Expert Systems With Applications 108 (2018) 170–182

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Unconstrained face identification using maximum likelihood of distances between deep off-the-shelf features

Andrey V. Savchenko[a,*], Natalya S. Belova[b]

[a] National Research University Higher School of Economics, Laboratory of Algorithms and Technologies for Network Analysis, 36 Rodionova St., Nizhny Novgorod, Russia

[b] National Research University Higher School of Economics, 20 Myasnitskaya St., Moscow, Russia

ARTICLE INFO

Article history:
Received 13 October 2017
Revised 2 April 2018
Accepted 29 April 2018
Available online 9 May 2018

Keywords:
Statistical pattern recognition
Unconstrained face recognition
Maximum likelihood estimation
CNN (Convolution neural network)
Kullback–Leibler divergence
Off-the-shelf deep features

ABSTRACT

The paper deals with unconstrained face recognition task for the small sample size problem based on computation of distances between high-dimensional off-the-shelf features extracted by deep convolution neural network. We present the novel statistical recognition method, which maximizes the likelihood (joint probabilistic density) of the distances to all reference images from the gallery set. This likelihood is estimated with the known asymptotically normal distribution of the Kullback–Leibler discrimination between nonnegative features. Our approach penalizes the individuals if their feature vectors do not behave like the features of observed image in the space of dissimilarities of the gallery images. We provide the experimental study with the LFW (Labeled Faces in the Wild), YTF (YouTube Faces) and IJB-A (IARPA Janus Benchmark A) datasets and the state-of-the-art deep learning-based feature extractors (VGG-Face, VGGFace2, ResFace-101, CenterFace and Light CNN). It is demonstrated, that the proposed approach can be applied with traditional distances in order to increase accuracy in 0.3–5.5% when compared to known methods, especially if the training and testing images are significantly different.

1. Introduction

Face identification is widely applied in various intelligent systems, such as video surveillance, border processing, virtual reality, search for a person in a social network, preventing voter fraud, conference socializing, driver license identification, law enforcement, etc. (Li, Wang, You, Li, & Li, 2013; Zhao, Liu, Liu, Zhong, & Hua, 2015). Though the face recognition has been thoroughly studied for several decades (Prince, 2012; Zhang, Yan, & Lades, 1997), it is still a challenging problem, which appears in many intelligent systems (Savchenko, 2016) due to the variability of individuals presented in images (Learned-Miller, Huang, RoyChowdhury, Li, & Hua, 2016). The latter task is usually solved with modern deep learning techniques (Goodfellow, Bengio, & Courville, 2016), which have recently reached a certain level of maturity. The most promising results are achieved with deep convolutional neural networks (CNN) (LeCun, Bengio, & Hinton, 2015). The CNNs have recently made it possible to achieve a near human-level performance in various computer vision contests including facial verification (Schroff, Kalenichenko, & Philbin, 2015; Taigman, Yang, Ranzato, & Wolf, 2014), especially when the training set contains plenty of reference images.

The situation becomes more complicated if the gallery set contains a small number of reference instances per each class. This small sample size problem (Chen, Liao, Ko, Lin, & Yu, 2000; Raudys, Jain et al., 1991) is especially crucial in face recognition, when sometimes only one reference image of each person is available (Tan, Chen, Zhou, & Zhang, 2006; Zhao et al., 2015). Such tasks with single reference photo per person of interest are typical for still-to-video face recognition (Dewan, Granger, Marcialis, Sabourin, & Roli, 2016; Parchami, Bashbaghi, & Granger, 2017; Savchenko, Belova, & Savchenko, 2018). In such case the transfer learning or domain adaptation methods can be applied (Cao, Wipf, Wen, Duan, & Sun, 2013; Goodfellow et al., 2016). In these methods the CNN is

Abbreviations: AUC, Area Under Curve; CNN, Convolution Neural Network; HOG, Histogram of oriented gradients; IJB-A, IARPA Janus Benchmark A; KL, Kullback-Leibler divergence; LBP, Local Binary Patterns; LDA, Linear Discriminant Analysis; LFW, Labeled Faces in the Wild; MAP, Maximum A Posteriori; MLP, Multi-Layered Perceptron; NN, Nearest Neighbor; PCA, Principal Component Analysis; SVM, Support Vector Machine; YTF, YouTube Faces.

