their high dimensionality. Additionally, in the cases that the edges are very thin, it is obvious that the edges information is difficult to hit (by the samples). Therefore, the proposed method is based on the distance function in that the information about its changes is not so important. In essence, we divide the image inside the sliding window into the blocks and cells (similarly as in HOG), but instead of the histograms of gradients we encode the values of distance function inside each cell. This leads to the reasonable dimensionality of the feature vector; furthermore, the values of distance function can be easily obtained by sampling. The feature vector that contains the distance function values is then used as an input for the SVM classifier.

2 Related Work

As was mentioned in the previous section, the three types of features are considered as the state-of-the-art in the area of feature based detectors; HOG, Haar, and LBP. In essence, the HOG descriptors can be considered as the dense version of SIFT [17,18]. The sliding window is divided into the cells in that the histograms of oriented gradients are calculated. The cells are normalized across the large blocks. The vector of features that is obtained from each sliding window is then used as input for the SVM classifier. In recent years, many modifications and applications of classical HOG were presented. In [26], the authors proposed the fast way of calculating the HOG features with the use of the integral image. The authors also integrated the HOG features into the Viola and Jones cascade framework. In [5], the authors presented the pyramid of histogram of orientation gradients (PHOG) descriptors in that the HOG descriptors are combined with the image pyramid representation of Lazebnik et al. [13]. In [12], the authors applied the Principal Component Analysis (PCA) to the HOG feature vector to obtain the PCA-HOG vector. The part-based detector based on HOG was proposed by Felzenszwalb et al. in [9].

The Haar-like features which are similar to Haar basis function were proposed by Papageorgiou and Poggio [20] and popularized by Viola and Jones [22]. Viola and Jones combined the Haar-like features with the integral image representation, AdaBoost algorithm, and cascade of classifiers. The extension of the Haar feature set has been presented by Lienhart et al. [16]. For example, the multi-view face detection system was presented by Wu et al. [23], the front-view car and bus detector based on the Haar-like features was proposed in [24].

The LBP operator was proposed by Ojala et al. [19] for the texture analysis, hoverer, the operator was successfully used in many detection and recognition

tasks. In the basic form of LBP operator, every pixel is compared with its neighbors to encode the local image structures such as lines, edges, spots, and flat areas. In [11], LBP were used for face detection problem in low-resolution images. The face recognition problem was solved using LBP in [1,2]. Multi-block Local Binary Patterns (MB-LBP) for face detection were proposed in [25].

Since the geodesic distance is used in the paper, it is important to mention works in that this distance was used in the area of image processing. Image segmentation and object detection methods based on geodesic distance were presented in [3,6,8,21].

3 Proposed Method

The proposed method is based on the fact that the properties of the image (especially the properties of the objects) can effectively be described by the distance function. The goal is to obtain more meaningful values for recognition than the classical state-of-the-art method. The usefulness of this function can be described in the following way.

Suppose the simple theoretical image that contains one object of constant brightness (Fig. 1(a)). The appearance (shape) of this object can be described using the gradient of this object (Fig. 1(b)). In the classical sliding window methods (HOG, Harr, LBP), the samples (e.g. blocks, rectangular features) must hit the places with the intensity differences (edges) to obtain the information about the object. In the situation that edges can be very thin (theoretically infinity thin), it is difficult to hit the places with the edge information, and many samples contain the redundant information without the gradient (edges) information.

Suppose the case that the samples (e.g. blocks, rectangular features) are placed inside the image in the way as is depicted in Fig. 1(c). In such a case, the samples do not detect any important information; the values of gradient sizes and directions are null (HOG principle), as well as the intensity differences inside the samples (Haar principle). This situation was motivation to use another way how to encode the appearance of the objects inside images.

