Video-based age and gender recognition in mobile applications

A S Kharchevnikova[1] and A V Savchenko[1]

[1]National Research University Higher School of Economics, Myasnitskaya str. 20, Moscow, Russia, 101000

Abstract. In this paper we develop the age and gender recognition mobile system using deep convolutional neural networks for mobile applications. The brief literature survey on the age/gender problem in retail applications is presented. The comparative analysis of classifier fusion algorithms to aggregate decisions for individual frames is provided. In order to improve the age and gender identification accuracy we implement the video-based recognition system with several aggregation methods. We provide the experimental comparison for IJB-A, Indian Movies, Kinect and EmotiW2018 datasets. It is demonstrated that the most accurate decisions are obtained using the geometric mean and mathematical expectation of the outputs at softmax layers of the convolutional neural networks for gender recognition and age prediction, respectively. As a result, the off-line application of the proposed system is implemented on the Android platform.

1. Introduction

Due to the rapid growth of interest in video processing, the modern face analysis technologies are oriented to identify various properties of an observed person. In particular, age and gender characteristics can be applied in retail for contextual advertising for particular group of customers [1]. Often, people ignore the advertisements because the information is irrelevant, uninteresting for them at the current moment. Consequently, companies incur huge losses from investing in contextual advertising, which turns out to be inefficient and ineffective. Therefore, one of the key tasks of video analytics in retail is to provide relevant information that meets the interests of a specific target audience. For instance, depending on the automatically detected customer data, the application could provide relevant information that corresponds to a specific target audience. Consequently, the video-based age and gender recognition would improve the efficiency of the contextual advertising and increase sales. The necessary video frames for the following recognition can be obtained from digital screens or interactive panels in the shops. Such applications, running in real time, should perform the recognition task at the required speed on platforms that are limited by power and memory resources. Therefore the advanced decision for mobile platforms is required. Despite the fact that over the past few years a large number of different algorithms for age and gender recognition have appeared [2,3], the reliability of existing solutions remains insufficient for practical application [4].

Unlike traditional single-image processing systems, the video analysis lets us use addition information. For rather fast recognition algorithms one can obtain more than 100 frames of the classified object in the dynamics within a few seconds from the video stream [1,5]. It is sufficient to guarantee that at least several frames belong to the same class from the reference base. The intuition is that if each classifier makes different errors, then the total errors can be reduced by an appropriate combination of these classifiers. Thus, this research work is intended to consider the video-based age and gender recognition task as the problem of choosing the most reliable solution using the classifier fusion (or ensemble) methods [6,7,8]. After a set of solutions for each classifier is obtained, it is necessary to implement a combining function to make a single decision. The most obvious strategy is a simple vote, in which the decision is made in favor of the class with the maximum number of predictions. This paper compares the classifier fusion obtained by traditional averaging of individual decision rules [9] with solutions based on the principle of maximum a posteriori probability [10,11,12].

The rest of the paper is organized as follows. In Section 2 the brief literature survey on age and gender image recognition is presented. In Section 3 we provide classifier fusion solutions and describe the proposed recognition scheme. Experimental results and concluding comments are presented in Section 4 and Section 5 respectively.

2. Literature Survey

The task of classifying a video image of a person is as follows. Initially, each frame $\{X(t)\}, X(t)=\|x_{uv}(t)\|, t=\overline{1,T}$ is assigned to one of the L classes by feeding the RGB matrix of pixels of a facial image $X(t)$ to the CNN [4,13]. This deep neural network should be preliminarily trained using the very large dataset of facial images with known age and gender labels. For simplicity, we assume that the video contains only one classified person with a previously selected face area on the frame, so on each image $\{X(t)\}$ the face area is detected and left. Based on this, the task of recognizing gender is a typical example of a binary classification [10]. Despite the age prediction is an example of a regression problem, in practice the highest accuracy is achieved when it is assigned to the classification problem with the definition of several age categories (L = 8 in [14]).

