Energy-Transfer Features and their Application in the Task of Face Detection

Radovan Fusek, Eduard Sojka, Karel Mozdřeň, 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.st, milan.surkala.st}@vsb.cz

Abstract

*In this paper, we describe a novel and interesting approach for extracting the image features. The features we propose are efficient and robust; the feature vectors of relatively small dimensions are sufficient for successful recognition. We call them the energy-transfer features. In contrast, the classical features (*e.g. *HOG, Haar features) that are combined with the trainable classifiers (*e.g. *a support vector machine, neural network) require large training sets due to their high dimensionality. The large training sets are difficult to acquire in many cases. In addition to that, the large training sets slow down the training phase. Moreover, the high dimension of feature vector also slows down the detection phase and the methods for the reduction of feature vector must be used. These shortcomings became the motivation for creating the features that are able to describe the object of interest with a relatively small number of numerical values without the use of methods for the reduction of feature vector. In this paper, we demonstrate the properties of our features in the task of face detection.*

1. Introduction

In the feature-based detectors that are combined with the trainable classifiers, the extraction of relevant features has a significant influence on the successfulness of detectors. The large number of features slows down the training and detection phases; on the other hand, the very small number of features need not be able to describe the properties of object of interest. The quality of training set and the selection of classifier is also equally important.

The proposed features are slightly inspired by the image features that are based on the histogram of oriented gradients (HOG) that was presented by Dalal and Triggs [3]. In their approach, the sliding window is divided into the small regions (cells). The histogram of gradient directions is computed within the regions. These regions are normalized across the larger regions (blocks) to provide better illumination invariance. The HOG descriptors are computed in every position of the sliding window. In their paper, the authors used the classifier based on the support vector machine (SVM). Many works show that the HOG descriptors are very useful in the various detection tasks (further details may be found in Section 2). Nevertheless, the classical HOG descriptors suffer from the large number of features, which causes that the training and detection phases can be time consuming. The sufficient amount of training data is also needed to find a separating hyperplane by the SVM classifier. Sometimes, it is desirable to use the methods for the dimensionality reduction of feature vector.

We experimented with these features and these shortcomings became the motivation for creating a novel method for the extraction of image features that give rise to the lower number of relevant values with the preservation of illumination and noise invariance properties without having to use the methods to reduce the feature space. We call them the energy-transfer features; if we speak about energy transfer in this paper, we have in mind transfer of heat. In the proposed method, we divide the whole input image into regions. Inside each region, we define the source of temperature. We calculate the mean temperature in these regions (instead of the histogram of oriented gradient that is used in HOG). For detection, the mean temperatures are then used for composing the feature vector of sliding window. The features are calculated globally in the whole input image only once (for each scale of input image). The feature vector is then used as an input for the SVM classier. In this paper, we demonstrate the robustness of the proposed features for solving the problem of face detection.

The paper is organized as follows. We introduce the readers to the area of feature-based detectors. Then, we present the new method and we show its properties. Finally, we compare our approach with the state-of-the-art methods.

2. Related Work

In the area of feature-based detectors, the HOG features have become the very popular object descriptors [3]. Many detection methods with the HOG features have been successfully presented. In [10], the authors presented the

pedestrian detection system applied to the infrared images that is based on the HOG features combined with the support vector machine (SVM) classifier. Zhu *et al*. [13] presented the AdaBoost-based feature selection combined with the integral image representation to compute the HOG features. The Authors report that the near real-time object localization is obtained but the accuracy of descriptors is reduced. In [2], the authors proposed boosting HOG features that are combined to the final feature vector to train the SVM classifier for vehicle classification. Kobayashi *et al*. [4] applied Principal Components Analysis to reduce the dimensionality of HOG features for pedestrian detection.

In the area of face detection, Kurita *et al*. [5] presented that the selection of the most relevant Gabor features could improve the accuracy of face detection. Papageorgiou and Poggio [7] proposed the object detection system for static images. Their system uses Haar wavelets for the description of faces, cars, and people combined with the SVM classier. Schneiderman and Kanade [9] presented the trainable detector for detecting faces and cars at any location, size, and pose. Viola and Jones [11] proposed the object detection framework based on image representation called integral image, rectangular features, and AdaBoost algorithm. With the use of integral image, the rectangular features are computed very quickly. The AdaBoost algorithm selects the most important features that are used to train classifiers and the cascade of classifiers is designed for reducing computation time. In [12], their detection framework was successfully extended for moving-human detection.