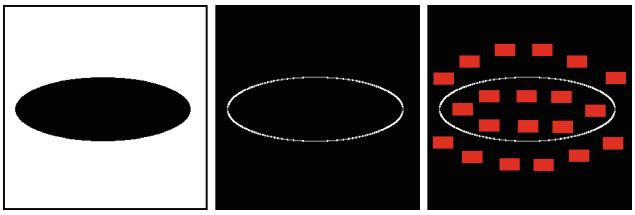
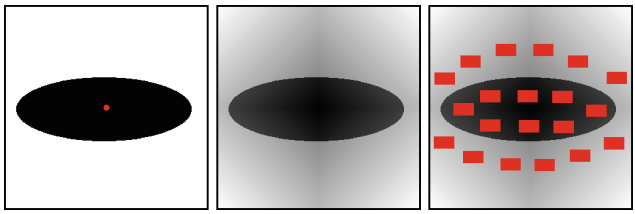


Fig. 1. The image with one object with constant brightness (a). The gradient of the image (b). The samples (red color blocks) in that the information about the object is encoded (c) (Color figure online).



(a) (b) (c)

Fig. 2. The image with one object with constant brightness that contains the centroid point (red color) (a). The visualization of the distance function (b). The values of distance function are depicted by the level of brightness. The samples (red color blocks) in that the information about the object is encoded (c) (Color figure online).

Suppose an arbitrary point that is placed inside the previously mentioned object. Say that in the gravity center of the object (Fig. 2(a)); this point can be called as the centroid of the object c_i. Let us compute the geodesic distance function d from the centroid c_i to all other points inside the image. The visualization of the distance function values is shown in Fig. 2(b). In this particular case, we use the geodesic distance, nevertheless, it is important to note that any appropriately distance function can be used in the proposed detection framework (e.g. resistance distance, diffusion distance). In general, the geodesic distance $d(c_1, c_2)$ between two points c_1, c_2 computes the shortest curve that connects booth points along the image manifold; the geodetic distance reflects the topology of the image.

Suppose the same distribution of the samples as in the previous case (Fig. 2(c)). The main contribution of using the distance function is that the values of this function are different inside and outside the object of interest. In essence, the values of distance function reliably reflect the image information and the appearance of objects, and the meaningful values can be reliably obtained by sampling. Even the simple samples in Fig. 2(c) can be used to describe the properties of the image; the sample values can be used to encode properties (shape) of the objects without a large number of redundant information.

It is clear that the situation is more complicated in the real images and one centroid will not be enough to cover more complicated image structures. Therefore, we divide the whole image into the cells. The gravity centers of each cell are defined as the centroid points c_i; the distance is computed from these points to all other points inside each cell.

The visualizations of geodesic distance values inside the cells of different sizes are shown in Fig. 3. Based on the cell sizes, information with various levels of details is obtained. To compress the information contained in the distance function in to a reasonable number of values, we use four values from each cell only. These values take into account the distance in four different directions (Fig. 4) and the values are then used in the feature vector. Each of four neighboring cells create a block in which into the large blocks (Fig. 4) in that the distance

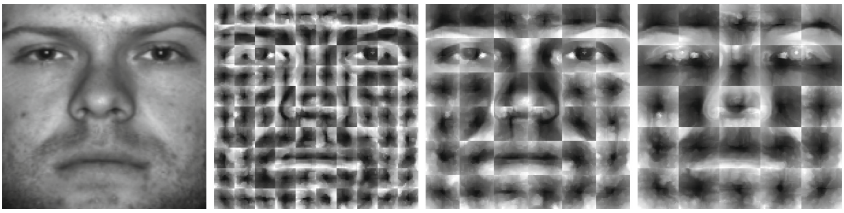


Fig. 3. The visualization of the distance function values inside each cell. The example of face image (a). The sizes of cell 15×15 (b), 25×25 (c), 35×35 (d).