The challenge of automatically extracting age and gender related attributes from facial images has received increasing attention in recent few years and the huge number of recognition algorithms has been proposed. Early age estimation methods are based on the calculation of the relationships between different dimensions of facial features. A detailed survey of such algorithms is presented by Kwon [15]. Since this solution requires an accurate calculation of the facial features location, that is a fairly complex problem, they are unsuitable for raw images, video frames. Geng [16] proposes a method for automatic age recognition - AGing pattErn Subspace (AGES), the concept of which is creating an aging pattern. However, the requirements of front alignment of images impose significant restrictions on the set of input parameters. The frequency-based approach is also known among the age identification algorithms. For instance, a combination of biological features of the image is studied by Guo et al. [17] (BIF - Biologically Inspired Features).

Gender recognition for a facial image is a much more simple task, because it includes only L = 2 classes. Hence, traditionally binary classifiers can be applied. Among them, such methods as SVM [20], boosting-based algorithms, and neural networks are widely used.

Unfortunately, the accuracy of traditional methods of computer vision and pattern recognition does not meet the requirements of practical application. With regard to the effectiveness of the convolution neural networks (CNN) implementation, in particular, to classification challenges, Levi [14] provides new insights into the process of solving age and gender recognition problems by applying this method. After that, several other papers have proved the efficiency of deep CNNs in these tasks [21,24,25]. Specifically, deep VGG-16 [22], trained to recognize gender and age by image, is described in [21]. Hence, we will use this deep learning approach in order to recognize age and gender for video data.

3. Proposed Algorithm

The output of the CNN is usually obtained in the Softmax layer that provides the estimation of posterior probabilities $P(l|X(t))$ for the *t-th* frame belonging to the *l-th* class label from the reference base [27]:

$$P(l|X(t)) = \text{softmax}\, z_l(t) = \frac{\exp(z_l(t))}{\sum_{j=1}^{L} \exp(z_j(t))}, l = 1,2,...,L \quad (1)$$

where $z_l(t)$ is the output of the l-th neuron in the last (usually fully connected) layer of the neural network. The decision is made in favor of a class with a maximum a posteriori probability (MAP) (1).

Due to the influence of diverse factors such as unknown illumination, quick change of camera angle, low resolution of video camera, etc., making a decision based on the MAP approach for every frame is usually inaccurate. Therefore, we will use the fusion of decisions for individual frames to increase recognition accuracy. The review and analysis of publications in the field of data processing shows that the synthesis of classifier fusion is one of the most effective approaches to increasing the accuracy and stability of classification [24,26,27]. According to aggregation algorithms, several criteria are used, each of which is able to assign a class label after that general classification result is formed on the basis of some principle [8]. In the task of video recognition, firstly the traditional problem of automatic image recognition with CNN is solved for each incoming $X(t)$ frame and then all individual solutions are combined into one common decision for a specific video recording. The most obvious approach is to use more complex algorithms for constructing classifier fusion based on algebraic methods [8, 26]. Most of these algorithms (such as weighted majority committee, bagging and boosting [10,28]) require a sufficient representative training sample. Unfortunately, in many image recognition cases, the existing database contains an insufficient number of standards for each class. In the present paper it is proposed to use known statistical methods of synthesis solutions [8] that do not require the test sample. So, we examine the following criteria [29, 11]:

1. *Simple voting*, in which each classifier votes on the class it predicts, and the class receiving the largest number of votes is the ensemble decision., in which the final decision is made in favor of the class [6, 11]:

$$l^* = \underset{l=\overline{1,L}}{\text{argmax}} \sum_{t=1}^{T} \delta(l^*(t) - l) \quad (2)$$

2. *Arithmetical mean* of posterior probability estimates (1), i.e. the sum rule [6]:

$$l^* = \underset{l=\overline{1,L}}{\text{argmax}} \frac{1}{T} \sum_{t=1}^{T} P(l|X(t)) \quad (3)$$