* Corresponding author.

E-mail addresses: avsavchenko@hse.ru (A.V. Savchenko), nbelova@hse.ru (N.S. Belova).

URL: http://www.hse.ru/en/staff/avsavchenko (A.V. Savchenko), http://www.hse.ru/en/staff/belova (N.S. Belova)

https://doi.org/10.1016/j.eswa.2018.04.039
0957-4174/

used to extract facial features, which can be processed using the known classifiers. It is known that most of existing face recognition algorithms suitable for application with small training samples are typically based on the similarity comparison between the image features in the training set and an observed image (Guo & Zhang, 2017; Taigman et al., 2014), which is essentially the nearest neighbor (NN) method.

In order to increase the accuracy of the NN classifier, in this paper we propose to exploit the idea of several approximate NN methods (Micó, Oncina, & Vidal, 1994; Savchenko, 2012): if there exists a *reliable* decision $\mathbf{x}^*$, for which the distance to the input feature vector $\mathbf{x}$ is low ($\rho(\mathbf{x}, \mathbf{x}^*) \ll 1$), then $\rho(\mathbf{x}, \mathbf{x}_r) \approx \rho(\mathbf{x}^*, \mathbf{x}_r)$ with high probability for an arbitrary rth reference point. This assumption is known to be asymptotically correct for the KL divergence and probabilistic model of each class (Kullback, 1997). Following this idea, we introduce the novel face recognition method, which is based on the probabilistic interpretation of recognition task (Savchenko, 2016). At first, several nearest reference images (instances from the gallery set) are obtained for an observed facial image. Next, computed distances to all instances are used to weight the recognition results based on the estimation of their reliability. The more is the likelihood of the computed vector of distances for particular individual, the more is the weight corresponding to this subject. The likelihoods (joint probabilistic densities) are computed using the idea of the maximum-likelihood approximate NN algorithm (Savchenko, 2017a; 2017c) by assuming that the Kullback-Leibler (KL) divergence (Kullback, 1997) is used to compare distances between deep high-dimensional off-the-shelf features. However, we demonstrate that our approach can be successfully applied with traditional Euclidean distance.

The rest of the paper is organized as follows. In Section 2 we briefly overview existing image recognition methods suitable for small training samples. In Section 3 we present a simple statistical formulation of the face recognition task (Shakhnarovich, Fisher, & Darrell, 2002) using the KL minimum discrimination principle and introduce the novel approach, in which the maximum a-posteriori (MAP) rule is regularized using the computation of the joint probability densities of distances based on the asymptotic properties of the KL divergence. Section 4 presents the experimental results in recognition of images from either IJB-A (IARPA Janus Benchmark A) (Klare et al., 2015) or YTF (YouTube Faces) datasets (Wolf, Hassner, & Maoz, 2011) with the still images from the LFW (Labeled Faces in the Wild) dataset (Learned-Miller et al., 2016). Finally, concluding comments are given in Section 5.

