

Face Recognition Based on the Coefficient Tree for Three Scale Wavelet Transformation

Wang Lyanpen^{a, *}, O. G. Petrosyan^a, and Du Jianming^a

^aBauman Moscow State Technical University, Moscow, 105005 Russia

*e-mail: jswanglp@gmail.com

Received September 11, 2018; revised September 28, 2018; accepted October 12, 2018

Abstract—A new method for face recognition is presented in the paper. The new method based on the coefficient tree for three-scale wavelet transformation is presented for solving a problem on feature separation. The hidden Markov model is used for classifying face image features.

Keywords: image feature, coefficients tree, hidden Markov model, recognition accuracy

DOI: 10.3103/S0146411619080315

INTRODUCTION

In crowded modern towns, it becomes necessary to ensure security. This problem correlates with the problem of face recognition and it is one of the main practical problems whose solution contributes to the development of the theory of pattern recognition [1]. This field appeared at the beginning of the 1980s, but its active development started in the 1990s, when information retrieval systems for face recognition and personal identification were created.

The process of person recognition is based on face recognition according to their images and consists in comparing the image with the features of facial images in a database. The process of face recognition can be separated into three main stages (Fig. 1) [2]: to detect faces in images, to separate features (specificities), and to classify features.

For solving the problem on detecting faces in images, we present an algorithm based on the Viola-Jones procedure [3], which is the most popular among similar methods [4] and is characterized by high speed and acceptable accuracy. The Viola-Jones face detector is based on the main three ideas [5]: to present the image integrally; to generate a classifier by using an adaptive boosting algorithm (AdaBoost); and to use a procedure for combining classifiers into a cascade structure. These ideas make it possible to generate a robust face detector able to operate in online mode [6].

The experimental simulation for face recognition is performed in the Python 3.5 environment. Data set FaceWarehouse: <http://gaps-zju.org/facewarehouse/>, 150 people, for each person, 24 images in a natural background, among which 10 images are for training the model of the person, and 14 other images for face recognition. It is known that for face recognition using neural networks, a good deal of training data is required. On the other hand, in practice, in training the model for face recognition, it is not always necessary to have large volumes of training data; the application of neural networks for face recognition leads to retraining. The procedure presented in our paper can solve this problem. Faces in a database are very different over race, age, and sex. For each person in the database, there are different facial expressions with illumination and contrast features, which is useful for verifying that the algorithm is not sensitive to

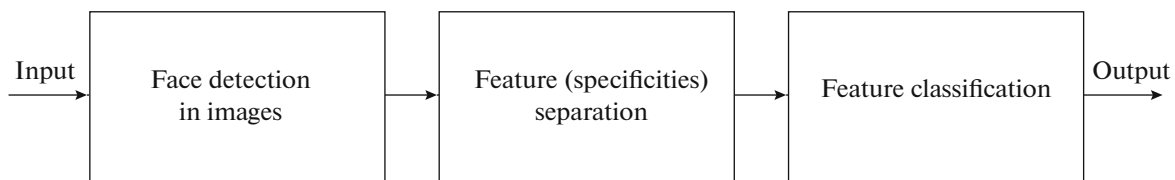


Fig. 1. Block diagram of face recognition.

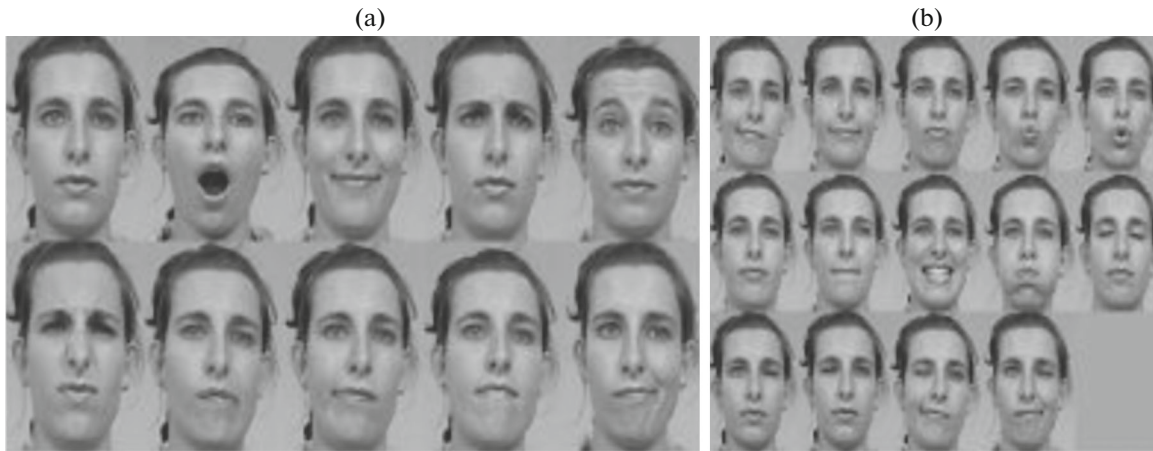


Fig. 2. Face images obtained using the Viola-Jones method: (a) training images; (b) images for testing. Faces detection in images.

these factors, but has a great ability to generalize. All face images are separated with the help of the Viola-Jones procedure.

Image features are separated into global and local. The essence of the algorithm for early face recognition is as follows: to separate and classify global features. There are two algorithms for face recognition using global features: Eigenface and Fisherface. Algorithm Eigenface uses the principal component method for face recognition. Algorithm Fisherface – is the method of linear discriminant analysis [7]. Algorithms for recognizing global features are sensitive to image illumination and contrast and to facial expression. Therefore, if the environment changes, the reliability and accuracy of face recognition decreases.

