Energy-Transfer Features for Pedestrian Detection

Radovan Fusek, Eduard Sojka, Karel Mozdřeň and Milan Šurkala

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

Abstract. In this paper, we propose an interesting and novel method for computing the image features that are useful for object detection. The method is interesting and novel in the terms of the feature vector dimensionality and object information capturing. In the proposed method, the areas of objects (that contain the important information useful for recognition) are described by the distribution of energy. The energy is transfered through the energy sources that are placed into the image and the distribution of energy is encoded into a vector of features. The vector is then used as an input for the SVM classifier. Using this approach, the objects of interest can be successfully described with a relatively small set of numbers if compared with the state-of-the-art descriptors that are based on the histograms of oriented gradients. We show the robustness of the features in the task of pedestrian detection.

1 Introduction

The objects of interest can be described using a lot of image information (e.g. shape, texture, colour). In the area of feature based detectors, the image features are carriers of this information. The goal is to design such features that are able to successfully describe the different objects of interest with a relatively small set of numbers. In this area and especially in the area of human detection, the features that are based on the Histograms of Oriented Gradients (HOG) [1] are dominant in the recent years. In HOG, the information about the distribution of gradient magnitudes and directions is used for the object description. The histograms of gradients are computed for each position of the sliding window that is dived into the blocks that consist of small connected cells. A feature vector is composed from these histograms and the vector is then used as an input for the trainable classifiers (e.g support vector machine). The detection methods based on these descriptors have been successfully presented in many papers (see Section 2).

Nevertheless, the classical HOG descriptors suffer from the high dimensionality of feature vector and sometimes it is useful to apply the methods for reducing the feature space (e.g. principal component analysis). The high dimensionality of feature vector negatively affects the speed of detection and training phases

and the large training set must be used for perfect object detection in the variable surroundings (streets, buildings, airports, etc.). In additional to that, the features that are based on the edge information (e.g. length, magnitude, orientation, localization) are also sensitive to the noise due to the fact that noise have a negative effect to the image quality (quality of edges). The noise must be suppressed, but the images can lose the important information about the object edges after the filtering. These shortcomings create the motivation for developing the novel approach for computing the image features that can be successful with a relatively small dimensionality of feature vector and with the filtering step that is directly included in the extraction of proposed features.

Basically, the proposed method is inspired by HOG but instead of the distribution of gradient magnitudes and directions the method captures the object information using the distribution of energy. In essence, the main idea of the proposed features is that the areas of objects (that consist of information about the shape) can be effectively described by the energy distribution. We will consider the transfer of energy as the transfer of heat in this paper. In our approach, the sliding window is also divided into the regions. The sources of temperature are defined inside each region. After the temperature transfer, the distributions of temperature inside the regions are encoded to the feature vector. Finally, the vector of features is used as an input for the SVM classifier. Since the temperature is transfered within the object areas, using this approach, we are able to describe the object areas with the positive filtration abilities that are obtained from the diffusion equation.

The next parts of the paper are organized as follows. The related works are described in Section 2. The process of extraction of the proposed features is described in Section 3. Finally, the results are shown in Section 4.

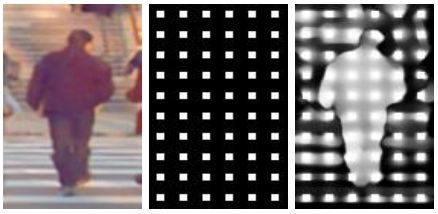
2 Related Works

In the area of feature based detectors, the methods that are based on the Histograms of Oriented Gradients [1] have been successfully presented in the recent years. Pedestrian detection method using infrared images and histograms of oriented gradients combined with the SVM classifier was presented in [2]. Near real-time human detection system using the cascade-of-rejectors with the histograms of oriented gradients was proposed in [3]. In [4], the authors applied the principal component analysis to the HOG feature vector to obtain the PCA-HOG vector. This vector contains the subset of HOG features and such vector is used as an input for the SVM classifier. Their method was used for pedestrian detection with the satisfactory results. The method for vehicle detection in low-altitude airborne videos using boosting HOG features was presented in [5]. In [6], the authors proposed Augmented Histograms of Oriented Gradients (AHOG) feature for human detection from a non-static camera. Their approach extended the classical HOG features by adding the human shape properties. The authors reported that the method achieved a good performance at many views of targets. The feature set that contains the combination of Histograms of Ori-

ented Gradients and Local Binary Pattern (HOG-LBP) for human detection was presented in [7]. Pyramid of Histogram of Orientation Gradients (PHOG) was proposed in [8]. This method uses the combination of the image pyramid representation and the histograms of orientation gradients. The very popular method for object detection was presented by Viola and Jones in [9]. Their method uses the integral image, rectangular features, and AdaBoost algorithm. The method was used for moving-human detection in [10].