2. Literature survey

One possible research direction in face recognition with small training samples is the usage of traditional computer vision methods (Szeliski, 2010), i.e., classification of either local features, e.g., SIFT, or a global descriptor, e.g., HOG (histogram of oriented gradients) (Dalal & Triggs, 2005). These methods usually partition each face image into several patches/blocks, and then perform feature extraction on them. Hence, they are sometimes called patch/block based approach (Zhu, Yang, Zhang, & Lee, 2014). After that, the NN methods are used with an appropriate similarity measure (Savchenko, 2016) between such features extracted from the observed image and all reference instances. A variant of such approach was proposed by Zhang, Yang, and Qian (2012), who performed PCA (principal component analysis) for feature extraction, chose k nearest neighbors of a given testing sample globally, and then used these neighbors to represent the testing sample via ridge regression. Similar ideas were introduced in the discriminative multi-manifold analysis method Lu, Tan, and Wang (2013), in which discriminative features are learned by maximizing the manifold margins of different persons. In this approach the facial images are segmented into disjoint sub-images, which form an image set for each sample per person. Another extension of these methods is the binary classification of the distances between images for intra/extra-personal classification (Zhang, Huang, Li, Wang, & Wu, 2004). It is suitable for the case when several reference facial images are available for each individual. Given the impossibility to train complex classifiers, the vector of distances between corresponding parts is assigned to one of two classes, depending on whether these distances are calculated between objects of the same or different classes. An observed image is segmented into a grid of blocks, then the distance vector is estimated for each instance from the training set. This distance vector is classified by a trained AdaBoost classifier, and the decision is made in favor of the class corresponding to the model with the highest confidence. In fact, it is the same NN rule, but the similarity measure is the AdaBoost's confidence.

The second group of methods includes enlarging of the training set by modifying images from the gallery set. After that statistical subspace-based approach can be implemented (Prince, 2012). These methods were most widely studied in literature devoted to face recognition from a single image per person (Tan et al., 2006). For instance, Zhao et al. (2015) proposed to automatically detect important local feature points by template matching and use a statistical model to learn the discriminative feature in the hidden space for each individual. Adaptive appearance model in still-to-video face recognition was considered by Dewan et al. (2016). Another well known technique is the singular value decomposition perturbations (Zhang, Chen, & Zhou, 2005) for the linear discriminant analysis (LDA), which enriches the eigenspace learned by the single training image. A linear generative model that creates a one-to-many mapping from an idealized identity space to the observed data space was introduced in Prince, Elder, Warrell, and Felisberti (2008) in order to deal with large pose variations.

One variant of such approach expands the training set using synthetic face generation (Mokhayeri, Granger, & Bilodeau, 2015), in which multiple virtual face images are generated from each single reference (Dewan et al., 2016). Image synthesis aims to estimate the intra-class face variations by simulating extra samples for each subject and, hence, increasing the number of samples. For instance, Li et al. (2013) proposed to enlarge the training set based on inter-class relationship and extended LDA in order to extract features from the enlarged training set. Zeng et al. (2017) used similar ideas to introduce intra-class facial variations, which are assumed to be shared across different persons.

Another variant is the multiple face representations (Bashbaghi, Granger, Sabourin, & Bilodeau, 2014), in which different discriminant features are extracted from a reference image to enhance the face models (Dewan et al., 2016). Leonidou, Tsapatsoulis, and Kollias (1999) proposed to use multiple representations of the gallery images, including variation in scaling, content and luminance. Multiple representations in Zhang, Hu, Xiang, and Zhao (2017) were created by cropping the original image into a number of non-overlapping blocks, applying of certain operations to each block, and merging all the modified blocks.

Unfortunately, most of these methods were successfully applied only in *constrained* face recognition task (Phillips et al., 2003). However, modern intelligent systems require identification of faces observed in *unconstrained* conditions, i.e., various illumination, pose, presence of noise, etc. (Best-Rowden, Han, Otto, Klare, & Jain, 2014; Learned-Miller et al., 2016). Hence, nowadays enlarging the training set (Dewan et al., 2016) (or generic learning) becomes all the more popular. In these methods the face intra-class variation information is gathered from an external generic training set.

An interesting combination of such approach with the patch/block methods is described by Zhu et al. (2014). They proposed the local generic representations of facial images by extracting the neighboring patches from the gallery dataset and considering the different importance of different facial parts.