Face recognition algorithms based on local image features are generated to overcome these disadvantages. The best known are as follows:

1. Histogram of Oriented Gradients (HOG) [8], the main idea of the algorithm for forming the HOG descriptor is as follows: the object in image area can be described by edge directions or by the distribution of brightness gradients.
2. Local Binary Patterns (LBP) [9], which is a description of the neighborhood of the image pixel in binary form.
3. SIFT (scale-invariant feature transform, (SIFT) descriptor [10], this descriptor searches the reference points and generates a feature vector.
4. An improved version and combination of reference points detector FAST and binary descriptors BRIEF (oriented FAST and rotated BRIEF, ORB) [11], which describes function points by using a binary string based on FAST and BRIEF algorithms.

Figure 3 depicts the local features of face images, which are not sensitive to illumination variation, contrast and face expression. All of them are characterized by one common disadvantage: the dimensionality of feature vectors is too high (for HOG the dimensionality is 291060, for LBP-27000, for SIFT-15744, and for ORB-5280), which does not contribute to a later classification of features.

In the paper, we present the procedure for extracting image features on the basis of a tree of coefficients for a three-scale wavelet transform, which not only preserves the characteristics of the methods that use global features, but has advantages intrinsic to the methods based on local features.

The final step in face recognition is the classification of the features using an already trained classifier such as Random Forests (RFs) [12], Support Vector Machines, (SVMs) [13], or Adaptive Boosting, (AdaBoost) [14]. These three classifiers are the most usable and most accurate. It has been proved [15] that the RF method results in retraining for classification in the presence of noise, and in addition, there is no visual presentation of the decision-making process and it is difficult to interpret the solutions. For the SVM method, there are not enough parameters for tuning and when we fix the core, the only one varying parameter is the error coefficient C , and the model is trained more slowly. Very often Adaptive Boosting causes cumbersome compositions consisting of hundreds of algorithms. For such compositions, it is

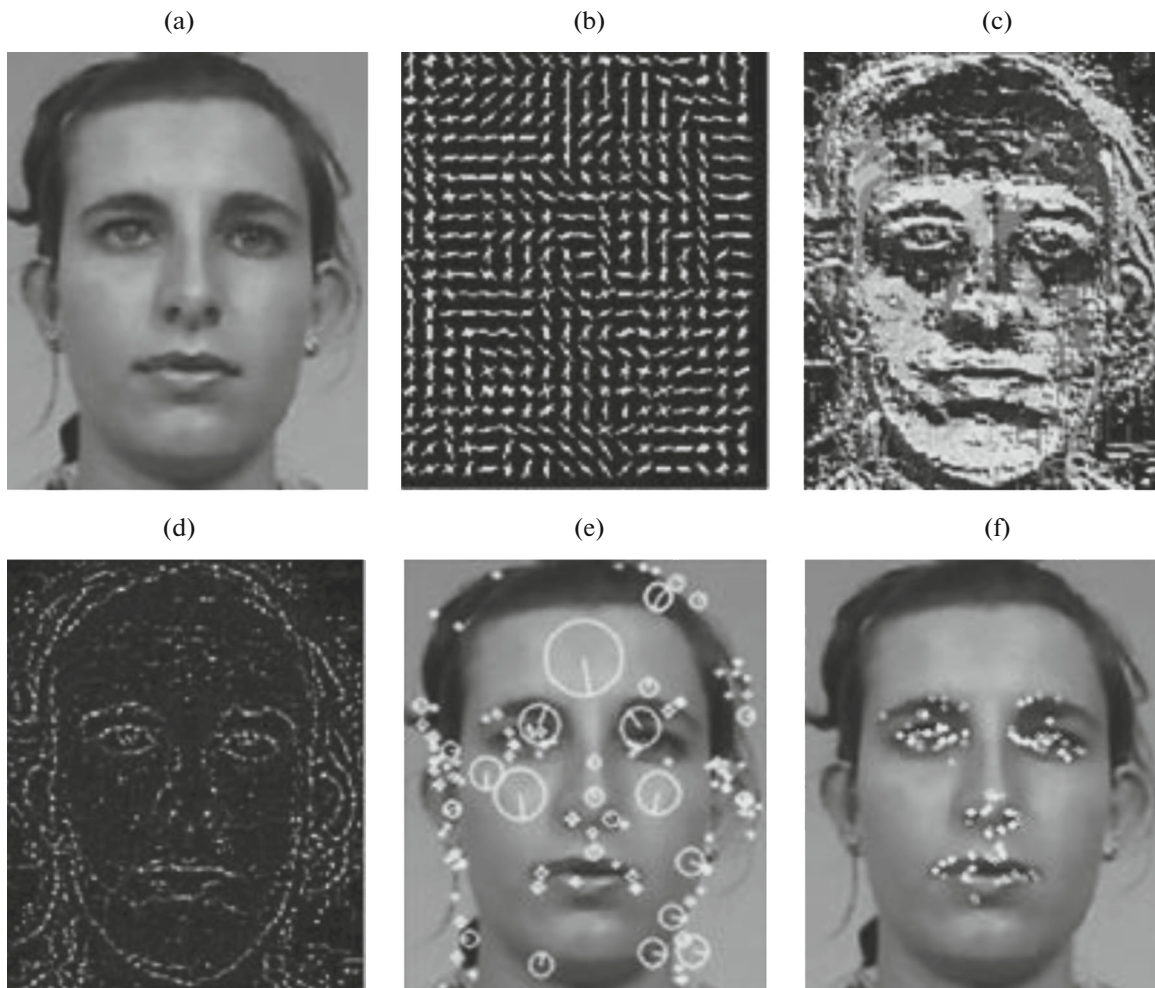


Fig. 3. Different known local features: (a) initial face image with resolution of 180×150 ; (b) HOG features (dimensionality is 291 060); (c) original LBP features (vicinity radius $r = 2$, number of points $N = 8$, dimensionality 27000); (d) improved LBR features (vicinity radius $r = 2$, number of points $N = 8$, dimensionality 27000); (e) SIFT features (dimensionality is 15744); (f) ORB features (dimensionality is 5280).

impossible to perform proper interpretations; they require large amounts of memory to store the basic algorithms and a significant time for calculation of classifications.

