Hierarchical Energy-Transfer Features

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*

Keywords: Object Detection, Recognition, SVM, Image Descriptors, Feature Selection.

Abstract: In the paper, we propose the novel and efficient object descriptors that are designed to describe the appearance of the objects. The descriptors are called as Hierarchical Energy-Transfer Features (HETF). The main idea behind HETF is that the shape of the objects can be described by the function of energy distribution. In the image, the transfer of energy is solved by making use of physical laws. The function of the energy distribution is obtained by sampling, after the energy transfer process; the image is divided into the cells of variable sizes and the values of the function is investigated inside each cell. The proposed descriptors achieved very good detection results compared with the state-of-the-art methods (e.g. Haar, HOG, LBP features). We show the robustness of the descriptors for solving the face detection problem.

1 INTRODUCTION

The area of computer vision includes many tasks that have been well researched in recent years. In this paper, we will focus on the problem of object description and detection. It is clear that the images contain many objects of interest. The goal of the object detection systems is to find the location of these objects in the images (e.g. cars, faces, pedestrians). For example, the vehicle detection systems are crucial for traffic analysis or intelligent scheduling, the people detection systems can be useful for automotive safety, and the face detection systems are the key part of face recognition systems. Typically, the detection algorithms are composed from two main parts in the area of feature-based detectors. The extraction of image features is the first part. The second part is created by the trainable classifiers that handle the final classification (object/non-object). In this paper, our contribution will be focused on the first part; on the extraction of image features.

The proposed features are based on the idea that the appearance of objects can be described by the function of the distribution of energy; if we speak about the energy transfer in this paper, we have the transfer of heat in mind. We suppose that the image is a rectangular plate with the thermal conductivity properties; big gradients represent the low conductivity and vice versa. In the process of obtaining the proposed features, the temperature sources are placed into the image. The transfer of temperature starts from the sources at the same time and the transfer is carried out during a chosen time. The distribution of temperature is investigated after the temperature transfer. The vector of features that is composed of this distribution is then used as an input for the SVM classifier. In essence, the proposed features are slightly inspired by HOG but instead of the histograms of oriented gradients we investigate the distributions of temperature. In this paper, we show the properties of the proposed features for solving the problem of face detection and we show that using the function of temperature distribution, the faces can be described with a reasonable number of features with good detection results. The preliminary version of the presented method (without the hierarchical improvement) was used for face detection in (Fusek et al., 2013).

The paper is organized as follows. In Section 2, we present an overview of the state-of-the-art feature-based detectors. In Section 3, we describe the extraction of the proposed features in detail, and we show the experiments and results in Section 4.

2 RELATED WORKS

In this section, we will especially focus on the image features that are extracted over the sliding window that are divided into the dense areas (due to the

fact that the proposed descriptors are also calculated within the dense regular blocks).

In recent years, the object detectors that are based on the edge analysis providing the valuable information about the objects of interest have been used in many detection tasks. In this area, the Histograms of Oriented Gradients (HOG) (Dalal and Triggs, 2005) are considered as the state-of-the-art method. In essence, the HOG descriptors can be regarded as a dense version of the SIFT features. In HOG, a sliding window is used for recognition. The window is divided into small connected cells in the process of obtaining HOG descriptors. The histograms of gradients are calculated for each cell. It is desirable to normalize the histograms across a large block of image. As a result, a vector of values is computed for each position of window. This vector is then used for recognition, e.g. by the Support Vector Machine (SVM) classifier (Boser et al., 1992). Many methods and applications that are based on HOG have been successfully presented in recent years. The PHOG descriptors that use the pyramid image representation were presented in (Bosch et al., 2007). The classical HOG-based detector was used for detecting upper bodies for automated upper body pose estimation in (Ferrari et al., 2008). Felzenszwalb et al. (Felzenszwalb et al., 2010) proposed the part-based detector that is based on HOG. In this method, the objects are represented using mixtures of deformable HOG part models and these models are trained using a discriminative method.