3 Proposed Features

The main idea behind the proposed features is that the appearance of objects can be efficiently described by the function of energy distribution. Especially, the information about the object areas that are useful for recognition are precisely described by this distribution in the presented method. The usefulness of energy distribution can be described in the following way.



(a) (b) (c)

Fig. 1. The real-life image (a). The regular grid of sources (b). The visualization of distribution of temperature from these sources (c). The value of temperature is depicted by the level of brightness.

Consider the simple theoretical image containing one object of constant brightness with the extremely thin edges; theoretically, the edges can be infinitely thin. In the case that the object appearance should be described by analyzing the function of intensity gradients and directions. The sample values of such a function will be difficult to obtain; theoretically, this is not possible in the case of the infinitely thin edges. Conversely, the information about the area of this object can be described without any difficulties. Suppose that the temperature source is placed into the previously mentioned object with extremely thin edges, and suppose that the transfer of temperature can be solved by making use of physical laws inside the image; the thermal conductivity properties are determined by the gradient of brightness (hight gradients indicate the low conductivity and vice versa). After the temperature transfer that is carried out during a certain chosen time, the area of this object will contain a certain distribution of energy. The function values of this distribution are approximately constant inside the object area. The information about the object area (that also

contains the information about the shape of object) can then be simply obtained by sampling, and it can be used for the recognition.

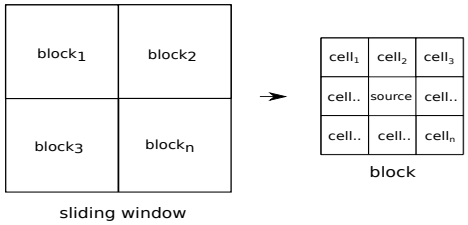


Fig. 2. The blocks and cells inside the sliding window.

It is clear that in the real images (Fig. 1(a)), the situation is more complicated, but the previous assumption can be extended also for these images. In the real images, the objects of interest consist of many areas and one temperature source will not be enough to cover all areas. Accordingly, suppose the locations of the sources in the form of a regular grid inside the image (Fig. 1(b)). The temperature transfer starts at the time $t = 0$ from all sources at once. The temperature of sources equals 1 for all $t > 0$. After the temperature transfer process inside the image (which is stopped at a suitable chosen time), the temperature distribution will reflect the areas of objects and also the shape of objects (Fig. 1(c)). After this process, the sample values from the distribution function can be used for recognition.

In the process of extracting the proposed features, the image inside the sliding window is divided into the regular blocks (Fig. 2). We use the gravity centers of these blocks as the places in which we put the temperature sources. For the purpose of obtaining the distribution of temperature, the blocks are divided into the small connected cells (Fig. 2).

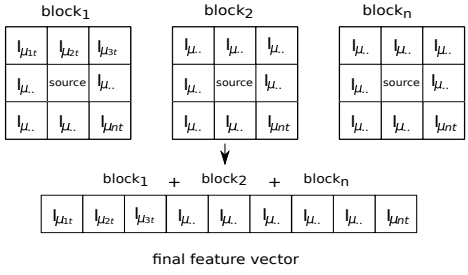


Fig. 3. The feature vector that is composed of the mean temperatures of cells.

Let $I(x, y, t)$ be a value of temperature at a time t and at a position (x, y). Inside each cell, the mean temperature $I\mu_{it}$ of the i-th cell at a time t can be calculated. The final feature vector is composed of these mean values (Fig. 3). We note that the temperature transfer is computed in the whole image inside the sliding window and temperature transfered from one source can influence every cell inside the sliding window; the blocks and cells are formed only for distribution measurement.