In our paper, we use a Hidden Markov Model (HMM), which is a powerful technique for classifying the features of different objects; it is characterized by a simple mathematical structure, high efficiency and recognition accuracy, and needs a short training time.

TREE OF COEFFICIENTS FOR A THREE-SCALE WAVELET TRANSFORM

According to the theory of wavelet analysis, any square-integrable function $f(x)$ can be restored with the help of an inversion formula [16]:

$$f(x) = C_{\Psi}^{-1} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \langle f(x), \Psi^{a,b}(x) \rangle \Psi^{a,b}(x) \frac{dadb}{a^2}, \quad (1)$$

where $\Psi^{a,b}(x) = |a|^{-1/2} \Psi\left(\frac{x-b}{a}\right)$ is the set of wavelet functions, operation $\langle \cdot, \cdot \rangle$ means scalar product in space L^2 and a, b are shift and compression parameters with the restriction $a \neq 0$. Constant C_{Ψ}^{-1} depends only on $\Psi^{a,b}(x)$ and is determined by the following formula:

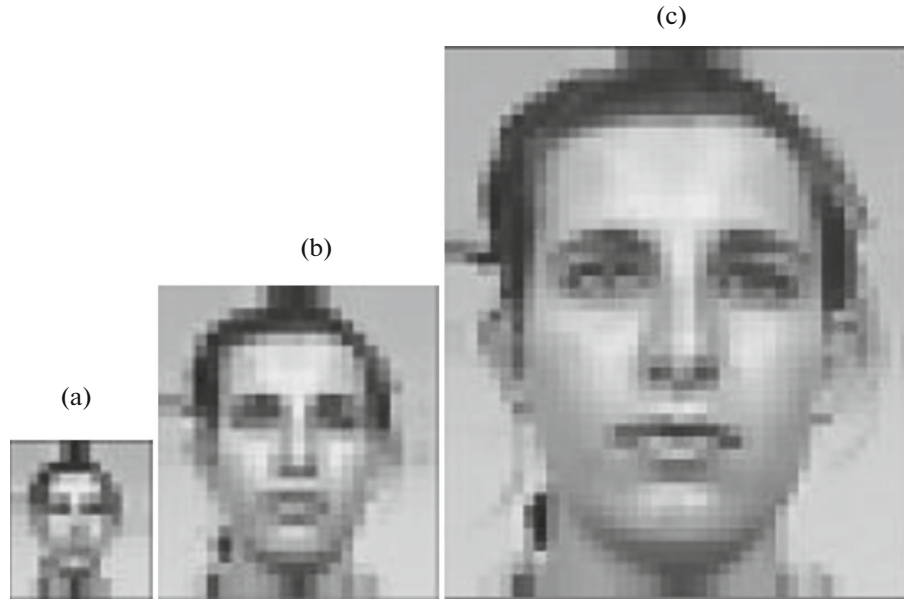


Fig. 4. Approximation coefficients under different scales: (a) in 4th scale (resolution is 15×14); (b) in 3rd scale (resolution is 30×28); (c) in 2nd scale (resolution is 60×56);

$$C_\psi = 2\pi \int_{-\infty}^{\infty} |\hat{\psi}(\xi)|^2 |\xi|^{-1} d\xi, \tag{2}$$

where $\hat{\psi}(\xi)$ is the Fourier transformation for the wavelet function.

Let us introduce a concept of 2D discrete wavelet transform for Daubechies function $f(x, y)$ [17] with dimensionality of $M \times N$:

$$W_\varphi(j_0, m, n) = \frac{1}{\sqrt{MN}} \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} f(x, y) \varphi_{j_0, m, n}(x, y), \tag{3}$$

$$W_\psi^i(j, m, n) = \frac{1}{\sqrt{MN}} \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} f(x, y) \psi_{j, m, n}^i(x, y), \quad i = \{H, V, D\}, \tag{4}$$

where $W_\varphi(j_0, m, n)$ characterizes the approximation coefficients for function $f(x, y)$ in scale j_0 , where j_0 is the arbitrary initial scale; $W_\psi^i(j, m, n)$ are wavelet transform coefficients, where $i = \{H, V, D\}$ (H are coefficients for horizontal, V —vertical, and D —diagonal details for scales $j \geq j_0$). Let us calculate the approximation coefficients $W_\varphi(j_0, m, n)$ in fourth, third and second scales (in this case $j_0 = 4, 3, 2$), as it is shown in Fig. 4.

According to the theory of fast wavelet transformation, the high-scale approximation coefficients are obtained by low-pass filtering and sparse sampling of the approximation coefficients in a neighboring low scale. Let us use the bilinear interpolation so that the total number of low-scale approximation coefficients is exactly twice the number of high-scale approximation coefficients in the long and wide directions.

The correlation between approximation coefficients in the neighboring three scales can be established as is shown in Fig. 5.

For each approximation coefficient in the fourth scale, it is possible to generate a 21-dimensional vector using the coefficient tree structure for a three-scale wavelet transform. Hereby, we obtain a 21-dimensional tensor corresponding to the approximation coefficients in the 4-th scale in Fig. 6 and the dimensionality of image features is $15 \times 14 \times 21 = 4410$.

