3D computer vision system for face recognition

Ivan A. Matveev, Alexander B. Murynin

Computing Center of Russian Academy of Sciences
Moscow, Russia

Abstract

One of the most popular and practically valuable problems in the field of pattern analysis and machine intelligence is human face recognition. A computer vision system for identification of human faces is presented. The system is designed as an automatic checkpoint. Two scenarios of system behavior are developed. The first one assumes verification of personal data, entered by visitor by a card reader. The other performs identification by only visitor's biometrics. The system performs remote measurements of face features of different types. In addition to stereoscopic images inputted to computer from cameras the models can use voice data and some person physical characteristics such as person's height, measured by imaging system. The diversity of employed characteristics makes the system reliable and tolerant.

Keywords: *computer vision, face recognition, 3D vision*

1. INTRODUCTION

Computer systems for automatic person recognition based on biometrics [1][2] are being developed intensively in connection with creation of computer security systems. Such systems based on recognition of human face and voice have a significant advantage comparing to systems, that use characteristics of fingerprints, iris etc., because recording of face and voice characteristics doesn't require any physical contact between the person and sensors of the system. It is urgent to create a recognition system, which first, ensures sufficient defense from unauthorized access, and second, implements measurements with minimal discomfort for users. For this purpose it is advisable to make use of methods, based on real time reconstruction of 3-D shape of human face [3]. The usage of 3-D surface increases reliability of the security system because a possibility is eliminated to pass the control by presenting a photo of some other person made in life-size.

The system presented contains special equipment and software that is a complex of various image processing and pattern recognition algorithms. Common structure of recognition system based on measurements of human biometrics is described and two laboratory models developed. Different modes of possible system behavior originating from two different types of recognition are presented.

2. SYSTEM STRUCTURE

The general structure of the recognition system is illustrated in Figure 1. The system may be treated as a compound of hardware (equipment) and software (algorithms). The authors have developed two laboratory models of system that are denoted here as 'visible-range' and 'infrared' ones. The software as well can be used in two modes: 'verification' and 'identification'. The detailed discussion of each point will follow and here it would be only emphasized, that any of two software branches can be used with any of hardware models with minimal adjusting of image processing parameters.

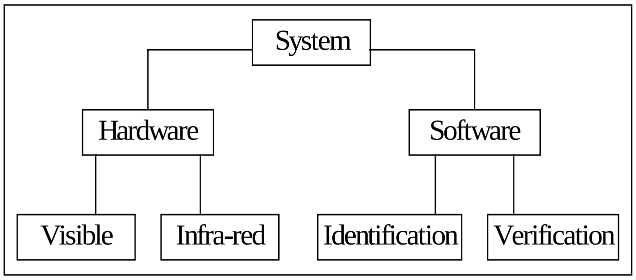

Figure 1 General system structure

Consider hardware structure of person recognition system based on remote measurements of some human biometrics. The model imitates an operation of an automatic entrance checkpoint that accomplishes control of visitors' access to some object. To carry out this task the system should be able to register data, process it and report results or perform some actions depending on them. Thus system should consist of the following functional blocks: input devices, analyzer block and effectors block. Input devices of the system described contain personal code reader, sound recorder, and video cameras. Analyzer consists of speech recognizer, images analyzer, decision-making algorithm. Effectors may include camera-positioning apparatus and installation like a turnstile for assuring limitations of physical access to a guarded object. Figure 2 sketches a principal scheme of the system. The system functionality can be extended by connecting additional equipment for biometrics measurement, for instance a weight sensor, an apparatus for fingerprint inputting, iris viewing and so on.