3. Energy-Transfer Features

The main idea of the proposed features is that the appearance of object of interest can be described by the distribution of temperature. The image can be considered as a rectangular plate with certain thermal conductivity properties that are determined by the gradient of brightness (big gradients indicate the low conductivity and vice versa). In the area of image, the distribution can be solved by making use of physical laws. We solve the distribution for the point sources of constant temperature that are appropriately located into the image. At $t = 0$, the temperature is zero in the whole area of image, except the temperature sources. We suppose that the heat transfer starts at $t = 0$ and, theoretically, it can be infinitely long. Nevertheless, we stop the transfer at a suitable time $t > 0$. During the whole time of transfer, the temperature at source points is held on a chosen initial value. After the transfer, the distribution of temperature is investigated. Since the contours of object correspond to the places with high gradients and since the values of gradients correspond to the value of thermal conductivity, we can conclude that the shapes of objects are encoded in the distribution of temperature that can be obtained by the process described above.

The usefulness and motivation to use the temperature distribution function can be described as follows. Suppose that the object of interest with the very thin edges is analyzed by the functions of gradient sizes and directions. The meaningful sample values of this function can be difficult to obtain; it is difficult to obtain (by the samples) the information about the thin edges (the samples need not hit the thin edges). On the other hand, the function of temperature distribution does not make problems during sampling. In this function, the areas with approximately constant temperature values are important and it is an easy matter to hit them by samples.

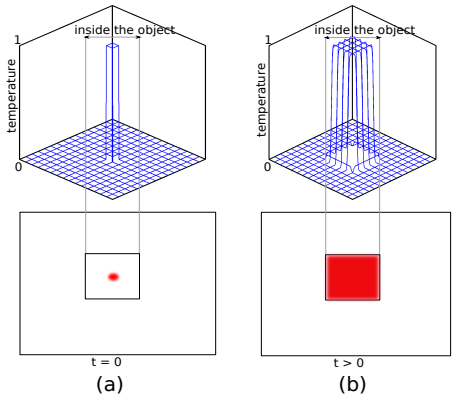
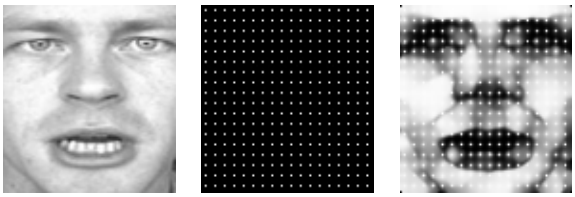


Figure 1. The image with one object and one source of temperature. The value of temperature is depicted by the intensity of red color.

For better understanding, let us firstly consider the following very simple theoretical image containing one rectangular object of constant brightness on the background (the gradient of brightness of this theoretical image is shown in the second row in Fig. 1). The problem of segmentation can be transformed into the problem of solving heat transfer as follows. At places where the size of gradient is zero, the thermal conductivity equals to infinity; where the size of gradient is greater than zero, the conductivity is zero. In this first example, we have only one source of temperature that is placed into a point lying inside the object (say into the center of gravity). For all $t >= 0$, the temperature at the source point is equal to 1. For $t = 0$, the temperature at all other places in the image is equal to 0 (Fig. 1(a)). After some time, $t > 0$ the distribution changes into the form as is depicted in Fig. 1(b). Clearly, the distribution of temperature reflects the shape of the object. It follows that the distribution of temperature can be used for recognizing. Generally, the distribution of temperature is a function with uncountably many values. For practical use, the function must be compressed into an acceptable amount of values. We solve the problem simply by sampling. We can imagine that the values of samples correspond to thermometers that monitor the values of temperature function at chosen loca-

tions of image. We can take the samples in a regular grid and we can use them as an input for recognition by SVM. We note that in this particular example, the time of transfer does not play a substantial role; the same distribution is achieved for every $t > 0$ due to the assumption about the infinite and zero conductivities.