Haar-based descriptors represent the next possibility for object description. The main idea behind the Haar-like features is that the features can encode the differences in mean intensities between two rectangular areas. For instance, in the problem of face detection, the regions around the eyes are lighter than the areas of the eyes; the regions below or on top the eyes have different intensities that the eyes itself. These specific characteristics can be simple encoded by one two-rectangular feature, and the value of this feature can be calculated as the difference between the sum of the pixels inside the two rectangles. The Haar-like features which are similar to Haar basis function have first been proposed by Papageorgiou and Poggio (Papageorgiou and Poggio, 2000). In their paper, the Haar-like features are combined with the SVM classifier. The authors used the three types of Haar features (vertical, horizontal, diagonal) that were able to encode the changes in the intensities at various locations, scales, and orientations. The authors also reported promising performance in the tasks of face, car, and people detection in their paper. Since then, the Haar-like features and their modifications have been used in many works. The paper of Viola and Jones (Viola and Jones, 2001) contributed to the popularity of Haar-like features. They proposed the object detection framework based on the image representation called the integral image combined with the rectangular features, and AdaBoost algorithm (Freund and Schapire, 1995). The extension of the Haar feature set has been presented by Lienhart et al.(Lienhart and Maydt, 2002). In this paper, the authors presented 45° rotated features that are able to reduce the false alarm and achieve more accurate face detection.

Local Binary Patterns (LBP) that were introduced by Ojala et al.(Ojala et al., 1996) for texture analysis can also be used for object description. The main idea behind LBP is that the local image structures (so called micro patterns such as lines, edges, spots, and flat areas) around each pixel can be efficiently encoded by comparing every pixel with its neighboring pixels; in the basic form, every pixel is compared with its neighbors in the 3×3 region. The result of the comparison is the 8-bit binary number for each pixel; in the 8-bit binary number, the value 1 means that the value of center pixel is greater than the neighbor and vice versa. The histogram of these binary numbers is then used to encode the appearance of the region. The important properties of LBP are resistance to lighting changes and low computational complexity. Duo to their properties, LBP have been used in many recognition tasks, especially for facial image analysis. In (Hadid et al., 2004), LBP were used for solving the face detection problem in low-resolution images. Multi-block Local Binary Patterns (MB-LBP) for face detection were proposed in (Zhang et al., 2007). The paper of Tan and Triggs (Tan and Triggs, 2010) proposed the face recognition method using robust preprocessing based on the Difference of Gaussian image filter combined with LBP in which the binary LBP code is replaced by the ternary LTP code.

In the next section, we will show the process of extraction of the proposed features.

3 HIERARCHICAL ENERGY-TRANSFER FEATURES

As was mentioned before, the main idea behind the proposed features is based on the fact that the appearance of objects can be efficiently described by the function of temperature distribution in the following way.

Suppose the theoretical image that contains one object (with one area), and suppose that this object

has very thin edges; theoretically, the edges can be infinitely thin (the gradient of brightness of this theoretical image is shown in the second row in Fig. 1). Suppose that this 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 (it may happen that the samples will not hit the thin edges). On the other hand, the function of temperature distribution does not make problems during sampling. Suppose that the source of constant temperature is placed inside the considered object; say that into the gravity center of object (Fig. 1(a)). Inside the image, the temperature transfer (from this source) can be solved by making use of physical laws. It means that the gradients of the object can be considered as the thermal insulator; high gradients indicate the low conductivity and vice versa. After the temperature transfer that is carried out during a chosen time, the area of this object will contain a certain distribution of temperature (Fig. 1(b)), and the function of temperature distribution inside this object can be investigated. In this function, the areas with approximately constant temperature values are important and it is an easy matter to hit them by samples.

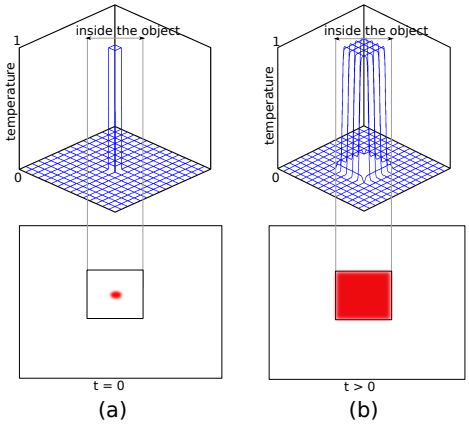

Figure 1: The image with one object and one source of temperature. The value of temperature is depicted by the intensity of red color.