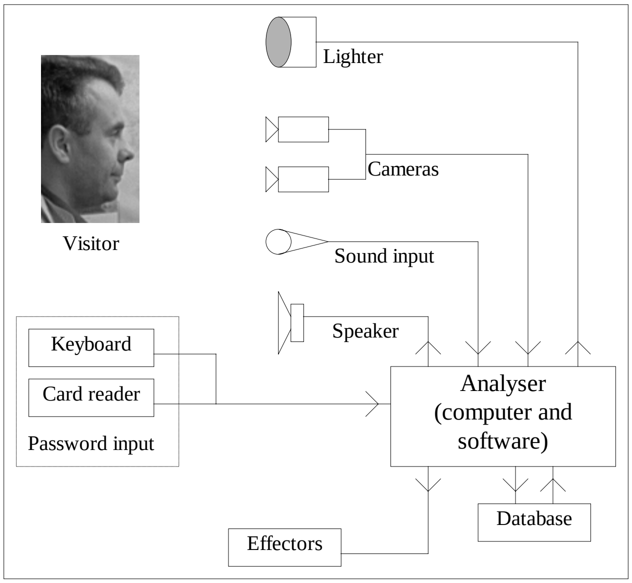

Figure 2 Possible components of system

The first question considering recognition is what does one actually mind saying "*recognition*". Indeed, recognition can be split at least into two types. Of course, no recognition can be performed without having some primarily registered and stored data that compounds a database (or knowledge base). It can be said that the storing procedure forms some informational space, in which all the following recognition processing is done. Moreover, elements of this space may be divided into several classes. In this terms one type of recognition can be described as follows: given some element of space and an identifier of some class in space determine whether this element belongs to the class or not. This type is called *verification* here. An example of verification is: the system is given a name of a person whom it does know and some photo and it should decide whether this is the photo of the person or not. Other type of recognition can be stated as follows: given an element of space the system should find a class, which it belongs to. This type is referred to as *identification*. An example: given a photo of person, the system should guess who is he (or maybe answer that it does not know him at all). It is interesting that identification can be performed by repetitive verifications: being given an element to identify one should carry out verifications of this element against all others in all classes and find a coincidence (or determine that the element does not belong to any of classes). Thus in general identification seems to be much more complex problem that verification, at least at calculations. The authors have developed two types of recognition system workflow, one performing verification, the other carrying out identification.

2.1 Visible-range model

This model system is complex - in addition to face recognition it verifies voice characteristics and passwords pronounced by the visitor. The system also measures other person's biometrics including height, weight if additional devices are connected. The model system hardware consists of: code input system based on a keyboard or smart-card reader, face images input subsystem, voice subsystem, electro-mechanical device adjusting camera position, some devices for measurement of other biometrics characteristics and finally, the core of recognition system that in turn is compounded of computer and database of standard characteristics. Images input subsystem contains two television cameras, lighter and two picture digitizers (frame-grabbers) each conjugated with a camera. Voice subsystem consists of microphone, sound digitizer and loud speaker that can be used to guide person to do some certain things or to report the recognition decision. The adjusting of camera position can be used as well to determine the height of person since the position of cameras in space is known at any moment. The system also includes a balance situated under a place where person should stay in order to be recognized. On the right side of Figure 3 you can see two cameras mounted on a column. The column can slide up- and downward adjusting cameras to better position for image grabbing and recognition thus measuring the height of person at the same time.

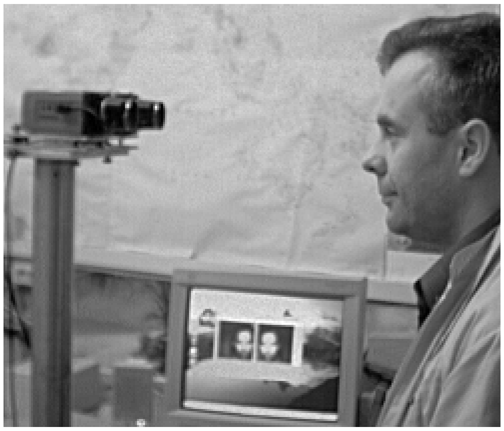

Figure 3 Appearance of visible range model

2.2 Infrared range model

This model is somewhat simpler as it includes fewer devices. It consists of cameras conjugated with video input cards (frame-grabbers), an infrared illuminator, two semi-transparent mirrors, speaker and microphone that all (with exception of frame-grabbers that are installed in computer) are mounted in a case. Mechanical adjustments of camera position and weight measurements are not performed here. Cameras and mirrors are specially oriented in order to allow users to view and position themselves in the way that is the most favorable for registering face images. One mirror is just a front panel of the case, cameras are situated closely behind it and the other mirror is lower and has different slope. A visitor should see his reflections simultaneously in both mirrors and at this moment he is in the best position for face registration. The range of wavelengths the system works at is near infrared. This allows achieving great intensity of illumination of face by lighter not disturbing the visitor. Figure 4 depicts the appearance of infrared model.