For the practical realization of the method, it is important to mention that the thermal field over one position of sliding window can be solved by making use of the following equation [11]

$$\frac{\partial I(x, y, t)}{\partial t} = \mathrm{div}(c \nabla I), \tag{1}$$

where I represents the temperature at a position (x, y) and at a time t, div is a divergence operator, ∇I is the temperature gradient and c stands for thermal conductivity. For the source points and arbitrary time $t \in [0, \infty)$, we set $I(x_s, y_s, t) = 1$, where (x_s, y_s) are the coordinates of source points (i.e. we hold the temperature constant during the whole process of transfer, which is in contrast with the usual diffusion approaches). In all remaining points, we take into account the initial condition $I(x, y, 0) = 0$. We solve the equation iteratively. The conductivity in Eq. 1 is determined by

$$c = g(\|E\|), \tag{2}$$

where E is an edge estimate. We define the edge estimate E as the gradient of original image $E = \nabla B$, where B is the brightness function. The function $g(\cdot)$ has the form of [11]

$$g(\|\nabla B\|) = \frac{1}{1 + \left(\frac{\|\nabla B\|}{K}\right)^2}, \tag{3}$$

where K is a constant representing the sensitivity to the edges [11]. Once the temperature field over the input image (inside the sliding window) is obtained (at a chosen time t), the mean cell temperature $I\mu_{it}$ can be obtained by making use of the formula

$$I\mu_{it} = \frac{\iint\limits_{M} I(x, y, t) dx dy}{|M|}, \tag{4}$$

where M stands for the cell area, and $|M|$ is its size.

In the next step, the SVM classifier is trained over the proposed descriptors. Let us consider a training data set (x_i, y_i) where x is the vector of proposed descriptors from training samples and y is the class label (+1 for pedestrian, -1 for non-pedestrian). The linear SVM determine hyperplane $w \cdot x + b$ where w is a weight vector, x is the vector of features and b is a constant. The goal is to find the optimal decision function that maximizes the distance between the nearest point x_i and the hyperplane. In the case when it is difficult separate examples in

a linear manner, the non-linear SVM can be used. The non-linear SVM maps the original space in a high-dimensional space using a kernel function that separate training samples. The optimal hyperplane for non-linear SVM is obtained by the function $f(x)$:

$$f(x) = \sum_{i=0}^{N} y_i \alpha_i k(x, x_i) + b, \tag{5}$$

where N represents the number of training patterns, y_i is a class indicator (+1 for pedestrian, -1 for non-pedestrian) for each training pattern x_i, α_i and b are learned weights and $k(.,.)$ is a kernel function. In our case, we use Gaussian radial basis function kernel:

$$k(x, y) = e^{\frac{|x-y|^2}{2\sigma^2}}. \tag{6}$$

4 Experiments

We collected 2500 positive samples and 10000 negative samples for the training phase. For the positive set, we combined the pedestrian images from the CBCL Pedestrian Database [12] with the images from the Daimler benchmark [13]. For the negative images, the examples were randomly sampled from the INRIA Person Dataset [1]. For the proposed method, each sample was resized to the size of 91×151 pixels. The visualization of the proposed features of positive samples is shown in the Fig. 5 (the parameters will be discussed later). The size of sliding window was set to the size of training samples. In the detection phase, we created the different resolutions of input image in which the sliding window was moving across these images.

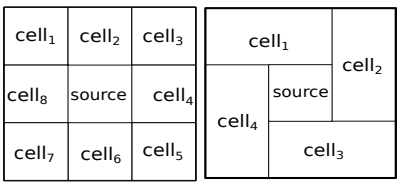


Fig. 4. The design of every block in the $Energy_{480}$ configuration (a). The design of every block in the $Energy_{240}$ configuration (b).

We experimented with the parameters of proposed features and we suggested the two optimal following configurations: $Energy_{480}$, $Energy_{240}$. The $Energy_{480}$ configuration was designed with the size of block = 15×15 pixels; the size of cells = 5×5 pixels (the number of cells inside each block = 8); the size of sources = 5×5 pixels; the time for the temperature transfer = 100 (the number of iterations).

Fig. 5. The visualization of proposed descriptors of pedestrian images. The value of temperature is depicted by the level of brightness. The features are designed with the following parameters: the size of blocks 15×15; the time $t = 150$ (the number of iterations for the transfer of temperature), the size of temperature sources $= 5 \times 5$.

This configuration consists of 480 descriptors for one position of sliding window. The design of each block of the $Energy_{480}$ configuration is shown in Fig. 4(a). The $Energy_{240}$ configuration was designed with the size of block = 15×15 pixels; the size of cells = 10×5 and 5×10 pixels (the number of cells inside each block = 4); the size of sources = 5×5 pixels; the time for the temperature transfer = 100 (the number of iterations). This configuration consists of 240 descriptors for one position of sliding window. The design of each block of the $Energy_{240}$ configuration is shown in Fig. 4(b).