The presented example also shows that one source point will not be sufficient for the real-life images (Fig. 3(a)). The reason can be easily understood. If we have more objects, if the objects are more complicated, and if we drop the assumption that the conductivity can only be either zero or infinity, more sources are apparently needed. Generally, the sources can be placed into a regular grid (Fig. 3(b)). The transfer of temperature starts from all sources at the same time. After the transfer that was carried out during a suitably chosen time, we obtain a temperature distribution. The distribution reflects the presence of objects and their parts, which is the main idea of method we propose. The visualization of temperature distribution is shown in Fig. 3(c).



(a) (b) (c)

Figure 3. 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.

For the purpose of recognition, as was said before, the function of temperature should be sampled. We can either simply take the values at a point grid or to carry out the sampling by integration. We regard the second approach as more robust. For this purpose, we divide the input image into cells and we investigate the mean temperature in each cell. Generally, the position of the sources of temperature can be chosen arbitrarily and independently on the cells. In the following text, however, we put the sources into each cell. As the position of sources, we use the gravity centers of cells.

Let us express the things more formally. Let $I(x, y, t)$ stand for the value of temperature at a position (x, y) and at a time t; the mean temperature in the i-th cell is denoted by $I\mu_i$. We can compute these mean temperatures for all cells in the whole input image. In the process of recognition, we use a sliding window. The vector of features for each position of the sliding window is assembled from the mean temperatures in the cells that fall into the window in its actual position (Fig. 4). The i-th item in the feature vector is the mean temperature $I\mu_i$ in the i-th cell in the window. The vector of features is then used in the SVM classifier.

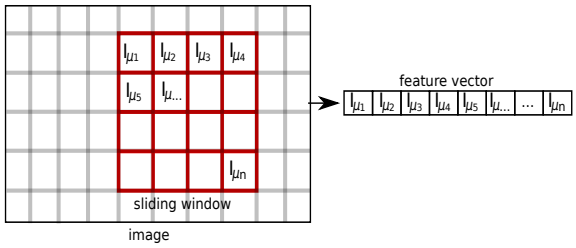


Figure 4. The vector of features for a momentary position of sliding window.

For practical realization of the method, it is important to mention that the thermal field over the input image can be solved by making use of the following equation [8]

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

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 [8]

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

where K is a constant representing the sensitivity to the edges [8]. Once the temperature field over the input image is obtained (at a chosen time t), the mean cell temperature $I\mu_i$ can be obtained by making use of the formula

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

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

After computing the vector of features, the SVM classifier is trained. The function of SVM classifier has the following form

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

Figure 2. The visualization of energy-transfer features (ETF). The firs row represents the original face images. The second row represents the visualization of ETF from these images. In the visualization of ETF, the features are designed with the following parameters: the size of cells 5×5, time $t = 200$ (the number of iterations for the transfer of temperature), and the constant $K = 10$. The value of temperature is depicted by the level of brightness.

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

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

where σ defines the kernel width. This kernel is very often used in SVM.

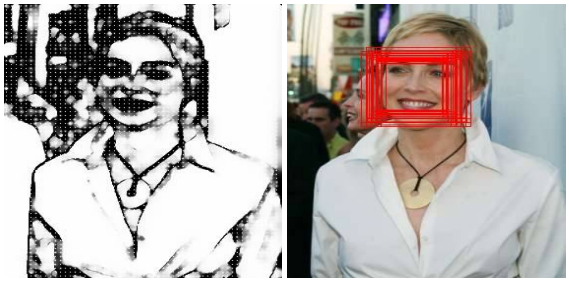


Figure 5. The example of visualization of proposed features. The left image shows the temperature function inside the whole input image (the value of temperature is depicted by the level of brightness). The right image shows the detection result of ETF_{324} configuration without the postprocessing (the detection results are not merged).

4. Face Detection

For the training phase, our positive set consists of 1700 faces. We used the face images from the BIOID database (https://www.bioid.com/downloads/software/bioid-face-database.html) combined with the Caltech face dataset (http://www.vision.caltech.edu/html-files/archive.html). We manually cropped these images on the area of faces only. The negative set consists of 3000 images that was obtained from the MIT-CBCL database (http://cbcl.mit.edu/software-datasets/FaceData2.html).