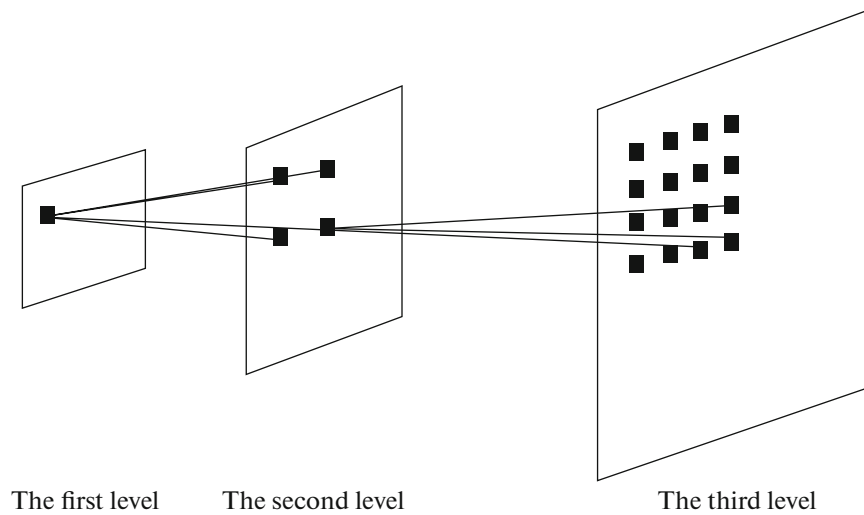


Fig. 5. The structure of coefficient tree for three-scale wavelet transform.

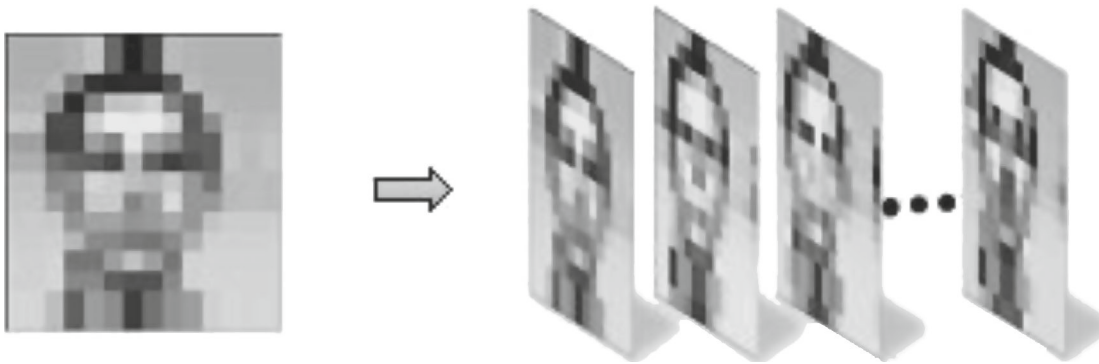


Fig. 6. 21-dimensional feature tensor corresponding to the approximation coefficients in 4th order scale.

HIDDEN MARKOV MODEL

The hidden Markov process is an arbitrary process, which at each moment of time $t \in \{1, \dots, T\}$ is in one of the states $s \in \{S_1, \dots, S_N\}$ and transits to the new state according to transition probabilities [18]. These states are hidden from the observer and are seen only in several patterns, which are a sequence of observations generated in the hidden states. Figure 7 depicts a simple hidden Markov model (HMM) with three hidden states:

The hidden Markov model is described by a matrix of transition probabilities, a matrix of observed symbols and by initial state probabilities [19]: $\lambda = (A, B, \Pi)$, whose definitions are presented below:

The matrix of transition probabilities $A = \{a_{ij}\}$, $i, j = 1, \dots, N$, the transition probability $a_{ij} = P\{q_t = S_j \mid q_{t-1} = S_i\}$, N is the number of hidden states in the model, S are hidden states $S = \{S_1, S_2, \dots, S_N\}$, q_t is the hidden state at time t ($q_t \in S, 1 \leq t \leq T$) and T is the length of the observed sequence. The restrictions:

$$\begin{cases} 0 \leq a_{ij} \leq 1 \\ \sum_{j=1}^N a_{ij} = 1 \end{cases} \quad (5)$$

The matrix of observed sequence probabilities is written as follows: $B = \{b_j(o_t)\}$, $j = 1, \dots, N$, where $b_j(o_t) = P\{o_t \mid q_t = S_j\}$, where o_t is a symbol observed at the time $t = 1, \dots, T$.

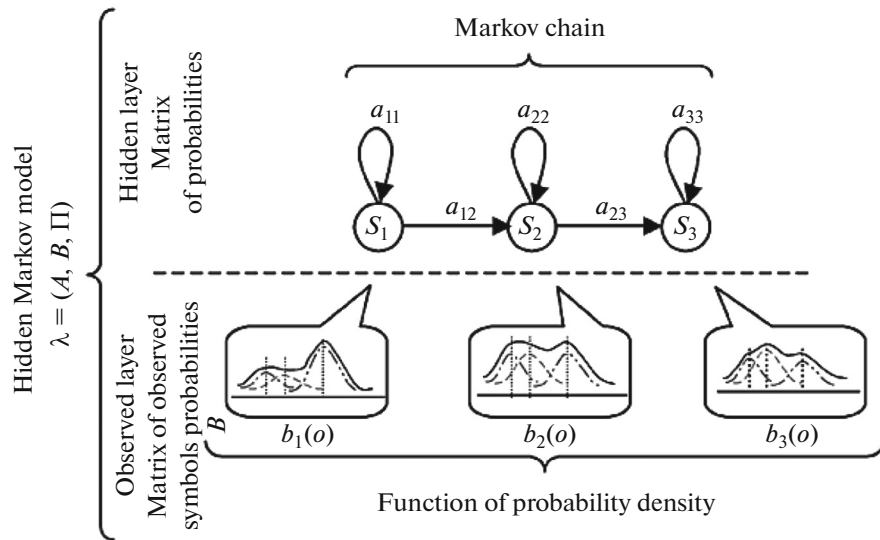


Fig. 7. Hidden Markov model.