3. If we follow the "naive" assumption about the independence of all frames [29], then the decision should be taken according to the *geometric mean* of posterior probabilities, or the product rule [6]:

$$l^* = \underset{l=\overline{1,L}}{\text{argmax}} \prod_{t=1}^{T} P(l|X(t)) = \underset{l=\overline{1,L}}{\text{argmax}} \sum_{t=1}^{T} \log P(l|X(t)) \quad (4)$$

In addition, we recall that the age prediction task is an essential regression problem. Hence, in this case it is possible to compute an *expected value* (mathematical expectation):

$$l^* = \sum_{l=1}^{L} P(l|X(t)) \cdot l \quad (5)$$

The general data flow in the proposed video-based age and gender recognition system is presented in Fig. 1.

The first step implies supplying images from video camera to the input of the system. Isolated frames are selected from the video stream with a fixed frequency (about 10-20 times per second) in the frame selection block. Then it is important to fix and leave only the face area that is performed in the corresponding block. Face detection is conducted using the cascade method of Viola-Jones and the Haar features from the OpenCV library [31]. To speed up the work, known procedures for tracking a person identified in previous frames can be used [30,35].

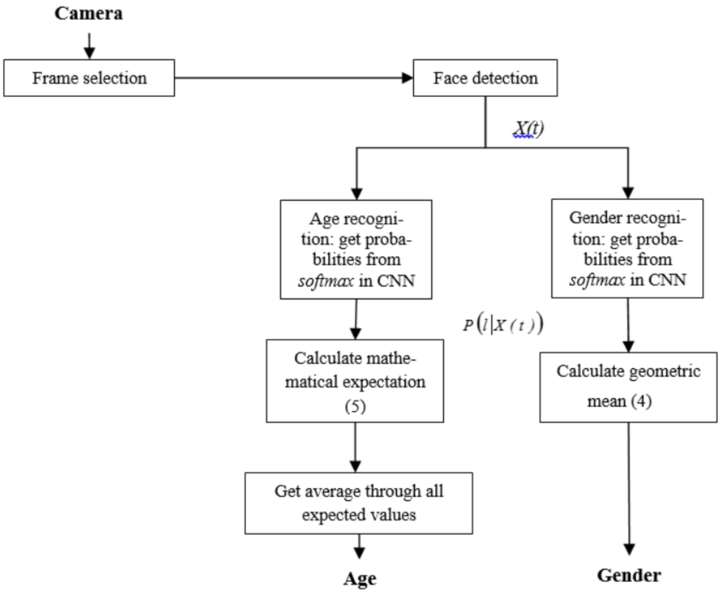

Camera
Frame selection
Face detection
X(t)
Age recognition: get probabilities from softmax in CNN
Gender recognition: get probabilities from softmax in CNN
$P(l|X(t))$
Calculate mathematical expectation (5)
Calculate geometric mean (4)
Get average through all expected values
Age
Gender

Figure 1. Proposed data flow.

At the next stage all the received images of persons on one frame are reduced to a single scale. In addition, often the subtraction of the mean image (*Image mean subtraction*) from each face image is applied [14]. The next step is supposed to provide the recognition of each frame, where the CNN is used. As the result, the estimates of posterior probabilities are obtained from the *softmax* layer (1). Recognition is performed using the model the Tensorflow library functionality. Based on the data that is the output of the classifier fusion block (2)-(4), a final recognition solution is implemented in favor of the corresponding class.

4. Experimental results

The experimental study of the proposed age/gender recognition algorithm scheme (Fig. 1) implementation is carried out in IDE Pycharm using “Python 3.6”. The characteristics of the machine: Intel Core i5-2400 CPU, 64-bit operating system Windows 7, with video card NVIDIA GeForce GT 440. The described approach taking into account the experimental results is also implemented on Android platform. The user interface of this application is presented in (Fig. 2).