We resized all training images to the size of 90×90 pixels. The visualization of energy-transfer features (ETF) is shown in Fig. 2 (the effect of parameters will be discussed later). In the detection phase, we create the eight different resolutions of input image; the proposed features are computed for each resolution. The size of sliding window is set to the size of training images (90×90 pixels). We experimented with the parameters of our method and we suggested the following configurations: ETF_{100}, ETF_{324}. The ETF_{100} configuration is designed with the size of cells 9×9, time chosen for the transfer of temperature $t = 350$, $K = 10$. This configuration consists of 100 features. The ETF_{324} configuration is designed with the size of cells 5×5, time chosen for the transfer of temperature $t = 200$, $K = 10$. This configuration consists of 324 features. In the ETF_{100} configuration, the higher time parameter is necessary due to the larger cell size. The larger cells require more iterations for the transfer of temperature to affect the adequate area of image. The example of visualization of temperature function inside the whole input image with the positive detections of ETF_{324} configuration is shown in Fig. 5.

For comparison, we used the detectors that are based on the HOG features, LBP (Local Binary Patterns) features ([6]) and Haar features (Viola-Jones detection framework). In essence, the HOG and LBP features are similar to the proposed features in the sense that they also compute values over regular regions (*e.g.* square blocks) and, therefore, these features are suitable for comparing (in contrast with SURF, SIFT features that are based on arbitrarily located feature points). The Viola-Jones detector that is based on the Haar features was used because it is considered as a state-of-the-art detector in the area of face detection.

Figure 6. The detection results of our approach (ETF_{324}) without the postprocessing (the detection results are not merged).

We experimented with the parameters of HOG descriptors and we suggested the following configurations: HOG_{324}, HOG_{1296}. The HOG_{324} configuration was designed with the same number of feature values like in the ETF_{324} configuration. The parameters were as follows: The size of block = 32×32, size of cell = 16×16, horizontal step size = 32, number of bins = 9. This configuration gives 324 HOG features. The HOG_{1296} configuration was designed with the size of block = 16×16, size of cell = 8×8, horizontal step size = 16, number of bins = 9. This configuration gives 1296 HOG features. For the HOG features combined with the SVM classifier, we resized the training images to the size of 96×96 and we used the same training images that we used for the proposed features with SVM (1700 faces, 3000 non-faces).

For the detectors based on the Viola-Jones detection framework with Haar features and with the features that are based on LBP, we created the cascade classifiers. For these classifiers, we resized the training images (1700 faces, 3000 non-faces) to the size of 19×19. The resulting cascade classifiers had 11 stages for the LBP features and also for Haar features (we note that the number of stages need not be sufficient and it incurred as a result of a relatively small amount of training data, however, we wanted to test all approaches with the same training data). To calculate the performance of approaches, we collected the set of 80 images that contains 117 faces from the Faces in the Wild dataset [1]. Before the process of performance calculation, the positive detections were merged to one if at least 6 positive detections hit approximately one place in the image. In Table 1, the detection results are shown. The ETF_{324} configuration achieved the best result with the 324 features (F1 score = 90.08%). With this size of feature vector, the proposed approach was able to describe the main features of faces and this configuration detected faces with the lower number of false positive detections than the booth configurations of HOG features, Haar based detector and LBP based detector. In the ETF_{100} configuration with the 100 features, the proposed approach correctly detected most of the faces, however, with the high number of false positive detections. This size of feature vector is not able to describe all non-faces samples correctly (F1 score = 49.25%).

	Precision	Sensitivity	F1 score
ETF_{100}	32.9%	97.46%	49.25%
ETF_{324}	**87.90%**	**92.32%**	**90.08%**
HOG_{324}	67.25%	98.29%	79.89%
HOG_{1296}	41.97%	97.41%	58.25%
Haar	74.45%	87.18%	80.31%
LBP	62.07%	76.92%	68.70%

Table 1. The detection performance.