However, a generic training set of auxiliary images is much more often gathered for domain adaptation methods (Goodfellow et al., 2016), when the large external datasets of celebrities are used to train the deep CNN (Parkhi, Vedaldi, & Zisserman, 2015; Wen, Zhang, Li, & Qiao, 2016; Wu, He, Sun, & Tan, 2015). Nowadays there exist several attempts to introduce CNNs into described above recognition methods for small training samples. For example, multi-pose deep representations were studied in AbdAlmageed et al. (2016) by training several pose-specific CNNs to generate multiple pose-specific features. Synthetic images with varying poses were generated in Hong, Im, Ryu, and Yang (2017) using a 3D face model in order to train the single sample per person domain adaptation network. In paper Masi, Tran, Hassner, Leksut, and Medioni (2016b) an auxiliary medium-sized dataset was enriched with important facial appearance variations (pose, shape and expressions) by manipulating the faces it contains in order to train ResNet-101 deep network. Parchami et al. (2017) fine-tuned the CNN for still-to-video recognition task using synthetically-generated faces based on still and videos of non-target individuals. Ding and Tao (2018) encouraged the CNN to learn blur-insensitive features automatically by using training data composed of both still images and artificially blurred data. Mokhayeri et al. (2015) generated multiple synthetic face images per reference based on camera-specific capture conditions to deal with illumination variations.

Despite the significant number of such studies, in many practical cases the trained CNN is simply applied to extract off-the-shelf features of the training images from the limited sample of subjects of interest using the outputs of one of the last layers (Savchenko, 2017b). Such deep learning-based feature extractors (Sharif Razavian, Azizpour, Sullivan, & Carlsson, 2014) make it possible to create a classifier that ideally performs nearly as well as if rich dataset of photos of these individuals were present (Cao et al., 2013). For example, the DeepFace method solves the face verification task by training a CNN on an external face dataset (Taigman et al., 2014). The outputs of the last layer of the CNN for the query image and every model from the gallery set are matched with the conventional chi-squared or Euclidean distance. Similar idea is implemented in the DeepID2 method (Sun, Chen, Wang, & Tang, 2014) which learns features suitable both for verification (Taigman et al., 2014; Zhou, Cao, & Yin, 2015) and identification (Liu, Deng, Bai, Wei, & Huang, 2015) tasks.

In fact, practically all existing unconstrained face identification algorithms suitable for application with very small training samples typically include the similarity comparison between deep CNN features in the training set and an observed image (Guo & Zhang, 2017), which is essentially the NN method. The only one exception is the usage of the multinomial logistic regression, which estimates the personsâ identity from his/her face feature (Guo & Zhang, 2017). In this paper an additional regularization (underrepresented-classes promotion) term was introduced in the objective function which aligns the norms of the weight vectors of the one-shot classes to those of the normal classes with many examples. Though the results are very promising, this method can be applied only if a rather large number of normal classes is available in the gallery set. As this situation is not typical for practical face recognition systems, in this paper we focus on the improvements of accuracy of the NN methods for deep learning-based feature extractors.

3. Materials and methods

3.1. Statistical face recognition

The task of the closed-set unconstrained face identification is to assign an observed image to one of $C > 1$ classes of subjects (identities). The classes are specified by the gallery set of $R \geq C$ facial images. We consider the supervised learning case, when the class label (subject id) $c(r) \in \{1, \ldots, C\}$ of the rth photo is known. At first, facial regions are obtained in each image using any appropriate face detector, e.g., either traditional fast AdaBoost multi-view cascade classifier (Viola & Jones, 2001) and HOGs (Dalal & Triggs, 2005) or more accurate CNN-based methods (Chen, Hua, Wen, & Sun, 2016; Hu & Ramanan, 2017; Zhang, Zhang, Li, & Qiao, 2016). Next, all facial images are resized in order to have the same width W and height H. We assume that the training sample is rather small ($C \approx R$) to train complex classifier (e.g. deep neural network). Hence, the domain adaptation can be applied (Goodfellow et al., 2016): each image is described with the off-the-shelf feature vector using the deep CNN, which has been preliminarily trained for face identification from large dataset, e.g., CASIA-WebFace, VGG-Face or MS-Celeb-1M. The L_1-normalized outputs at the one of last layers of this CNN for the input image and each rth reference image are used as the D-dimensional feature vectors $\mathbf{x} = [x_1, \ldots, x_D]$ and $\mathbf{x}_r = [x_{r;1}, \ldots, x_{r;D}]$, respectively.