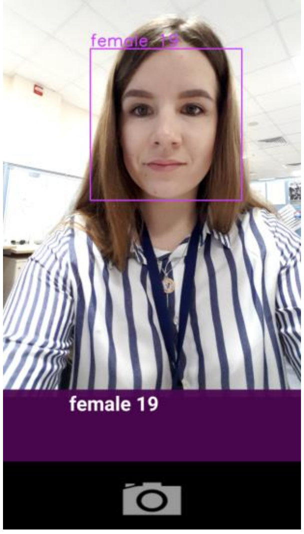
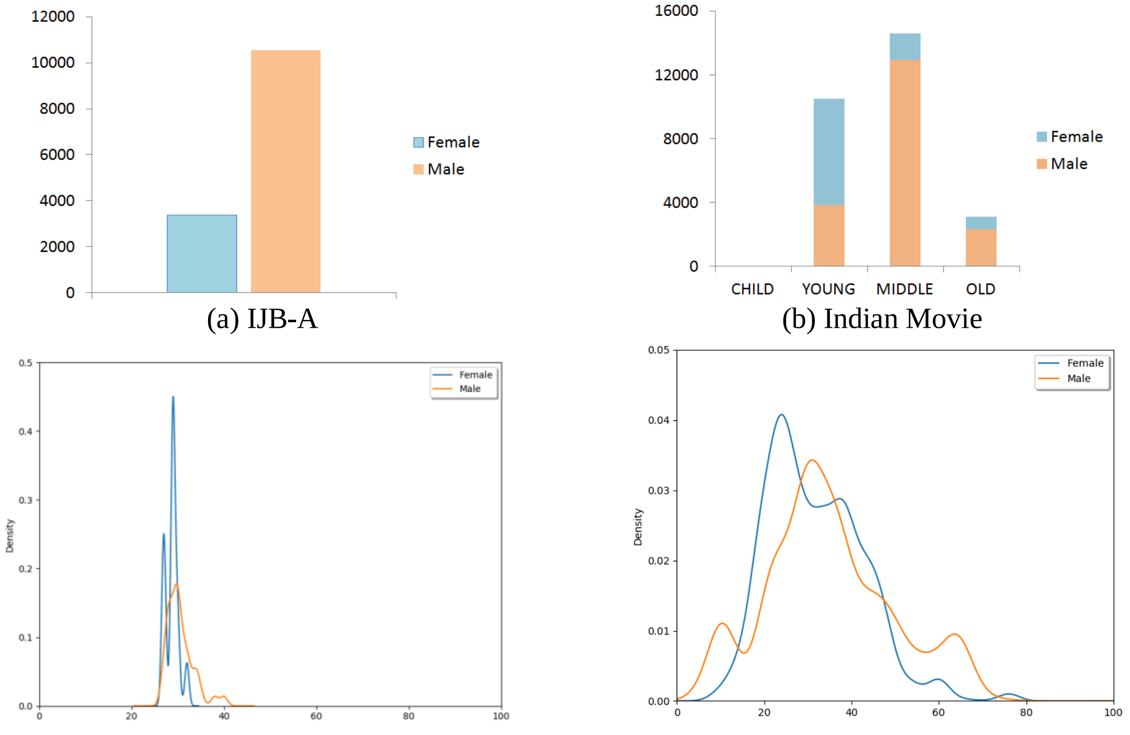


Figure 2. GUI of the proposed system.

According to the provided GUI the main part of the screen is designed to display video. There also is a text field to demonstrate the recognition and aggregation solution results in real time mode. By default it is defined as "No face detected", this indicates that the face has not been detected. Otherwise, the gender and age of the person on the frame are displayed.