The vector of initial state probabilities $\Pi = \{\pi_i\}, i = 1, \dots, N$, where $\pi_i = P\{q_1 = S_i\}$, and q_1 is the hidden state at the initial time.

HIDDEN MARKOV MODEL TRAINING AND FACE RECOGNITION IMPLEMENTATION

The process of model training and face recognition includes the following steps:

Step 1. To separate features of face images using the coefficients tree for three-scale wavelet transforms as an observed sequence O .

Step 2. To de-correlate features with the help of Principal Component Analysis (PCA).

Step 3. To generate a general model $\lambda = (A, B, \Pi)$, to determine the number of hidden states N , for initial state probabilities $\Pi = \{\pi_i\}$ we determine $\pi_1 = 1, \pi_i = 0 (i \neq 1)$, for the matrix of transition probabilities $A = \{a_{ij}\}$ we determine $a_{ij} = 1 (i = j), a_{ij} = 0 (i \neq j)$, the matrix of observed sequences probabilities $B = \{b_j(o_t)\}$ [20]:

$$b_j(o_t) = \sum_{k=1}^M c_{jk} (2\pi)^{-N/2} |\Sigma_{jk}|^{-1/2} \exp[-(o_t - \mu_{jk})^T \Sigma_{jk}^{-1} (o_t - \mu_{jk})] / 2, \tag{6}$$

where M is the dimensionality of the observation alphabet, c_{jk} is the i th weight coefficient for the mix of normal distributions in j th hidden state, μ_{jk} and Σ_{jk} are the mean value and the covariation matrix of the k th mixture component in the j th hidden state, $j = 1, \dots, N, k = 1, \dots, M$:

$$\mu_{jk} = \left(\sum_{t=1}^{E_{jk}} o_t^{(j,k)} \right) / E_{jk}, \tag{7}$$

$$\Sigma_{jk} = \left[\sum_{t=1}^{E_{jk}} (o_t^{(j,k)} - \mu_{jk})(o_t^{(j,k)} - \mu_{jk})^T \right] / E_{jk}, \tag{8}$$

where E_{jk} is the length of the observed sequence, respectively, of the k th component of the normal distribution mixture in the j -th hidden state, and $o_t^{(j,k)}$ are the corresponding observed data.

Figure 8a depicts the result of reducing the dimensionality of the 21-dimensional tensor to three. Let us divide each image in the feature tensor into several parts according to the number of hidden states ($N = 5$). The area marked by the dashed line in Fig. 8b corresponds to the observed sequence of the first hidden state.

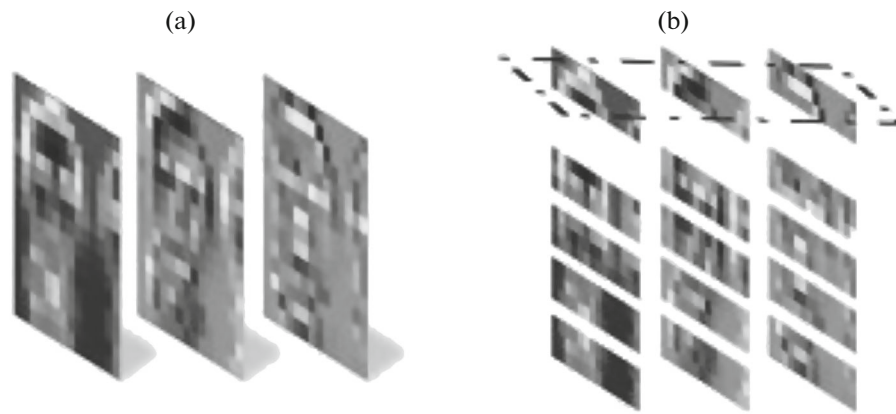


Fig. 8. (a) 3D feature tensor, dimensionalities are reduced by PCA method; (b) feature segmentation according to the number of hidden states ($N = 5$).

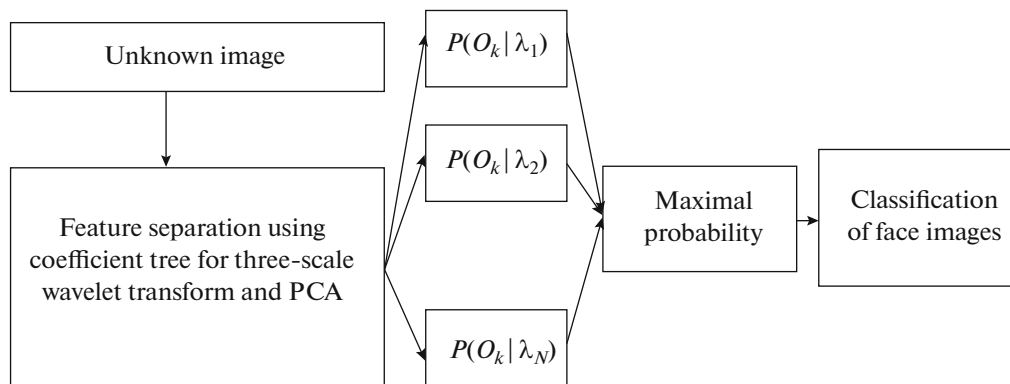


Fig. 9. Block diagram of face recognition by hidden Markov model.

Step 4. To train the model using the Baum-Welch algorithm [21], from which we obtain the model $\lambda_1 = (A_1, B_1, \Pi_1)$ corresponding to the first person.

Step 5. To repeat steps 2, 3 and 4, and we obtain the trained models for all faces.