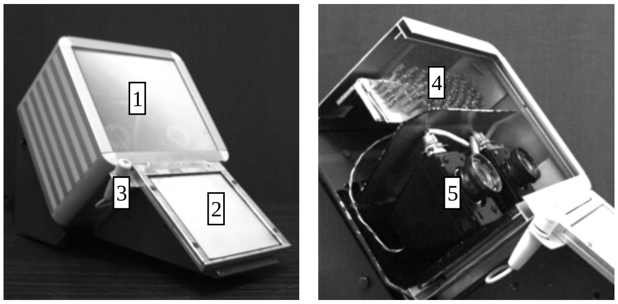

Figure 4 Appearance of infrared range model

1 *front panel mirror,* 2 *additional mirror,* 3 *microphone,* 4 *lighter,* 5 *camera*

3. FACE DETAILS

Further processing of images requires introduction of some informative model of the object to be recognized. We develop a model of human face based on information known a priori about 3-D surface of the object and spatial distribution of light scattering characteristics. To estimate 3-D shape of the face we use elevation maps derived from disparity distributions calculated from stereo images. Methods and algorithms for solution of this problem are presented in authors papers presented in [1][2][3].

In order to build a model of face, we assume, that there are some details of human face that must exist on each full face image: eyes, nose, mouth, eyebrows etc. The aggregate of templates of these face features their relative sizes and positions is treated as a parametric model of face and the problem of face recognition is formulated as a problem of calculating this aggregate of parameters and comparing them. Different techniques can be applied to find these parameters the following sequence is used in the system. Each step relies on some model assumption, which allows distinguishing the current object of interest from its environment and calculating its parameters.

The first step is, of course, the searching of face on the image. The model assumption here is that since we use lighters face is a bright object in the darker environment. We assume also, that face shape is close to elliptic and thus, search this “big bright elliptic object” on the image. The sophisticated method that takes into account different possible situations and negates various errors is presented in [4].

To other face features in the images we use template correlation technique applied to both 2-D spatial distributions of brightness and elevation maps estimated from disparities. The spatial position of a face part is determined by maximum of correlation coefficient defined as

$$C = \frac{< BT > - < B >< T >}{\sigma(B)\sigma(T)} \qquad (3)$$

where B is current image, T is template, < > is averaging operator, σ(.) is the mean square deviation in the region of search, BT is product of matrices B and T. Commonly a set of different templates T is used for increasing reliability of search results.

4. IMAGE SPACE

Image recognition problem deals with visual, audio or other type of information. In the case of visual information input image is a vector function of two coordinates $\vec{B}(x, y)$. $\vec{B}$ is a vector that may consist of brightness, hue, saturation, and relief in a 3-D case and the like. Current system works with stereo-images and relieves (elevation maps), reconstructed from these pairs. Also, bound the problem to discrete case, which is most important for automatic recognition systems. Thus, *x* and *y* take integer values in a limited range and image is a matrix which elements also take discrete or integer values in a limited range:

$$I = \begin{pmatrix} B_{11} & B_{21} & \cdots & B_{m1} \\ B_{12} & B_{22} & & \vdots \\ \vdots & & \ddots & \\ B_{1n} & \cdots & & B_{mn} \end{pmatrix}.$$

This 2-D matrix may be presented as 1-D vector: $\vec{I} = \begin{pmatrix} B_1 \\ B_2 \\ \vdots \\ B_N \end{pmatrix}$,

where $N = m * n$. Image space has N dimensions. Variety of images of all objects I_F is a subset of a set of all possible images or image space I (*Figure 5*).

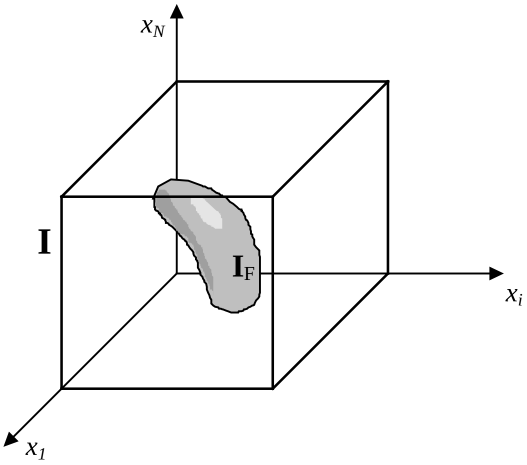

Figure 5. Set of objects' images I_F in the image space I.