The choice of datasets for recognition accuracy and performance testing is an inherently challenging problem. The reason for this is that just the limited number of databases provides such personal information as age, gender or both about a person on an image. Furthermore the video-based approach is considered in this work. In this case the databases that are used to train and test described CNNs architectures could not be applied. Hence, in this research testing the accuracy of recognition is conducted using the facial datasets IARPA Janus Benchmark A (IJB-A) [31], Indian Movie [32], EURECOM Kinect [33] and EmotiW2018[30], which gender and age information is available for, and the video frames of a single track are stored. The first dataset consists of 2,043 videos, where only the gender information is available. The data distribution for this dataset is presented in Fig 3a.

The next database is a collection of video frames assembled from Indian films. In total there are about 332 different videos and 34,512 frames of one hundred Indian actors, whose age is divided into four categories:”Child”, ”Young”, ”Middle” and ”Old”. In this example, the verbal description of age is replaced by provided specific age intervals: 1-12, 13-30, 31-50, 50+ respectively. In the following experiments the intersections of the recognition results at the given intervals will be estimated. The data distribution for the dataset can be found in Fig. 3b. The Kinect dataset contains 104 videos with 52 people (14 women and 38 men). The database provides information about the gender and the year of birth that simplifies the estimation of age (Fig. 3c). The EmotiW2018 database is a collection of videos taken from various films and serials. When implementing the algorithm, age is considered in the range with addition and subtraction of 5 years, since it is necessary to identify the accuracy of the intersection with the recognized age interval. The gender of the actor on the video is provided, as well as his age. In total, the database consists of 1,165 videos. Since this dataset does not contain the final video images but video files, it became necessary to split the video into frames and subsequently detect the face area. Information about data in EmotiW2018 is presented in Fig. 3d.



(a) IJB-A

(b) Indian Movie

(c) Kinect

(d) Emotiw2018

Figure 3. Data distribution in experimental datasets.

We compare two publicly available CNNs architectures: Age net and Gender net models [14] and deep VGG-16 [22] neural network trained for age/gender prediction [21]. Moreover, we implement image normalization techniques, namely, mean image subtraction to cope with illumination effects, low camera resolution, etc.

First of all, average inference time of individual CNNs for testing machine on Android platform is presented in Table 1. The best results are shown in bold.

Table 1. Average inference time and models' size.

	Time (sec)	Size (MB)
Gender/Age net	**1.076**	**43.55**
VGG-16	273	513.75

As expected, the most resource-intensive model is deep vgg16, which occupies almost 10 times more memory space than the Gender/Age nets. This fact imposes significant restrictions on the storage of such architecture, as well as on the recognition speed. It took about four minutes to make a decision for one frame using the vgg16 model.

Evaluation of the CNN quality and aggregation algorithms is carried out with the help of accuracy metric, since the problem of gender determination is a classic example of binary classification, and age recognition is considered to the investigation of several intervals-the multiclass classification. Accuracy is the proportion of correct algorithm responses.

In this paper, all proposed classifier solution methods are compared with the traditional recognition solution for each frame (2). Next, to visualize the results, we introduce the following abbreviations for discussed aggregation techniques:

FBF – *frame by frame* (2).
SV – *simple voting* (3).
SR – *sum rule* (4).
PR – *product rule* (5).
ME – *mathematical expectation* (6).

The comparison of CNN architectures and classifier fusion algorithms for gender task is presented in Table 2. The best results are in bold**.**

Table 2. Gender recognition accuracy (%).

	FBF	**SV**	**SR**	**PR**
	IJB-a			
Gender_net	51	**60**	**59**	**59**
VGG-16	72	81	81	**82**
	Kinect			
Gender_net	55	73	75	**77**
VGG-16	69	**84**	**84**	**84**
	Indian Movie			
Gender_net	61	71	72	**75**
VGG-16	75	81	87	**88**
	EmotiW2018			
Gender_net	72	**75**	**75**	**75**
VGG-16	71	78	80	**81**

Thus, based on the conducted experiments, it can be concluded that the classifier fusion implementation increases the accuracy of the gender recognition in comparison with the traditional approach. The difference is 3-10%. The product rule shows its efficiency among the aggregation algorithms almost in all cases. The deep Vgg16 is more accurate than Gender_net. For instance, the difference between models is considered to be about 9% for Kinect dataset.