If the ReLU activation functions (Krizhevsky, Sutskever, & Hinton, 2012) are used in the CNN, extracted feature vectors are positive and sum to 1 due to the L_1-normalization. As a result, they can be treated as the discrete probability distributions (Dalal & Triggs, 2005). In this section we assume that the feature vectors of input facial image and each rth instance are the estimates of multinomial distribution of (hypothetical) random variables X and $X_r, r \in \{1, \ldots, R\}$, respectively. The face recognition task is reduced to a problem of testing of simple statistical hypothesis $W_c, c \in \{1, \ldots, C\}$ about distribution of X. The optimal MAP decision is obtained the class corresponded to the maximal posterior probability $P(W_c|\mathbf{x})$ (Bishop, 2006). In this paper we focus on the case of full prior uncertainty, hence, the prior probabilities of observing each subject here are equal. It is typical for the small sample size problem to assume that each instance in the training set represents the mode in the probability density of the whole class. Hence, unknown posterior probability can be estimated using the Bayes rule:

$$P(W_c|\mathbf{x}) = \frac{\max\limits_{r \in \{1,\ldots,R\}, c(r)=c} f_r(\mathbf{x})}{\sum_{i=1}^{C} \max\limits_{r \in \{1,\ldots,R\}, c(r)=i} f_r(\mathbf{x})}. \quad (1)$$

Here $f_r(\mathbf{x})$ is the unknown probability density function of X_r. Following the ideas of representation learning (Goodfellow et al., 2016), the outputs of the last layers of deep CNN, and, as a consequence, the individual features can be assumed to be independent random variables. Hence, the likelihood $f_r(\mathbf{x})$ can be estimated with the following expression:

$$f_r(\mathbf{x}) = \prod_{d=1}^{D} (x_{r;d})^{WHx_d}, \quad (2)$$

where the sample size used to estimate the distribution of random variable X is computed as the total number of pixels in the input image WH. By substituting (2) in (1) and dividing its numerator and denominator to $\prod_{d=1}^{D} (x_d)^{WHx_d}$, one can obtain the following estimate of the posterior probability:

$$P(W_c|\mathbf{x}) = \frac{\exp\left(-WH\rho_c(\mathbf{x})\right)}{\sum_{i=1}^{C} \exp\left(-WH\rho_c(\mathbf{x})\right)}, \quad (3)$$

where $\rho_c(\mathbf{x})$ is the KL divergence between input image and the cth identity. In case of small training sample this distance can

be defined using the idea of the single-linkage cluster analysis (Ross, 1969):

$$\rho_c(\mathbf{x}) = \min_{r\in\{1,\dots,R\},c(r)=c} \rho_{KL}(\mathbf{x},\mathbf{x}_r). \quad (4)$$

Here $\rho_{KL}(\mathbf{x}, \mathbf{x}_r)$ is the KL divergence between feature vectors $\mathbf{x}$ and $\mathbf{x}_r$ (Kullback, 1997). Thus, the MAP criterion for the face identification task (Shakhnarovich et al., 2002) in the case of full prior uncertainty is equivalent to the KL minimum information discrimination principle (Kullback, 1997):

$$\min_{c\in\{1,\dots,C\}} \rho_c(\mathbf{x}). \quad (5)$$

In fact, this criterion implements conventional NN rule, which is typical for the small sample size problem (Raudys et al., 1991).