Step 6. To separate features of face image O_k in the test base using the coefficient tree for the three-scale wavelet transform and the PCA method. We calculate the probability $P(O_k | \lambda_i)$ that the model λ_i generates O_k . If the n th model λ_n is characterized by the highest probability of generating sequence O_k , assign the corresponding image to the n th person. A block diagram for face recognition using the hidden Markov model is shown in Fig. 9.

Figure 10 depicts the accuracy of face recognition when varying the number of hidden states S and the dimensionality of feature tensor K .

Axis x in Fig. 10 characterizes the dimensionality of feature tensor K , axis y shows percentages. Line $\cdots*$ shows the accuracy of face recognition when the number of hidden states is three, line $-o$ corresponds to the case when the number of hidden states is four, line \rightarrow corresponds to the case when the number of hidden states is five, line $-■$ corresponds to the case when the number of hidden states is six. It is seen that if the number of hidden states is five, and the dimensionality of the feature tensor is three, the accuracy of face recognition is maximal at 95.71%.

WEIGHT FUNCTION OF A 2D NORMAL DISTRIBUTION

For each coefficient in the fourth scale with resolution of 15×14 there is a 21-dimensional vector. Any point in 4-th scale corresponds to a 21-dimensional vector. Let us fill all 21-dimensional data in a 5×5

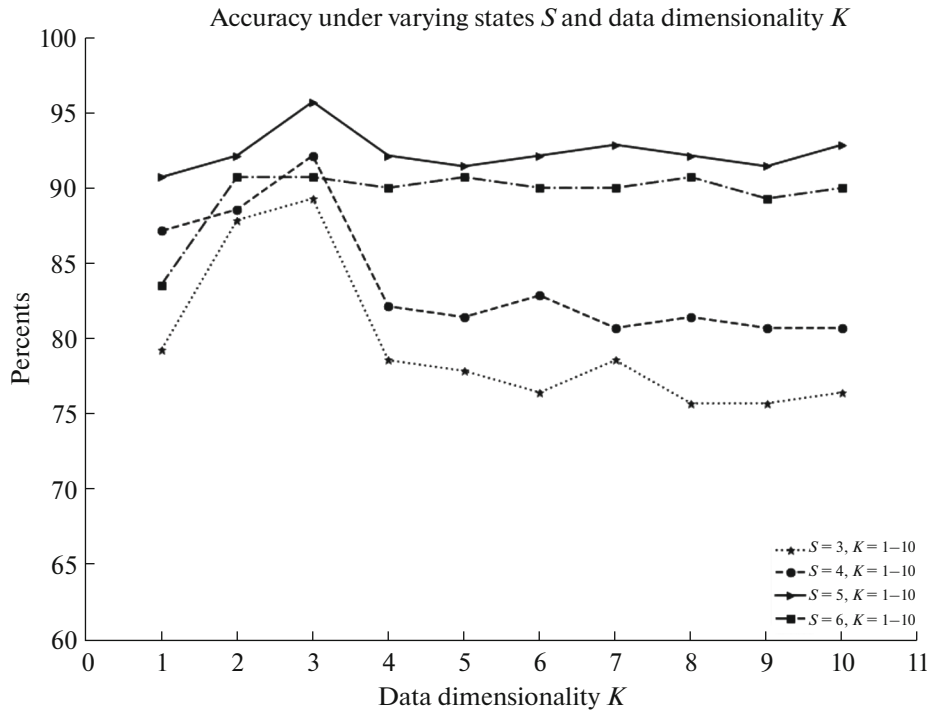


Fig. 10. Recognition accuracy when varying the hidden states and dimensionalities of feature tensor.

square (Fig. 11). Numbers in the square indicates the point's level in the coefficient tree of the wavelet transform. Values of four points in the square corner can be calculated by means of bilinear interpolation. Let us assume that all surrounding points introduce a contribution to this area, the nearer to the center, the higher the contribution, and the weight coefficients of the contribution are determined by a function of a 2D normal distribution:

$$f(x, y) = \frac{(1 - \rho_{xy}^2)^{\frac{1}{2}}}{2\pi\sigma_x\sigma_y} \exp \left\{ -\frac{1}{2(1 - \rho_{xy}^2)} \left[\frac{(x - \mu_x)^2}{\sigma_x^2} - \rho_{xy} \frac{2(x - \mu_x)(y - \mu_y)}{\sigma_x\sigma_y} + \frac{(y - \mu_y)^2}{\sigma_y^2} \right] \right\}, \quad (9)$$

where, ρ_{xy} is the correlation coefficient of x and y , μ_x, μ_y —the mathematical expectations, and σ_x, σ_y are variances. We determine $\rho_{xy} = 0, \mu_x = \mu_y = 0, \sigma_x = \sigma_y = \sigma$. If condition $(6\sigma + 1)(6\sigma + 1) = 5 \times 5$ is true, the sum of all weight coefficients is equal to unity, and $\sigma = 0.67$. Figure 12 depicts the weight function.

If the weighted features are used for training models, the accuracy of face recognition increases. Figure 13 depicts the recognition accuracy under different numbers of hidden states and different numbers of training images.

In Fig. 13, axis x depicts the number of training face images K , axis y depicts the recognition percentage. Line $\cdots*$ is the accuracy of face recognition when the number of hidden states S is three, line $-o-$ corresponds to the case when the number of hidden states is four, line \rightarrow corresponds to the case when the number of hidden states is five, line $\rightarrow\blacksquare$ corresponds to the case when the number of hidden states is six. It is seen that the recognition accuracy increases if the number of training face images increases, and when $K = 10$ and $S = 5$, the accuracy of face recognition is maximal at 98.57%.

COMPARISON WITH OTHER ALGORITHMS

Table 1 depicts the accuracy of face recognition by different algorithms.