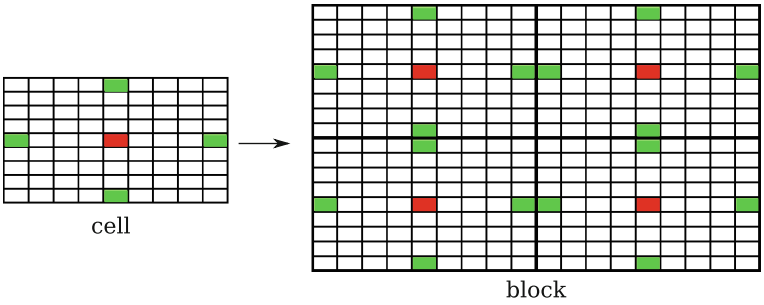


Fig. 4. An example of 9×9 cells that are grouped into one block. In this particular case, from each cell, four values (depicted by green color) are used in the feature vector (Color figure online).

values are normalized. In our experiment, we use the overlapping blocks; the second half of one block correspond with the first half of the next block. The final feature vector is then used as an input for the SVM classifier.

4 Experiments

In this section, we demonstrate the properties of the proposed method for the case of solving the problem of face detection using the classical sliding window technique. For this task we collected 2300 faces and 4300 non-faces. The faces were obtained from the BIOID database combined with the Extended Yale Face Database B [14]. The negative set consists of 3000 images that were obtained from the MIT-CBCL database combined with the 1300 hard negative examples. In the detection process, the sliding window scanned 10 different resolutions of input image. We experimented with many sizes of sliding windows, cells, blocks of the proposed method and we suggested the following configurations.

The configuration $Dist_1$ is designed with the size of sliding window $= 70 \times 70$, size of blocks $= 14 \times 14$, size of cells $= 7 \times 7$, horizontal block step size $= 7$. This configurations consists of 1296 descriptors for one position of sliding window; each window consists of 81 overlapping blocks and each block consists of 4 cells,

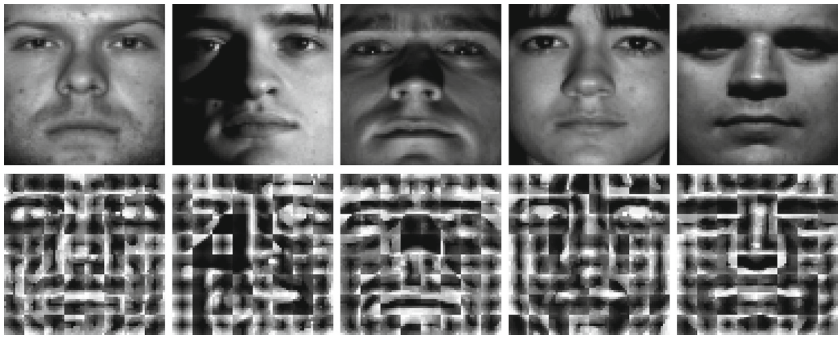


Fig. 5. The visualization of the distance function values inside each cell of $Dist_2$ configuration.

i.e. 324 cells are defined in each window. Finally, each cell is described using 4 distance values, i.e. 1296 descriptors are used (324×4).

The configuration $Dist_2$ is designed with the size of sliding window $= 72 \times 72$, size of blocks $= 18 \times 18$, size of cells $= 9 \times 9$, horizontal block step size $= 9$. This configurations consists of 784 descriptors for one position of sliding window; each window consists of 49 overlapping blocks and each block consists of 4 cells, i.e. 196 cells are defined in each window. Finally, each cell is described using 4 distance values, i.e. 784 descriptors are used (196×4).

The configuration $Dist_3$ is designed with the size of sliding window $= 88 \times 88$, size of blocks $= 22 \times 22$, size of cells $= 11 \times 11$, horizontal block step size $= 11$. This configurations consists of 784 descriptors for one position of sliding window; each window consists of 49 overlapping blocks and each block consists of 4 cells, i.e. 196 cells are defined in each window. Finally, each cell is described using 4 distance values, i.e. 784 descriptors are used (196×4). The examples of visualization of distance function values of training images are shown in Fig. 5.