Objects of recognition in our concern are human faces. Let us consider possible variations of features relating to one object of recognition. They are caused by different light conditions, different position of object, including shifting along x, y, z axes (z-shifting or zooming is equivalent to scaling, if size of object is small relatively to distance from camera) and rotating around this axes (tilt, turn, rotation), different facial expression, some features, changing with time (hairs, closing part of face, glasses, mustache and beard, etc.). Some of them may be and actually are compensated by special techniques. For example finding face border and particular face features can fully compensate shifting and scaling, and even rotation. However, tilt and turn are harder to compensate. Different light conditions can be partly compensated by special filters: normalizing brightness and/or contrast, equalizing and so on. Selection of algorithms applied depends on implementation. Detailed discussion of these methods is beyond the scope of this article [1]. We should note however, that in terms of optimization these methods clearly serve as clusterization, feature selection or both. For example, "finding face" procedure excludes features not related to object of recognition, and compensates differences between images, in which object is positioned differently. Hence the procedure serves as clusterization and feature selection. But the result of these procedures is still a raster image that is a bulk data and this image is still subject to further preprocessing.

4.1 Recognition in a single image space

Usually similar objects (for instance, faces) have many generic features and their images differ weakly if compared to differences among all possible objects. In this case set of objects' images I_F is a very narrow subset of I and it can be concluded that image space is far from optimal for describing objects in terms of their images. One can present input raster images (vectors of a huge number of dimensions) using vectors of lesser dimensionality. This is possible because: a) highly correlated features duplicate information, thus some of the features may be omitted; b) features that do not change significantly while shifting from one object to another also may be omitted as they don't yield substantial information. While dimensionality of image space reduces, image processing time and quantity of saved data decreases but informativity does not fall significantly.

PCA, based on Karhunen-Loeve expansion, is an approach to reducing the dimensionality of image space in such a way that a basis in a new space reflects properties of a variety of recognized classes in optimal way. It has following optimal properties: *a)* it minimizes error of approximation, thus working as optimal features selection (this property assures that the error of reconstruction by any fixed number of components is the smallest possible among any reconstructions made by the same number of components.) and *b)* it shows behavior typical to clusterization. Karhunen-Loeve expansion yields statistically uncorrelated components. These components are calculated as eigenvectors of autocorrelation matrix. That is why they are called further just as 'eigenvectors'. Eigenvectors corresponding to maximal dispersion of training image set are called *principal components* (PC). Choosing principal components for representation of faces provides first optimal property of PCA.

Let us illustrate how PCA works for object recognition. Optimal implementation of PCA presumes, that mean of all images is zero vector, hence mean vector is calculated for training set and then subtracted from all images treated in this set. *A priori* one knows only images, constituting training set, so it is reasonable hypothesis to suppose, that mean of all images is equal (or close enough) to mean of training set images (*Figure 6*). Principal components calculated in image spaces of photo-images and elevation maps are shown on *Figure 7* and *Figure 8* respectively. Every image (photo-image or elevation map) is represented in the basis of corresponding principal components. Vectors in this presentation are input data for decision rule. Performance of decision rule algorithm is very high, since dimensionality of vectors is very small by proposed procedure and distances in principal component spaces are calculated very fast.

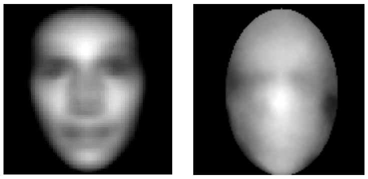

Figure 6. Mean vectors of training sets of photo-images and elevation maps respectively.

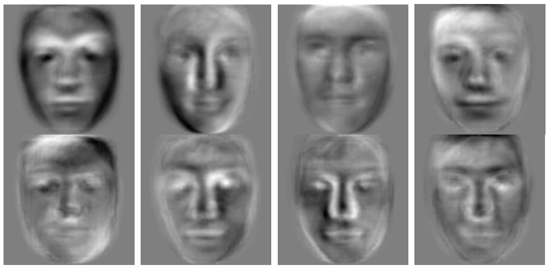

Figure 7. First eigenvectors of photo-images – principal components.

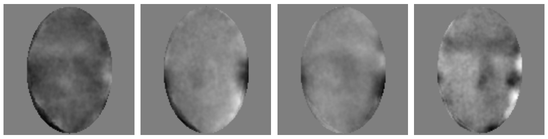

Figure 8. First eigenvectors of elevation maps – principal components

Every processed image is represented in the principal component basis. Principal component space I$_{PC}$ is a subspace of image space I. This implies that vector representing an arbitrary recognized image could be situated beyond the principal component space. Thus, vector reconstructed using Karhunen-Loeve expansion can differ from original vector. If this difference is too high one can make a decision that present image does not belong to the variety of recognized classes.