WF – Wavelet Features.

WC – Weighted Coefficients.

Multi-NN – Multilayer Neural Network (network parameters: 4 layers, 1200 neurons/layer, Dropout 0.2).

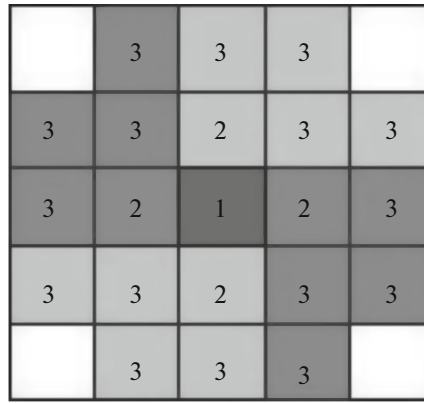


Fig. 11. Position of 21-dimensional feature tensor in 5×5 square.

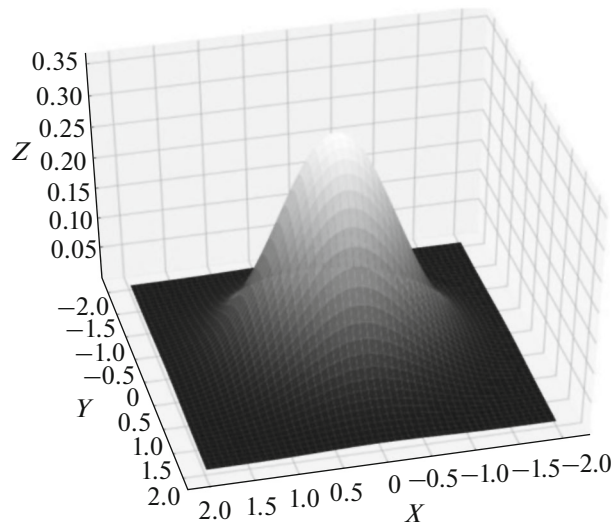


Fig. 12. 2D weight function.

CNN – Convolutional Neural Network (network parameters: 7 layers with regularization L2, two $32 \times 3 \times 3$ convolutional layers, two $64 \times 3 \times 3$ convolutional layers).

From the table it is seen that since the number of training models is small, the model of convolutional neural network is retrained and the recognition accuracy for testing is very unstable. Maximal recognition accuracy obtained in numerous experiments is 99.14%, and minimal recognition accuracy is 96.43%.

In reality, the data volume used for training is insufficient and it is easy to trigger retraining when using a convolutional neural network, as shown in Table 1. For face recognition with small samples, the procedure presented in this paper has advantages in training speed and recognition accuracy. Table 2 depicts the

Table 1. Accuracies of face recognition by different algorithms

Recognition method	Eigenface	Fisherface	LBP+SVM	WF + WC + SVM	WF + WC+ RF
Recognition accuracy	84.57%	90.86%	94.57%	96.28%	97.86%
Recognition method	WF + WC + Adaboost	WF + PCA + HMM	WF + WC + HMM	WF + WC + Multi-NN	CNN
Recognition accuracy	90.71%	95.71%	98.57%	96.86%	max-99.14% min-96.43%

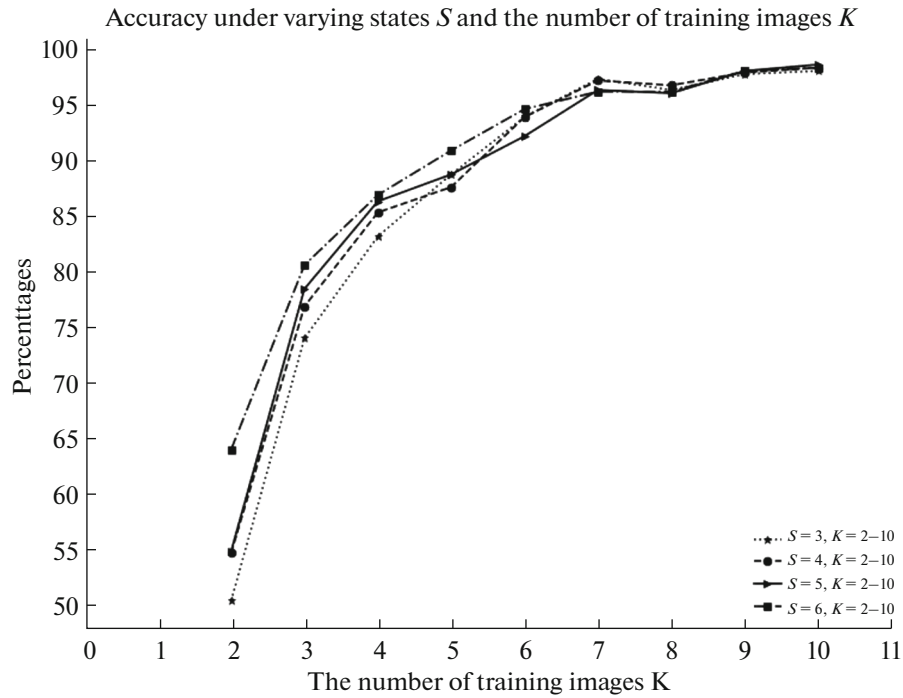


Fig. 13. Accuracy of face recognition when varying hidden states and the number of training images.

time needed for training for different algorithms. It is seen that the CMM method is characterized by the shortest time for model training, 351.5 ms, and the CNN method is characterized by the highest time for training, 1323.7 s, with video card GT650M; if super video card Tesla P100 is used, the time for model training decreases to 228.6 s.

CONCLUSIONS

The new method for face recognition based on a coefficient tree for a three-scale wavelet transform and on a hidden Markov model (HMM) is presented in the paper. The wavelet transform and the weight function of a 2D normal distribution are used sequentially for extracting features. The hidden Markov model is used for comparing features.