Finally, the temperature distribution reflects the presence of objects and their parts and the appearance of object of interest can be described by the distribution (with a relatively small amount of descriptors), which is the main idea of the method we propose.

It is clear that the real images consist of objects with different areas (Fig. 2(a)) and one temperature source will not be enough to cover all areas. Therefore, we chose the location of temperature sources in the form of regular grid (Fig. 2(b)). At the time $t = 0$,

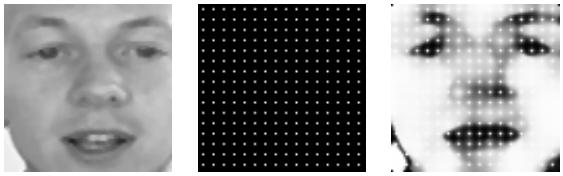

Figure 2: The real-life image (a). The regular grid of sources (b). The visualization of distribution of temperature from these sources (c). The value of temperature is depicted by the level of brightness.

the temperature of constant value 1.0 is attached to the source locations; other places inside the image have zero temperature at the time $t = 0$. The temperature transfer that starts from all sources at the same time is carried out during a chosen time t. Once the temperature transfer inside the image is obtained, the function of temperature distribution inside the image is investigated (the example of temperature distribution is shown in Fig. 2(c)). For this purpose, the image is iteratively divided into the finer spatial cells; i.e. we recursively divide the image into the cells of varying size (Fig. 3). In general, the image at hierarchical level l has 4^l cells. Inside each cell, the distribution is investigated.

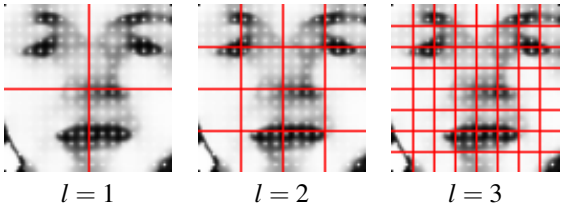

Figure 3: The different hierarchical levels of the cells.

Let $I(x,y,t)$ be the function of temperature (at location (x,y) and the time t) that is determined. We can compute the mean temperature in every cell; $I\mu_{it}$ stands for the mean temperature of the i-th cell at the time t. We use the mean cell temperatures as the values in the feature vector. For the additional information and for the precise description of the temperature distribution, we use the histogram of the temperature distribution that is also determined inside the cells. Each cell is represented by the $\sqrt{|M|} + I\mu_{it}$ dimension vector, where $|M|$ is the size of the cell, and $\sqrt{|M|}$ represents the number of histogram bins; the final vector of features is composed of the mean temperature and from the histograms of each cell at each hierarchical level. For example, for one cell of size 400 pixels (20×20), we compute the 20 histogram bins and the mean temperature of the cell $I\mu_{it}$, and the feature vector of this cell is the vector with dimension

$d = 21$. The process of composing the final feature vector for the levels $l = 0$ and $l = 1$ is shown in Fig. 4.

In the detection phase, we use the sliding window technique. The size of detection window is set to the size of training samples. Once the temperature field inside the whole input image is computed, the detection window scans this field and the feature vector is composed inside the window. The vector is then used as an input for the SVM classifier. It is important to mention that the classical HOG descriptors are not rotationally invariant. Since the proposed descriptors are similar to the HOG descriptors (in the sense that the features are computed in a grid in both approaches), this limitation also occurs in the proposed descriptors. Similarly as in HOG, the scale invariance is achieved by rescaling the images.