Following recognition scheme for one image space is proposed by Turk and Pentland [8]. There are four possible situations for every vector in an image space. They are presented in the following table and illustrated in *Figure 9*:

	PC Space I$_{PC}$	**Known images**	**Decision**
1	near	Near	Is a known object
2	near	Far	Is an unknown object
3	far	Near	Is not an object of given variety
4		Far	

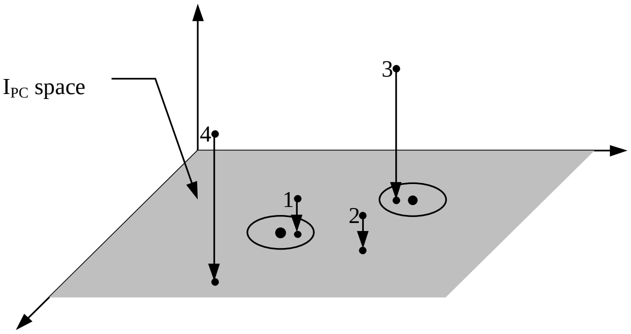

Figure 9. Possible situations for vector in image space.

4.2 Recognition in Multiple Image Spaces

Current version of system implements PCA in spaces of photo-images of objects and their relief maps. So, the implementation of proposed recognition scheme is not straightforward as it is in the case of one space.

Obviously, heterogeneous images (i.e. images belonging to different spaces) have different statistical characteristics. PCA distinguishes first eigenvectors as vectors corresponding to maximal dispersion of image set. If one will use vectors consisting of features from heterogeneous images, first principal components will reflect structure of a homogeneous subspace, which has maximal dispersion of features. Structure of other subspaces will be reflected by vectors of higher order and this structure will be hidden by noises of the former subspace. That’s why it is undesirable to use PCA on vectors consisting of heterogeneous features (for instance, vector that includes photo-image and relief simultaneously).

Thus, heterogeneous images are processed separately, i.e. for each type of them separate principal components space is built. Problem arises of synthesizing decision rule, which will consider information obtained in different image spaces. The following measure is proposed:

$$\frac{1}{R^2(c,v^k)}=\sum_i W_i\frac{1}{R_i^2(c_i,v_i^k)}\,, \qquad (1)$$

where R_i is a measure (Euclidean distance) in i-th image space, c is a compound image to be classified (image consisting of heterogeneous images is called here compound image), v^k is k-th compound image of training set, W_i are weights of heterogeneous spaces, calculated as:

$$W_i = D_i\frac{D_i}{D_i^c}\,, \qquad (2)$$

where D_i is dispersion of all images in i-th space and D_i^c is the averaged dispersion by classes in i-th space. $R(c,v^k)$ can be considered as a normalized distance for compound images. In terms of normalized distance decision rule may be defined as:

$$N=\arg\min\left\{R(c,v^1),\ldots,R(c,v^K),T\right\}. \qquad (3)$$

In case that T is not the minimum, N is a number of image, closest to classified image c. If distance exceeds threshold T, then it is concluded that image c does not belong to any of known classes. Otherwise classified image is considered to belong to the class that closest image belongs to. The level of the threshold depends on implementation. For example, the use of threshold is unnecessary if one knows a priori that image does belong to one of existing classes.

As one can see, a direct sum of measures (or squares of measures) in separate image spaces was not used. It is made by the following reason. Consider heterogeneous image of object, consisting of pair of photo-images. Assume first image of pair is in situation 1 of one-image-space recognition scheme (see *Figure 9*) and second is in situation 2. Then for direct sum most of it will fall to the share of second "bad" image. As for the reciprocal, used in measure (1), the main share will be of first "good" image and second image will not worsen this measure too much. This argument is based on the experimental knowledge that usually there are two different situations for images of the same class and for images of different classes. Images of the same class sometimes happen to be very close to each other. As for the images of different classes, they may be situated near each other, like most of the same class images, but not so close as the same class images happen to be. The measure built must not reduce the effect of this behavior and the measure (1) even strengthens this effect, as it increases the probability of considering of such "good" images.