Age recognition results are provided in Table 3.

Table 3. Age recognition accuracy (%).

	FBF	SV	SR	PR	ME
			Kinect		
Age_net	52	41	43	45	69
VGG-16	58	60	66	**71**	**71**
			Indian Movie		
Age_net	56	68	45	48	32
VGG-16	38	29	29	29	**54**
			EmotiW2018		
Age_net	26	27	27	27	**30**
VGG-16	47	47	48	48	**52**

The estimation of the mathematical expectation (5) has shown the effectiveness in determining the age in most cases. Thus, it could be noticed that the VGG-16 architecture is ahead of Gender net and Age net models for the age accuracy. Here we have a general trade-off between performance and accuracy. The low accuracy of age recognition can be due to the complexity of the problem as a whole, since this biometric characteristic depends on many factors and cannot always be uniquely determined.

5. Conclusion

The video-based age and gender recognition algorithm with the implementation of the classifier committees is proposed in this work. The experimental results have demonstrated the increase of recognition accuracy of the proposed algorithm when compared to traditional simple voting decision. The method of finding the geometric mean (product rule) with normalization of the input video images is the most accurate in gender classification task. At the same time, the most accurate age prediction is achieved with the computation of the expected value. We have presented the results of comparing the following CNN architectures: Age net and Gender net [14] and VGG-16 [22] trained for age and gender prediction. Eventually, the accuracy of the VGG-16 architecture is about 10-20% higher for the gender recognition and age prediction than Age and Gender net models. However, the inference time of the VGG-16 is 4-9 times lower. A limiting factor of VGG-16 practical usage has been overcome with optimization techniques [1,26,33,34]. As a result, a prototype of the age and gender recognition system (Fig. 1) for retail needs has been implemented in the Android application (Fig. 2). The intuition is that this application can improve the efficiency of contextual advertising.

6. References

[1] Savchenko A 2016 *Search techniques in intelligent classification systems* (Springer International Publishing)

[2] Chao W, Liu J and Ding J 2013 Facial age estimation based on label-sensitive learning and age-oriented regression *Pattern Recognition* **46** 628-641

[3] Rybintsev A V, Konushin V S and Konushin A S 2015 Consecutive gender and age classification from facial images based on ranked local binary patterns *Computer Optics* **39(5)** 762-769 DOI: 10.18287/0134-2452-2015-39-5-762-769

[4] Wang H, Wang Y and Cao Y, 2009 Video-based face recognition: A survey *World Academy of Science, Engineering and Technology, International Journal of Computer, Electrical, Automation, Control and Information Engineering* **3(12)** 2809-2818

[5] Kalinovskii I A and Spitsyn V G 2016 Review and testing of frontal face detectors *Computer Optics* **40(1)** 99-111 DOI: 10.18287/2412-6179-2016-40-1-99-111

[6] Kittler J and Alkoot F 2003 Sum versus vote fusion in multiple classifier systems *IEEE Transactions on Pattern Analysis and Machine Intelligence* **25** 110-115

[7] Tresp V 2001 *Committee machines. Handbook for Neural Network Signal Processing* 135-151

[8] Rudakov K 1999 On methods of optimization and monotonic correction in the algebraic approach to the problem of recognition *RAS Papers* 314-317 (in Russia)

[9] Mazurov V 1990 *Method of committees in problems of optimization and classification* (Moscow: Science) p 248

[10] Theodoridis S, Koutroumbas C 2009 *Pattern Recognition* (Elsevier Inc.) p 840

[11] Savchenko A 2012 The choice of the parameters of the image recognition algorithm on the basis of the collective of decision rules and the principle of maximum a posteriori probability *Computer Optics* **36(1)** 117-124