For comparison, we used the detectors that are based on the HOG features, LBP (Local Binary Patterns) features [15] and Haar features (Viola-Jones detection framework). For the HOG features, we used the classical parameters of HOG; the size of block $= 16 \times 16$, size of cell $= 8 \times 8$, horizontal step size $= 8$, number of bins $= 9$. The training images (for HOG) were resized to the size of 80×80 pixels (the size of sliding window was also set to this size) and the sliding window scanned 10 different resolutions of input image. This configuration of HOG consists of 2916 descriptors for one position of sliding window, and it is denoted as HOG. For the detectors that are based on the Haar and LBP features, the cascade classifiers was created and the training images were resized to the 19×19 pixels for these detectors; the detector based on Haar is denoted as $Haar$, the detector based on LBP is denoted as LBP.

We used the identical training set (2300 positive and 4300 negative samples) and testing set for all detectors. To test the detectors, we collected 300 images from the Faces in the Wild dataset [4]. Before the process of performance

Table 1. The face detection results.

	Precision	Sensitivity	F1 score
$Dist_1$	99.14 %	88.69 %	93.62 %
$Dist_2$	96.05 %	93.83 %	94.93 %
$Dist_3$	98.53 %	86.38 %	92.05 %
HOG	68.85 %	94.71 %	79.73 %
$Haar$	85.88 %	81.28 %	83.52 %
LBP	72.60 %	70.67 %	71.62 %

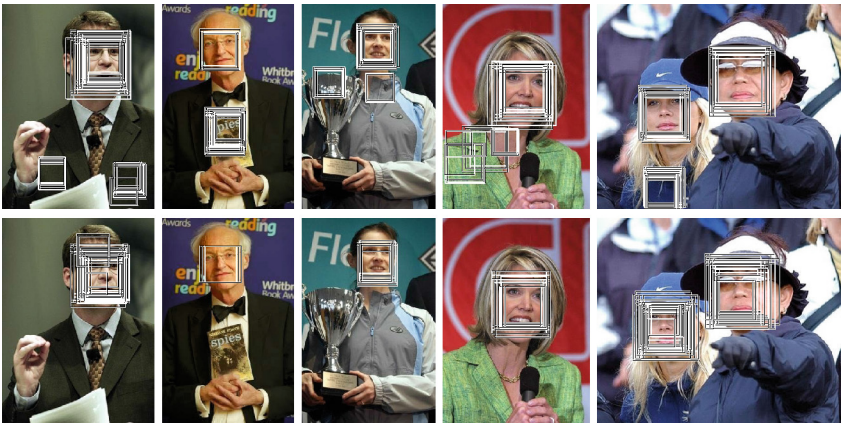


Fig. 6. The differences between the detection results. The first row: the detection results of HOG detector. The second row: the detection results of proposed detector based on the $Dist_2$ configuration. The results are without the postprocessing (the detection results are not merged).

calculation, the positive detections were merged to one if at least 5 positive detections hit approximately one place in the image. In Table 1, the detection results are shown.

The HOG based detector achieved the higher true positives rate (Sensitivity 94.71 %). It means that this detector achieved the large numbers of true positives and the detector had relatively small numbers of false negatives. On the other side, the positive predictive value (Precision 68.85 %) is quite low. It means that the number of false positives is rather large. This is caused by the large dimensionality of feature vector that is created by the 2916 values for one position of sliding window. Since we used the relatively small set of training data (2300 faces and 4300 non-faces) and dimensionality of feature vector of HOG is relatively large, the detector based on HOG detected the faces in the wrong places. Overall detection rate (F1 score) of HOG detector is 79.73 %. This problem also appeared in Haar (F1 score = 83.52 %) and LBP (F1 score = 71.62 %) based

Fig. 7. The face detection results of $Dist_2$ configuration. The results are without the postprocessing (the detection results are not merged).

detector. Although, the Haar based detector achieved the low number of false positives, in general, this detector needs a larger training set; similarly to LBP.