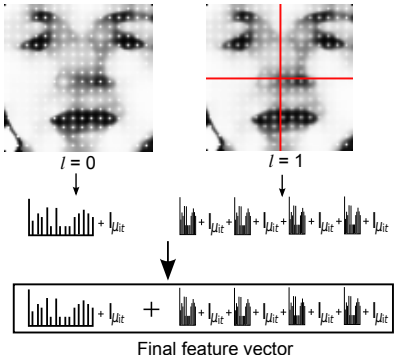

Figure 4: The process of composing the final feature vector. In the case that the image has size 80×80 pixels, the vector dimension is $d = 81$ at the level $l = 0$ (80 bin histogram + mean temperature $I\mu_{it}$). The level $l = 1$ contains 4 cells with size 40×40; the 40 bin histogram with the mean temperature $I\mu_{it}$ is calculated in each cell (41 values for each cell) and the vector dimension is $d = 164$. In this case, the final feature vector is composed as the sum of vector at the level $l = 0$ and $l = 1$.

For the practical realization of the method, it is important to mention that the thermal field inside the image can be solved by making use of the following equation (Perona and Malik, 1990)

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

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 the 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 (Perona and Malik, 1990)

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

where K is a constant representing the sensitivity to the edges (Perona and Malik, 1990). Once the temperature field over the input image is obtained (at a chosen time t), the mean cell temperature $I\mu_{it}$ can be obtained by making use of the formula

$$I\mu_{it} = \frac{\iint_M I(x,y,t)dxdy}{|M|}, \tag{4}$$

where M stands for the cell area, and $|M|$ is its size. The Support Vector Machine classifier with the radial basis function kernel is trained over the proposed descriptors in the next step to create the final classifier.

4 EXPERIMENTS

For the training phase, the positive set consists of 2300 faces and 4300 non-faces. We used the face images from the BIOID database (https://www.bioid.com/downloads/software/bioid-face-database.html) combined with the Extended Yale Face Database B (Lee et al., 2005). We manually cropped these images on the area of faces only. The negative set consists of 3000 images that were obtained from the MIT-CBCL database (http://cbcl.mit.edu/software-datasets/FaceData2.html) combined with the 1300 hard negative examples. The training images (for the proposed method) were resized to the size of 80×80 pixels. The visualization of temperature distribution of faces is shown in Fig. 5.

As we said before, we use the sliding window technique in the detection phase. The size of detection window is set to 80×80 pixels (the size of training samples). We use the fixed size of window that scan the image in 12 different resolutions of input image. The thermal field is computed for each resolution. The example of visualization of temperature function inside the whole input image (of one resolution of input image) with the positive detections is shown in Fig. 6.

Figure 5: The visualization of distribution of temperature. The value of temperature is depicted by the level of brightness. The firs row represents the original face images. The second row represents the visualization of temperature distribution.

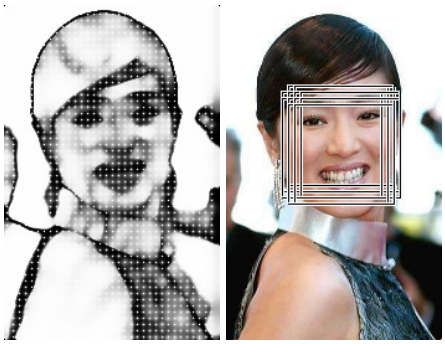

Figure 6: The example of visualization of the temperature field. The left image shows the temperature function inside the whole input image (the value of temperature is depicted by the level of brightness). The right image shows the detection results without the postprocessing (the detection results are not merged).

We experimented with the proposed method and we suggest the following configuration. The configuration of hierarchical energy transfer features is denoted as $HETF$. The size of temperature sources is 1 pixel and the distance between the sources is 5 pixels. The time for temperature transfer is 250; this number represents the number of iterations because we solve equation Eq. 1 iteratively. The mean temperature is computed inside each cell at levels $0,1,2,3,4$; it means that the mean temperature is computed inside the 341 cells (341 values in the feature vector). Additionally to that, the histogram of the temperature distribution is computed at levels $0,1,2,3$ (1200 values in the feature vector). Due to the fact that the cells at $l = 4$ are relatively small (5×5 pixels), we describe these cells with the mean temperature only. Finally, the final feature vector consists of 1541 descriptors for one position of sliding window; we experimented with the different settings of levels and this configuration achieved the best detection results. This configuration is also used for the visualization of proposed features in Fig. 2, 3, 4, 5, and 6.