[12] Savchenko A 2012 Adaptive Video Image Recognition System Using a Committee Machine, *Optical Memory and Neural Networks (Information Optics)* **21** 219-226

[13] Savchenko A V 2018 Trigonometric series in orthogonal expansions for density estimates of deep image features *Computer Optics* **42(1)** 149-158

[14] Levi G and Hassner T 2015 Age and gender classification using convolutional neural networks *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops* 34-42

[15] Kwon Y and da Vitoria Lobo N 1994 Age classification from facial images *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 762-767

[16] Geng X 2006 Learning from facial aging patterns for automatic age estimation *Proceedings of the 14th ACM International Conference on Multimedia* 307-316

[17] Guo G, Mu G and Fu Y 2009 Human age estimation using bio-inspired features *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 112-119

[18] Choi S 2011 Age estimation using a hierarchical classifier based on global and local facial features *Pattern Recognition* **44(6)** 1262-1281

[19] Makinen E, Raisamo R 2008 Evaluation of gender classification methods with automatically detected and aligned faces *IEEE Transactions on Pattern Analysis and Machines Intelligence* **30(3)** 541-547

[20] Shan C 2012 Learning local binary patterns for gender classification on real-world face images *Pattern Recognition Letters* **33(4)** 431-437

[21] Rothe R, Timofte R and Van L 2015 Deep expectation of apparent age from a single image *Proceedings of the IEEE International Conference on Computer Vision Workshops* 10-15

[22] Simonyan K and Zisserman A 2014 Very deep convolutional networks for large-scale image recognition *Preprint arXiv:1409.1556*

[23] Szegedy C 2015 Going deeper with convolutions *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 1-9

[24] Krizhevsky A, Sutskever I and Hinton G 2012 ImageNet classification with deep convolutional neural networks *Advances in neural information processing systems* 1097-1105

[25] Esmaeili M 2007 Creating of Multiple Classifier Systems by Fuzzy Decision Making in Human-Computer Interface Systems *Proceedings of the IEEE Conference on Fuzzy Systems* 1-7

[26] Savchenko A 2002 Adaptive Video Image Recognition System Using a Committee Machine *Optical Memory and Neural Networks (Information Optics)* **21(4)** 219-226

[27] Shan C 2010 Face recognition and retrieval in video *Video Search and Mining* 235-260

[28] Lienhart R and Maydt J 2002 An extended set of Haar-like features for rapid object detection *Proceedings of the IEEE Conference on Image Processing* **1**

[29] Dhall A, Joshi J, Sikka K, Goecke R and Sebe N 2015 The more the merrier: Analysing the affect of a group of people in images *Proceedings of the 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition*

[30] Klare B 2015 Pushing the frontiers of unconstrained face detection and recognition: IARPA Janus Benchmark A *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*

[31] *IMFDB dataset* (Access mode: http://cvit.iiit.ac.in/projects/IMFDB/)

[32] *Eurecom Kinect dataset* (Access mode: http://rgb-d.eurecom.fr/)

[33] Savchenko A 2017 Maximum-likelihood dissimilarities in image recognition with deep neural networks *Computer Optics* **41(3)** 422-430 DOI: 10.18287/2412-6179-2017-41-3-422-430

[34] Rassadin A and Savchenko A 2017 Compressing deep convolutional neural networks in visual emotion recognition *CEUR Workshop Proceedings* **1901** 207-213

[35] Nikitin M Y, Konushin V S and Konushin A S 2017 Neural network model for video-based face recognition with frames quality assessment *Computer Optics* **41(5)** 732-742 DOI: 10.18287/2412-6179-2017-41-5-732-742

Acknowledgements
The work was conducted at Laboratory of Algorithms and Technologies for Network Analysis, National Research University Higher School of Economics and supported by RSF (Russian Science Foundation) grant 14-41- 00039.