On the other side, the proposed method achieved very promising results. Since the detectors based on the $Dist_2$ and $Dist_3$ created the feature vector of a relatively small size (784), the selected training set (2300 faces and 4300 non-faces) was sufficiently large for them. Even the $Dist_2$ configuration (with 1296 values in feature vector) achieved better results than the state-of-the-art detectors (F1 score = 93.62 %). The best detection results achieved the detector based on the $Dist_2$ configuration (F1 score = 93.62 %). The sensitivity of this detector was lower than in the HOG detector (88.69 % vs. 94.71 %), nevertheless the proposed detector achieved considerably less false positives than the HOG, LBP, and Haar based detectors. The examples of differences between the detection results of the prosed detector and the HOG detector are shown in Fig. 6.

The achieved results confirmed the assumption that the distance function values can be effectively used to create the feature vector. Additionally, the values can be used to describe the information inside the sliding window better than the classical gradient-based approaches with a relatively small set of numbers. Finally, the calculation of geodesic distance and composition of the feature vector take approximately 2 ms for one position of sliding window in $Dist_2$ configuration on an Intel core i3 processor. The detection results of proposed detector based on the $Dist_2$ configuration are shown in Fig. 7.

5 Conclusion

In the paper, we presented an efficient way how the image information can be encoded into the feature vector, which can be used in sliding-window-based techniques of recognition. In essence, the method is based on the idea that the information contained in the window can be expressed by measuring distances along the image manifold. In the method, the sliding window is divided into the cells; inside each cell, a central point is selected. The distances are computed from the central point to all other points inside the cells. We also showed that the proposed method can be used for solving the problem of face detection with promising results and a relatively small size of feature vector. We will try to reduce the vector dimensionality using statistical methods for reducing the dimension of feature vector (e.g. PCA).

In the paper, we used the geodesic distance; however, it is worth mentioning that various distance metrics can be used. We leave experiments with various types of distances for future work.

Acknowledgments. This work was supported by the SGS in VSB Technical University of Ostrava, Czech Republic, under the grant No. SP2014/170.

References

1. Ahonen, T., Hadid, A., Pietikainen, M.: Face description with local binary patterns: application to face recognition. IEEE Trans. Pattern Anal. Mach. Intell. **28**(12), 2037–2041 (2006)
2. Ahonen, T., Hadid, A., Pietikäinen, M.: Face recognition with local binary patterns. In: Pajdla, T., Matas, J.G. (eds.) ECCV 2004. LNCS, vol. 3021, pp. 469–481. Springer, Heidelberg (2004)
3. Bai, X., Sapiro, G.: A geodesic framework for fast interactive image and video segmentation and matting. In: IEEE 11th International Conference on Computer Vision, ICCV 2007, pp. 1–8, October 2007
4. Berg, T.L., Berg, A.C., Edwards, J., Forsyth, D.: Who's in the picture. In: Saul, L.K., Weiss, Y., Bottou, L. (eds.) Advances in Neural Information Processing Systems 17, pp. 137–144. MIT Press, Cambridge (2005)
5. Bosch, A., Zisserman, A., Munoz, X.: Representing shape with a spatial pyramid kernel. In: Proceedings of the 6th ACM International Conference on Image and Video Retrieval, CIVR '07, pp. 401–408. ACM, New York (2007). http://doi.acm.org/10.1145/1282280.1282340