For the comparison, we designed the two configurations of classical HOG features HOG_{336}, HOG_{3780}. We use the classical version of HOG descriptors without the extensions that were mentioned in Section 3 (e.g. PCA, AdaBoost, LBP) due to the fact that the proposed features are also presented without these extensions. We note that some of these extensions can be also used in the proposed features in the future works. The training sets for the HOG features and proposed features were identical (2500 positive and 10000 negative samples). For the HOG descriptors, each training sample was resized to the size of 64×128 pixels. The HOG_{336} configuration was designed with the similar number of descriptors like in the proposed configurations. The parameters were as follows; the size of block $= 32 \times 32$ pixels; the size of cell $= 16 \times 16$ pixels, the horizontal step size = 16 pixels; the number of bins = 4. This configuration consists of 336 HOG descriptors. The HOG_{3780} configuration was designed with the typical parameters of HOG descriptors; the size of block $= 16 \times 16$ pixels; the size of cell $= 8 \times 8$ pixels; the horizontal step size = 8 pixels; the number of bins = 9. This configuration consists of 3780 HOG descriptors. For the testing, we collected 55 images from the [1]. The detection results are shown in Table 1.

The worst detection results were acquired with the HOG_{336} configuration with the 336 HOG descriptors. The high numbers of false positive detections were visible in this configuration compared with the proposed method. This

	Precision	Sensitivity	F1 score
$Energy_{240}$	85.29%	77.33%	81.12%
$Energy_{480}$	**87.14%**	**84.72%**	**86.56%**
HOG_{336}	51.72%	85.71%	64.86%
HOG_{3780}	86.97%	81.97%	84.03%

Table 1. The detection performance of proposed features and the features that are based on HOG.

negative effect was caused by the small dimensionality of feature vector of the HOG configuration. Many significant object details cannot be precisely described with the size of blocks and cells of the HOG_{336} configuration. The 336 HOG descriptors were not able to successfully distinguish between the positive and negative samples and the detector based on this configuration detected human in the images in which the people were not visible; for example, the HOG_{336} configuration detected the traffic signs like pedestrians (Fig. 6).

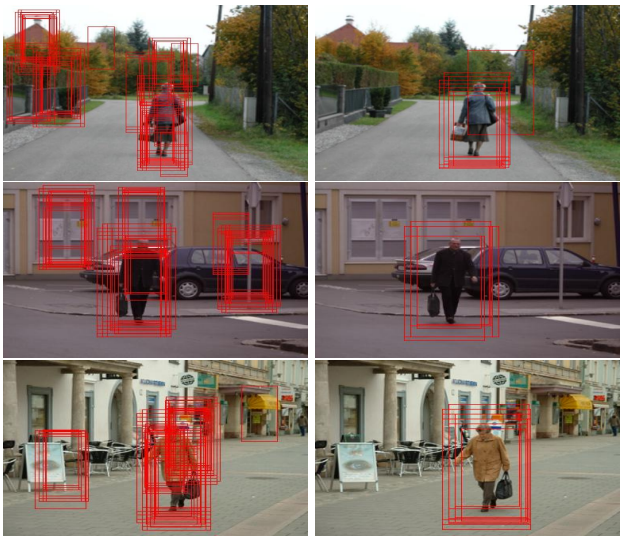


Fig. 6. The example of detection. The left images show the detection results of HOG_{336} configuration. The right images show the results of $Energy_{240}$ configuration. The detection results of approaches are shown without the post-processing (the detection results are not merged).

On the other hand, the $Energy_{240}$ configuration of the proposed method (with the relatively small set of descriptors = 240) was able to successfully describe the appearance of the objects of interest. The $Energy_{240}$ configuration of the proposed method achieved the better results (F1 score 81.12%) than the

HOG_{336} configuration (F1 score 64.86%). The $Energy_{240}$ configuration detected the objects of interest with the very promising detection rates. We tried to increase these rates by creating the second configuration with more descriptors. As we show in the results, the second proposed configuration ($Energy_{480}$) achieved the best detection rate (F1 score 86.56%). The detector that was based on this configuration successfully detected most of the pedestrians with the 480 descriptors. Compared with the HOG_{3780} configuration (F1 score 84.03%), the proposed features achieved the similar results, however the proposed method gives 7× less descriptors than the classical HOG_{3780} configuration.

Finally, the $Energy_{480}$ configuration shows that the pedestrians can be efficiently encoded with the reasonable dimensionality of feature vector without need for the methods for reducing the feature space. The detection results of the $Energy_{480}$ configuration are shown in Fig. 7.