For comparison, we used the detectors that are based on the HOG features, LBP (Local Binary Patterns) features (Liao et al., 2007) and Haar features (Viola-Jones detection framework).

For the HOG features we used the identical training sets like for the proposed features (2300 positive and 4300 negative samples) and we also used the identical size of the samples (80×80pixels). We used the typical parameters of HOG descriptors; the size of block = 16×16 pixels; the size of cell = 8×8 pixels; the horizontal step size = 8 pixels; the number of bins = 9. This configuration consists of 2916 HOG descriptors for one position of sliding window; this configuration is denoted as HOG. The Support Vector Machine classifier with Radial basis function kernel is trained over the HOG descriptors create the final classifier (similarly in the proposed descriptors).

For the detector based on the Viola-Jones detection framework and for the LBP-based detector, we also used the identical training sets like for the proposed features (2300 positive and 4300 negative samples) and we resized the training images to the size of 19×19. The detector based on the Viola-Jones detection framework is denoted as $Haar$, the detector based on the LBP features is denoted as LBP. It is important to mention that for these features, we created the cascade classifiers.

To calculate the performance of approaches, we collected the set of 200 images that contains 250 faces from the Faces in the Wild dataset (Berg et al., 2005). Before the process of performance calculation, the positive detections were merged to one if at least 3 positive detections hit approximately one place in the image. In Table 1, the detection results are shown.

The HOG-based detector achieved the good number of true positive detections (sensitivity 97.58%), nevertheless, the number of false positives is larger (precision 68.75%). For example, the spherical-like objects (e.g. balls, tennis rackets) are the problem for the HOG-based detector. In general, this detector detected a lot of objects in the images and many of these objects are not faces. In contrast, the proposed

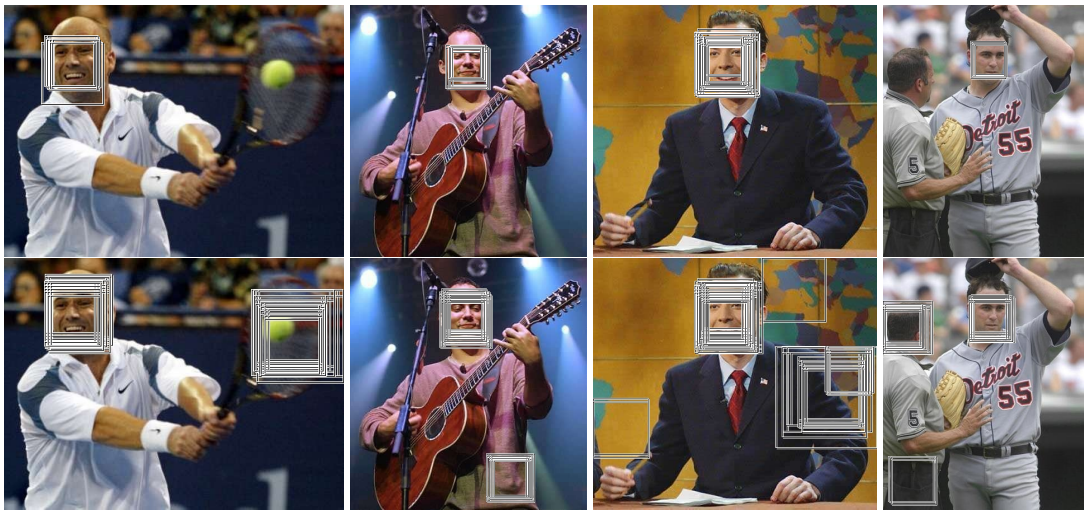

Figure 7: The differences between detection results. The first row: the results of $HETF$. The second row: the results of HOG. The results are without the postprocessing (the detection results are not merged).