6. Criminisi, A., Sharp, T., Blake, A.: GeoS: geodesic image segmentation. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part I. LNCS, vol. 5302, pp. 99–112. Springer, Heidelberg (2008)
7. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2005, vol. 1, pp. 886–893, June 2005
8. Economou, G., Pothos, V., Ifantis, A.: Geodesic distance and MST based image segmentation. In: European Signal Processing Conference, pp. 941–944 (2004)
9. Felzenszwalb, P.F., McAllester, D.A., Ramanan, D.: A discriminatively trained, multiscale, deformable part model. In: CVPR (2008)
10. Freund, Y., Schapire, R.E.: A decision-theoretic generalization of on-line learning and an application to boosting. In: Vitányi, P.M.B. (ed.) EuroCOLT 1995. LNCS, vol. 904, pp. 23–37. Springer, Heidelberg (1995). http://dl.acm.org/citation.cfm?id=646943.712093
11. Hadid, A., Pietikainen, M., Ahonen, T.: A discriminative feature space for detecting and recognizing faces. In: Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2004, vol. 2, pp. II-797–II-804 (2004)
12. Kobayashi, T., Hidaka, A., Kurita, T.: Selection of histograms of oriented gradients features for pedestrian detection. In: Ishikawa, M., Doya, K., Miyamoto, H., Yamakawa, T. (eds.) ICONIP 2007, Part II. LNCS, vol. 4985, pp. 598–607. Springer, Heidelberg (2008)
13. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 2, CVPR '06, pp. 2169–2178. IEEE Computer Society, Washington, DC, USA (2006). http://dx.doi.org/10.1109/CVPR.2006.68
14. Lee, K., Ho, J., Kriegman, D.: Acquiring linear subspaces for face recognition under variable lighting. IEEE Trans. Pattern Anal. Mach. Intell. **27**(5), 684–698 (2005)
15. Liao, S.C., Zhu, X.X., Lei, Z., Zhang, L., Li, S.Z.: Learning multi-scale block local binary patterns for face recognition. In: Lee, S.-W., Li, S.Z. (eds.) ICB 2007. LNCS, vol. 4642, pp. 828–837. Springer, Heidelberg (2007)
16. Lienhart, R., Maydt, J.: An extended set of haar-like features for rapid object detection. In: 2002 International Conference on Image Processing, vol. 1, pp. I-900–I-903 (2002)
17. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vis. **60**(2), 91–110 (2004). http://dx.doi.org/10.1023/B:VISI.0000029664.99615.94
18. Lowe, D.: Object recognition from local scale-invariant features. In: The Proceedings of the Seventh IEEE International Conference on Computer Vision, vol. 2, pp. 1150–1157 (1999)
19. Ojala, T., Pietikäinen, M., Harwood, D.: A comparative study of texture measures with classification based on featured distributions. Pattern Recogn. **29**(1), 51–59 (1996). http://dx.doi.org/10.1016/0031-3203(95)00067-4
20. Papageorgiou, C., Poggio, T.: A trainable system for object detection. Int. J. Comput. Vision **38**(1), 15–33 (2000). http://dx.doi.org/10.1023/A:1008162616689
21. Paragios, N., Deriche, R.: Geodesic active contours and level sets for the detection and tracking of moving objects. IEEE Trans. Pattern Anal. Mach. Intell. **22**(3), 266–280 (2000)

22. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2001, vol. 1, pp. I-511–I-518 (2001)
23. Wu, B., Ai, H., Huang, C., Lao, S.: Fast rotation invariant multi-view face detection based on real adaboost. In: Proceedings of Sixth IEEE International Conference on Automatic Face and Gesture Recognition, pp. 79–84 (2004)
24. Wu, C., Duan, L., Miao, J., Fang, F., Wang, X.: Detection of front-view vehicle with occlusions using adaboost. In: International Conference on Information Engineering and Computer Science, ICIECS 2009, pp. 1–4 (2009)
25. Zhang, L., Chu, R.F., Xiang, S., Liao, S.C., Li, S.Z.: Face detection based on multi-block LBP representation. In: Lee, S.-W., Li, S.Z. (eds.) ICB 2007. LNCS, vol. 4642, pp. 11–18. Springer, Heidelberg (2007). http://dl.acm.org/citation.cfm?id=2391659.2391662
26. Zhu, Q., Yeh, M.C., Cheng, K.T., Avidan, S.: Fast human detection using a cascade of histograms of oriented gradients. In: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2, pp. 1491–1498 (2006)