The configuration of HOG_{324} successfully detected the majority of faces, nevertheless, with the false positive de-

tections in some cases. For example, the 324 HOG features were not able to distinguish between the balloon and face (Fig. 7). In these cases, the HOG_{324} configuration detected most of the spherical objects like faces. The HOG based detector (HOG_{324}) is not able to describe all details of the faces with such a small amount of descriptors (F1 score = 79.89%). On the other hand, the HOG_{1296} configuration suffers from the high number of false positive detections (F1 score = 58.25%). We experimented with many configurations of HOG features, but increasing the number of features (without increasing training data and without the reduction of feature space) did not improve the detection accuracy.

Haar based detector had less false positive detections (precision = 74.45%) than the booth configurations of HOG features and LBP based detector. However, the Haar based detector and LBP based detector missed some of the faces (sensitivity = 87.18% and 76.92%, respectively). These detectors would also need to increase the amount of training data to achieve better results.

Finally, the ETF_{324} configuration shows that the faces can be described with a reasonable number of features with very good detection results and also with a relatively small set of training data without need for the methods for reducing the feature space.

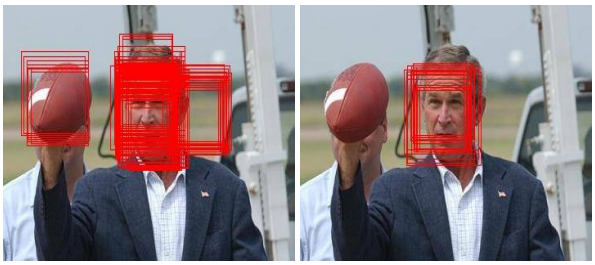


Figure 7. An example of detection. The left image shows the detection result of HOG_{324} configuration. The right image shows the result of ETF_{324} configuration. The detection results of approaches are shown without the postprocessing (the detection results are not merged).

5. Conclusion

In this paper, we proposed the novel approach for the computation of image features. We call them the energy-transfer features (ETF). The presented features are based on the distribution of energy (temperature). The vector of the features is used as an input for the SVM classifier. We demonstrated that ETF were able to describe the faces with a relatively small number of relevant features and with the high accuracy. In our future work, we will focus on the detection of other objects of interest (humans, cars) using 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] T. L. Berg, A. C. Berg, J. Edwards, and D. A. Forsyth. Whos in the picture. In *NIPS*, 2004.

[2] X. Cao, C. Wu, P. Yan, and X. Li. 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*, pages 2421 –2424, sept. 2011.

[3] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 886 –893 vol. 1, june 2005.

[4] T. Kobayashi, A. Hidaka, and T. Kurita. Neural information processing. chapter Selection of Histograms of Oriented Gradients Features for Pedestrian Detection, pages 598–607. Springer-Verlag, Berlin, Heidelberg, 2008.

[5] T. Kurita, K. Hotta, and T. Mishima. Feature ordering by cross validation for face detection. In *MVA*, pages 211–214, 2000.

[6] S. Liao, X. Zhu, Z. Lei, L. Zhang, and S. Z. Li. Learning multi-scale block local binary patterns for face recognition. In *ICB*, pages 828–837, 2007.

[7] C. Papageorgiou and T. Poggio. A trainable system for object detection. *Int. J. Comput. Vision*, 38(1):15–33, June 2000.

[8] P. Perona and J. Malik. Scale-space and edge detection using anisotropic diffusion. *IEEE Trans. Pattern Anal. Mach. Intell.*, 12:629–639, July 1990.

[9] H. Schneiderman and T. Kanade. Object detection using the statistics of parts. *Int. J. Comput. Vision*, 56(3):151–177, Feb. 2004.

[10] F. Suard, A. Rakotomamonjy, and A. Bensrhair. Pedestrian detection using infrared images and histograms of oriented gradients. In *in IEEE Conference on Intelligent Vehicles*, pages 206–212, 2006.

[11] P. Viola and M. Jones. 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, pages I–511 – I–518 vol.1, 2001.

[12] P. Viola, M. Jones, and D. Snow. Detecting pedestrians using patterns of motion and appearance. In *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, pages 734 –741 vol.2, oct. 2003.

[13] Q. Zhu, M.-C. Yeh, K.-T. Cheng, and S. Avidan. 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, pages 1491 – 1498, 2006.