Another improvement of measure (1), made recently, is that measures R_i are taken separately. Then formula looks like:

$$\frac{1}{R^2(c,k)} = \sum_i W_i \frac{1}{\min\limits_{j_k}\left[R_i^2(c_i, v_i^{j_k})\right]} . \qquad (4)$$

Expression in the denominator means that minimum of measure R_i is taken by all known images of given class in *i*-th image space. Thus, measure of image is taken relative to whole class of images, rather than to single compound image. This way the distance from class can be defined, rather than distance from image. As experiments show, implementation of this measure increases reliability of recognition.

The measure (4) was also implemented for continuous frame-grabbing, that gives a sequence of images (maybe, from a single camera) that changes in a given period of time, rather than a set of images from different points of view, but in a single moment in time. The measure is calculated for a set of images, taken at different moments. Weight (2) was revised to take into account different significance of moments of shooting simply by multiplication coefficient. For instance, images, shot just before processing are more significant than those shot some moments ago. The measures of "just-shot" images are added to the sum and decision rule is applied to corrected normalized distance. This allows proceeding with recognition in real-time mode using previously obtained information if first compound images do not yield reliable information about object of recognition.

5. CONCLUSION

Sets of images of about 100 stereo-images (that is 200 photo-images and 100 elevation maps) were used for determining optimal number of dimensions of Principal Component spaces in image spaces of photo-images and relieves. Number of Principal Components required for face recognition was determined to be 20-30 vectors in every image space. Thus, number of features was reduced in hundreds of times relative to raw input data quantity of about 10^4-10^5 values for each image. This allows classifying images on large databases in a real-time mode and storing processed images in a very compact form.

Experiments show that image preprocessing is of crucial importance for reliability of recognition. Therefore one should thoroughly choose and implement preprocessing methods before using our algorithm. Different methods supposed to have clusterization and feature selection characteristics been applied. They are: *a)* finding face border and orientation on photo-image with following procedures of clipping, scaling, shifting and rotating of original image, *b)* normalizing of brightness and contrast. The possibility of enhancing the face location procedure by adding algorithms of finding important face features as eyes, nose, mouth is studied.

Compound images consisting of pair of photo-images and elevation map, reconstructed from this pair are used for recognition in the system discussed. Also sets of images are used, consisting of frames, shot at different moments of time. It is possible to extend method by adding other types of images to the structure of compound image. It can be images of the most informative parts of human faces, particularly eyes and separately nose and mouth or face sketch which is a picture of edges obtained from original image by applying appropriate filters.

Recognition using compound images was tested on the database of about 600 stereo-images of 200 persons and recognition accuracy achieved was about 95%. This is about two times more reliable than using simple photo-images. Most wrong cases were due to difference in angle of view or facial expression.

6. REFERENCES

[1] **Alexander B. Murynin, Ivan A. Matveev, Victor D. Kuznetsov**. "*Automatic Stereoscopic System for Person Recognition*". SPIE , 1999, Vol. 3516.

[2] **Ivan A. Matveev, Alexander B. Murynin**,. *3-D Surface "Reconstruction in Automatic Recognition System"*. SPIE, 1999, Vol. 3516.

[3] **Victor D. Kuznetsov, Ivan A. Matveev, Alexander B. Murynin**, "*Optimization of Informative Components for 3-D Object Recognition*". SPIE , 1999, Vol. 3516.

[4] **V.D.Kuznetsov, I.A.Matveev** "*Face standard and region of interest selection in the problem of person identification*", Preprint of CCRAS, 1997

[5] **O.Nakamura, S.Mathur, T.Minami,** *"Identification of human faces on isodensity maps"*, Pattern Recognition,1991, V.24,N.3.

[6] **P.Fua,** "*Parallel stereo algorithm that produces dense depth maps and preserves image features*" Machine Vision and Applications, 1993, V.6.

[7] **B.Moghaddam, A.Pentland,** *"Face Recognition using View-Based and Modular Eigenspaces"* SPIE, 1994, V. 2277, July

[8] **M. Turk and A. Pentland**, "*Eigenfaces for Recognition*", Journal of Cognitive Neuroscience, Vol. 3, No.1, 71-86, 1991

Authors:

Alexander B. Murynin, Ph.D in Computer Science, Senior Researcher of Computing Center of RAS.
Address: Vavilova 40, Moscow, 119967, Russia
E-mail: murynin@ccas.ru

Ivan A. Matveev, the postgraduate student of Moscow Institute of Physics and Technology.
Address: Kerchenskaya 1A1, Moscow, 113303, Russia
E-mail: matveev@ccas.ru