	Precision	Sensitivity	F1 score
$HETF$	98.15%	86.18%	91.77%
$HETF_{PCA}$	**93.15%**	**93.90%**	**93.52%**
HOG	68.75%	97.58%	80.67%
$Haar$	85.77%	87.55%	86.65%
LBP	62.96%	69.39%	66.02%

Table 1: The detection performance.

method achieved the very promising numbers of true positive detections (sensitivity 86.18%) and the proposed descriptors also achieved the small number of false positive detections (precision 98.15%) using 2× less descriptors than in the HOG-based detector. We experimented with the different settings of HOG descriptors, however, without the increasing the datasets for training, the HOG-based detector was not able to achieve better results. The examples of the cases in which the HOG-based detector failed are shown in Fig. 7.

The cascade classifier of Haar features achieved better detection results than HOG descriptors (F1 score 86.65%). Nevertheless, this detector needed to increase the number of training samples to achieved better results. This problem also occurs in the LBP-based detector (F1 score 66.02%).

With respect to the fact that the proposed features had a low number of false positive detections (50 false positive detection windows of 7 million detection windows) and the number of descriptors (1541) achieved the good results (F1 score 91.77%), we reduced the number of descriptors using PCA (Principal Component Analysis). To determine the number of principal components, we used the 200 principal components (corresponding to the largest eigenvalues); the final feature vector is the vector with dimensionality $d = 200$. The detector that are based on this subset of the proposed features is denoted as $HETF_{PCA}$.

The detector based on the subset of the proposed features, achieved F1 score 93.52%; we note that the number of false positives is larger sing this subset nevertheless the detector achieved the best detection results (the examples in which the $HETF_{PCA}$ detector failed are shown in 8). The detector also shows that the appearance of the faces can be successfully described using the distribution of temperature with a relatively small set of the training samples and with a relatively small set of descriptors compared with the state-of-the-art descriptors. The detection results of the detector based on the $HETF_{PCA}$ configuration are shown in Fig. 9.

If we discuss about the computational time of the algorithm, the measurement of the time can be divided into two parts. The first part contains the time that is necessary for the temperature transfer inside the image. Since we solve the diffusion equation iteratively, the number of iterations has a major impact on the computational time of this part. We have developed the GPU (CUDA) and CPU (SSE/AVX) versions for solving the temperature transfer process. The computational time of GPU version is 40 milliseconds, the time of CPU version is 150 milliseconds for 150 iterations and for the size of input image 640×480 pixels.

The second part contains the time that is required for composing the feature vector. The mean temperature inside the rectangular cells is computed in a con-

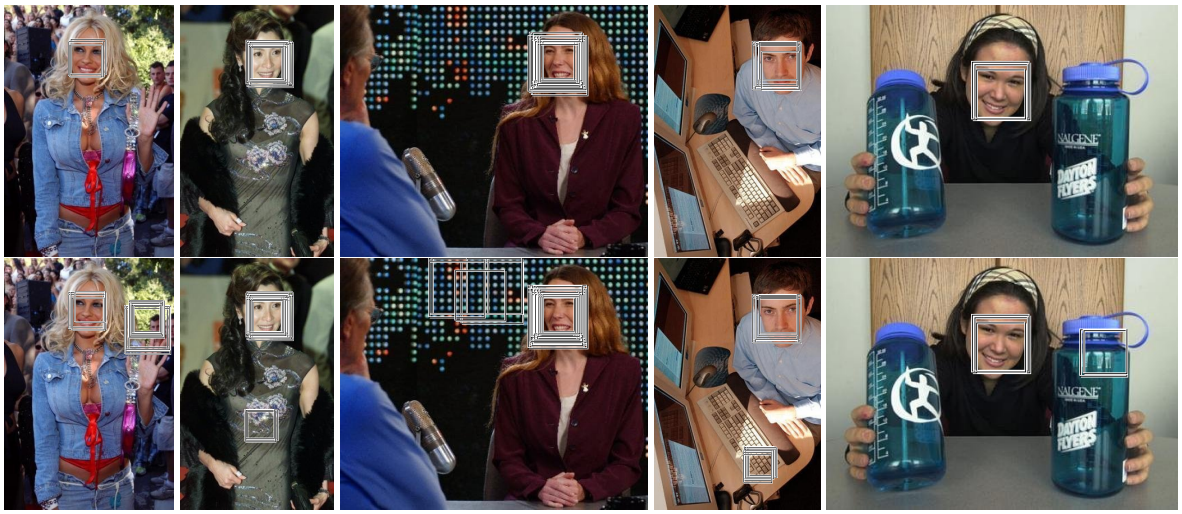

Figure 8: The detection results of $HETF_{PCA}$ in which the detector failed. The first row: the results of $HETF$ (without PCA). The second row: the results of $HETF_{PCA}$ (with PCA). The results are without the postprocessing (the detection results are not merged).

stant time using the integral image. The calculation of histogram of each cell took 1 millisecond for one position of sliding window (80×80 pixels). Finally, the recognition time of feature vector depends on the chosen classifier.