The presented procedure is used for simpler models (with respect to other methods). It is characterized by the following advantages:

- (1) The dimensionality of image features is lower.
- (2) It is not necessary to retrain the model if new samples are added, i.e., only the new samples are trained individually.
- (3) The training rate increases.
- (4) The accuracy of face recognition is higher.

Table 2. Time for model training for different algorithms

Recognition method	WF + WC + HMM	WF + WC + RF	WF + WC + AdaBoost	WF + WC + SVM
Time for model training	351.5 ms	703.5 ms	57.1 s	425.2 ms
Recognition method	WF + WC + Multi-NN	CNN		
Time for model training	84.6 s	1323.7 s (GT650M) 228.6 s (Tesla P100)		

As a result of the experiment, we found that: if the bio-orthogonal wavelet set bior 1.3 is used, the accuracy of face recognition is 98.57%, if the number of hidden states is five, and the number of training images is ten.

For this method the time required for recognizing an image feature of each model is 0.12 ms with processor Intel Core i7-3630QM @ 2.40 GHz, meaning that the recognition rate is acceptable if the total number of models is less than 1000. Since each model is independent, the present method can be used in the majority of practical applications, if parallel calculation is possible.

CONFLICT OF INTEREST

The authors declare that they have no conflicts of interest.

REFERENCES

1. Mizyukin, A.V. and Moskvina, D.A., Assessment of the complexity of solving the face recognition problem using the principles of fractal image compression, *Probl. Inf. Bezop., Komp'yut. Sist.*, 2013, no. 3, pp. 27–39.
2. Zegzhda, D.P., Moskvina, D.A., and Bosov, Yu.O., Pattern recognition based on fractal compression, *Probl. Inf. Bezop., Komp'yut. Sist.*, 2012, no. 2, pp. 86–90.
3. Moskvina, D.A. and Gluhov, V.V., Using principles of fractal image compression for complexity estimation of the face recognition problem, *Nonlinear Phenom. Complex Syst.* (Dordrecht, Neth.), 2014, vol. 17, no. 3, pp. 231–241.
4. Maslii, R.V., Use of local binary patterns for face recognition in grayscale images, *Inf. Tekhnol. Komp'yut. Tekh.*, 2008, no. 4, pp. 104–110.
5. Viola, P. and Jones, M.J., Robust real-time face detection, *Int. J. Comput. Vision*, 2004, vol. 57, no. 2, pp. 137–154.
6. Spitsyn, V.G., Bolotova, Yu.A., and Shabalina, N.V., Face recognition based on the method of principal components using Haar and Daubechies wavelet descriptors, *Nauchn. Vizualizatsiya*, 2016, no. 5, pp. 103–112.
7. Rogozin, O.V. and Kladov, S.A., Comparative analysis of face recognition algorithms in the problem of visual identification, *Inzh. Zh.: Nauka Innovatsii*, 2013, no. 6, pp. 74–98.
8. Yashchina, M.V. and Tolmachev, A.A., Pattern recognition methods for assessing the characteristics of pedestrian flows, *Telekommun. Transp.*, 2017, no. 8, pp. 45–51.
9. Dong Chen Khe and Li Van, Texture block, texture spectrum, and texture analysis, *Geofiz. Distantionnoe Zondirovanie*, 1990, no. 4, pp. 509–512.
10. Lowe, D.G., Distinctive image features from scale-invariant keypoints, *Int. J. Comput. Vision*, 2004, vol. 60, no. 2, pp. 91–110.
11. Rublee, E., Rabaud, V., Konolige, K., and Bradski, G., ORB: An efficient alternative to SIFT or SURF, *2011 International Conference on Computer Vision*, 2011, pp. 2564–2571.
12. Chistyakov, S.P., Random forests: A review, *Tr. Karel. Nauchn. Tsentra Ross. Akad. Nauk*, 2013, no. 1, pp. 117–136.
13. Ben-Hur, A., Horn, D., Siegelmann, H.T., and Vapnik, V., Support vector clustering, *J. Mach. Learn. Res.*, 2001, vol. 2, pp. 125–137.
14. Wei Gao and Zhi-Hua Zhou, On the doubt about margin explanation of boosting, *Artif. Intell.*, 2013, vol. 203, pp. 1–18.
15. Gehrke, J., Ramakrishnan, R., and Ganti, V., Rainforest: A framework for fast decision tree construction of large datasets, *Data Mining Knowl. Discovery*, 2000, vol. 4, no. 2, pp. 127–162.
16. Daubechies, I., *Ten Lectures on Wavelets*, SIAM, 1992.
17. Gonzalez, R.C. and Woods, R.E., *Digital Image Processing*, Pearson Education, 2011.
18. Chzhu Sian and Tsao Lin, *Wavelet Analysis and Its Application in Digital Image Processing*, Beijing, 2012.
19. Gul'tyaeva, T.A., Popov, A.A., and Uvarov, V.E., The use of hybrid computing to optimize the recognition of sequences described by hidden Markov models, in *Sbornik nauchnykh trudov novosibirskogo gosudarstvennogo tekhnicheskogo universiteta* (Collection of Scientific Papers of Novosibirsk State Technical University), Novosibirsk, 2015, pp. 42–55.
20. Andrev, Zh.V., Error limits for convolutional codes and an asymptotically optimal decoding algorithm, *Osn. Tsifr. Besprovodn. Mira*, 1966, no. 2, pp. 41–50.
21. Baum, L.E., Petri, T., and Suulz, D., Maximization technique arising in the statistical analysis of probability functions of Markov chains, *Letopis' Mat. Stat.*, 1970, no. 1, pp. 164–171.