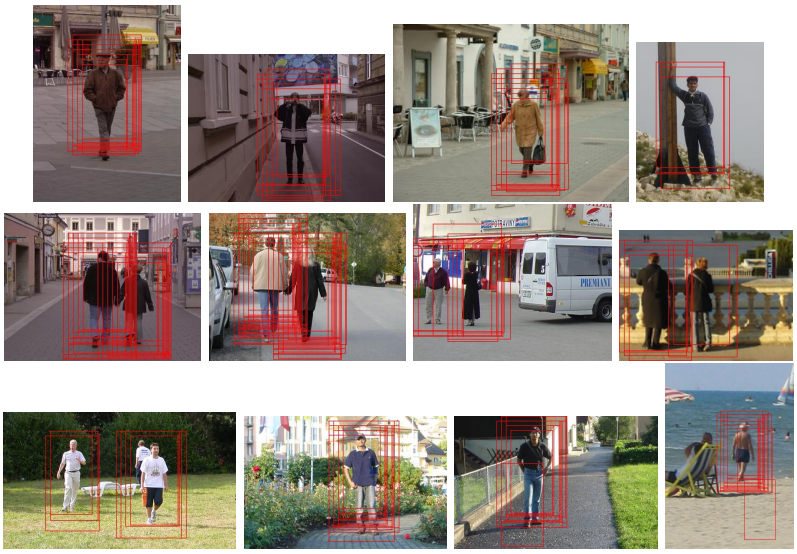


Fig. 7. The detection results of the $Energy_{480}$ configuration without the post-processing (the detection results are not merged).

5 Conclusion

In this paper, we proposed the efficient method for the computation of image features. The proposed features are based on the energy distribution. Using this distribution, the appearance of objects can be effectively described in the sense of dimensionality of the feature vector. The detection results that were achieved with this dimensionality are very promising for the future works in which we

will focus on the detection of other objects of interest (faces, cars) and we will also focus on the time complexity of computation of the proposed features.

Acknowledgments

This work was supported by the SGS in VSB Technical University of Ostrava, Czech Republic, under the grant No. SP2013/185.

References

1. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on. Volume 1. (2005) 886 –893 vol. 1
2. Suard, F., Rakotomamonjy, A., Bensrhair, A., Broggi, A.: Pedestrian detection using infrared images and histograms of oriented gradients. In: Intelligent Vehicles Symposium, 2006 IEEE. (2006) 206 –212
3. Zhu, Q., Yeh, M.C., Cheng, K.T., Avidan, S.: Fast human detection using a cascade of histograms of oriented gradients. In: Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on. Volume 2. (2006) 1491 – 1498
4. Kobayashi, T., Hidaka, A., Kurita, T.: Neural information processing. Springer-Verlag, Berlin, Heidelberg (2008) 598–607
5. Cao, X., Wu, C., Yan, P., Li, X.: Linear svm classification using boosting hog features for vehicle detection in low-altitude airborne videos. In: Image Processing (ICIP), 2011 18th IEEE International Conference on. (2011) 2421 –2424
6. Chuang, C.H., Huang, S.S., Fu, L.C., Hsiao, P.Y.: Monocular multi-human detection using augmented histograms of oriented gradients. In: Pattern Recognition, 2008. ICPR 2008. 19th International Conference on. (2008) 1–4
7. Wang, X., Han, T., Yan, S.: An hog-lbp human detector with partial occlusion handling. In: Computer Vision, 2009 IEEE 12th International Conference on. (2009) 32–39
8. 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, New York, NY, USA, ACM (2007) 401–408
9. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on. Volume 1. (2001) I–511 – I–518 vol.1
10. Viola, P., Jones, M., Snow, D.: Detecting pedestrians using patterns of motion and appearance. In: Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on. (2003) 734 –741 vol.2
11. Perona, P., Malik, J.: Scale-space and edge detection using anisotropic diffusion. IEEE Trans. Pattern Anal. Mach. Intell. **12** (1990) 629–639
12. Center for Biological and Computational Learning: MIT CBCL Pedestrian Database #1 (2013) http://cbcl.mit.edu/software-datasets/PedestrianData.html.
13. Enzweiler, M., Gavrila, D.: Monocular pedestrian detection: Survey and experiments. Pattern Analysis and Machine Intelligence, IEEE Transactions on **31** (2009) 2179–2195