5 CONCLUSION

In the paper, we proposed the interesting method for object description. The method is based on the fact that the appearance of the objects can be described using the function of temperature distribution. The key part of the proposed method is the process of temperature transfer that is carried out during a chosen time inside the image. This process is the most time-consuming part of the method and we will focus on the time complexity of the method in the future works. However, compared with the state-of-the-art descriptors, the proposed descriptors achieved the very good detection results and we will also focus on detection of other objects of interest using this method.

ACKNOWLEDGMENTS

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

REFERENCES

Berg, T. L., Berg, A. C., Edwards, J., and Forsyth, D. (2005). Who's in the picture. In Saul, L. K., Weiss, Y., and Bottou, L., editors, *Advances in Neural Information Processing Systems 17*, pages 137–144. MIT Press, Cambridge, MA.

Bosch, A., Zisserman, A., and Munoz, X. (2007). Representing shape with a spatial pyramid kernel. In *Proceedings of the 6th ACM international conference on Image and video retrieval*, CIVR '07, pages 401–408, New York, NY, USA. ACM.

Boser, B. E., Guyon, I. M., and Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. In *Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory*, pages 144–152. ACM Press.

Dalal, N. and Triggs, B. (2005). 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.

Felzenszwalb, P., Girshick, R., McAllester, D., and Ramanan, D. (2010). Object detection with discriminatively trained part-based models. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(9):1627–1645.

Ferrari, V., Marin-Jimenez, M., and Zisserman, A. (2008). Progressive search space reduction for human pose estimation. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8.

Freund, Y. and Schapire, R. E. (1995). A decision-theoretic generalization of on-line learning and an application to boosting. In *Proceedings of the Second European Conference on Computational Learning The-*

Figure 9: The detection results of $HETF_{PCA}$. The results are without the postprocessing (the detection results are not merged).

ory, EuroCOLT '95, pages 23–37, London, UK, UK. Springer-Verlag.

Fusek, R., Sojka, E., Mozdren, K., and Surkala, M. (2013). Energy-transfer features and their application in the task of face detection. In *Advanced Video and Signal Based Surveillance (AVSS), 2013 10th IEEE International Conference on*, pages 147–152.

Hadid, A., Pietikainen, M., and Ahonen, T. (2004). A discriminative feature space for detecting and recognizing faces. In *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, volume 2, pages II–797–II–804 Vol.2.

Lee, K., Ho, J., and Kriegman, D. (2005). Acquiring linear subspaces for face recognition under variable lighting. *IEEE Trans. Pattern Anal. Mach. Intelligence*, 27(5):684–698.

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

Lienhart, R. and Maydt, J. (2002). An extended set of haar-like features for rapid object detection. In *Image Processing. 2002. Proceedings. 2002 International Conference on*, volume 1, pages I–900–I–903 vol.1.

Ojala, T., Pietikäinen, M., and Harwood, D. (1996). A comparative study of texture measures with classification based on featured distributions. *Pattern Recognition*, 29(1):51–59.

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

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

Tan, X. and Triggs, B. (2010). Enhanced local texture feature sets for face recognition under difficult lighting conditions. *Image Processing, IEEE Transactions on*, 19(6):1635–1650.

Viola, P. and Jones, M. (2001). 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.

Zhang, L., Chu, R., Xiang, S., Liao, S., and Li, S. Z. (2007). Face detection based on multi-block lbp representation. In *Proceedings of the 2007 international conference on Advances in Biometrics*, ICB'07, pages 11–18, Berlin, Heidelberg. Springer-Verlag.