3.2. Proposed approach

In this paper we examine the potential application of the idea of the maximal-likelihood approximate NN method (Savchenko, 2017b; 2017c) in unconstrained face identification task. This idea exploits the known property of the KL divergence between two densities, which can be considered as the information for discrimination in favor of the first density against the second one (Kullback, 1997). Hence, we propose to use slightly different recognition criterion, which looks for the maximum posterior probability $P(W_c|\rho_1(\mathbf{x}),\dots,\rho_C(\mathbf{x}))$ based on the joint distribution of the C-dimensional random vector of *distances* $[\rho_1(\mathbf{x}),\dots,\rho_C(\mathbf{x})]$. To compute this posterior probability with the Bayes rule (1) it is necessary to estimate conditional joint density of distances between reference images and the input image $f(\rho_1(\mathbf{x}),\dots,\rho_C(\mathbf{x})|W_c)$. By using a natural assumption about independence of all classes from the training set, we estimate this joint density of distances as follows:

$$f(\rho_1(\mathbf{x}),\dots,\rho_C(\mathbf{x})|W_c) = f(\rho_c(\mathbf{x})|W_c)\cdot\prod_{i=1,i\neq c}^{C} f(\rho_i(\mathbf{x})|W_c). \quad (6)$$

At first, let us consider each term in the second multiplier. We propose to estimate the conditional density function $f(\rho_i(\mathbf{x})|W_c)$ of the distance between the image from the cth class and the ith instance using the known asymptotic properties of the KL divergence. It is known (Kullback, 1997) that $2WH$-times KL divergence $\rho_{KL}(\mathbf{x}, \mathbf{x}_i)$ is asymptotically distributed as the non-central chi-squared with $D-1$ degrees of freedom and the non-centrality parameter $2WH\cdot\rho_{KL}(\mathbf{x}_r, \mathbf{x}_i)$, if the input object (with estimated probability distribution $\mathbf{x}$) has the same distribution as $\mathbf{x}_r$. This asymptotic distribution is also known to be achieved by such probabilistic dissimilarities as the chi-squared distance, the Jenson–Shannon divergence, etc. (Kullback, 1997; Savchenko, 2017b). As the number of features D is usually large, it is possible to approximate the non-central chi-squared distribution with the Gaussian distribution (Savchenko, 2017c). Hence, we use the asymptotically normal distribution of the distance between the facial image of the rth subject and the ith reference photo:

$$\mathcal{N}\left(\rho_{KL}(\mathbf{x}_r,\mathbf{x}_i)+\frac{D-1}{WH}, \frac{4WH\rho_{KL}(\mathbf{x}_r,\mathbf{x}_i)+(D-1)}{2(WH)^2}\right). \quad (7)$$

In order to use such probability estimation in (6), one should define the distance between images from c-th and i-th identities (classes). In this paper we simply compute an average distance between instances from these classes:

$$\rho_{c;i} = \frac{1}{R_c R_i}\sum_{r=1}^{R}\sum_{r_i=1}^{R}\delta(c-c(r))\delta(i-c(r_i))\rho(\mathbf{x}_r,\mathbf{x}_{r_i}), \quad (8)$$

where $\delta(c)$ is the discrete delta function (indicator) and $R_c = \sum_{r=1}^{R}\delta(c-c(r))$ is the total number of photos of the c-th subject. Thus, the conditional densities in the second multiplier in (6) are estimated as follows:

$$f(\rho_i(\mathbf{x})|W_c) = \exp(-WH\phi_{c;i}(\mathbf{x})), \quad (9)$$

where we denote $\phi_{c;\,i}(\mathbf{x})$ for

$$\phi_{c;i}(\mathbf{x}) = \frac{1}{2WH}\ln\left(4\rho_{c;i}+\frac{\pi+D-1}{WH}\right)+\frac{\left(\rho_i(\mathbf{x})-\rho_{c;i}-\frac{D-1}{WH}\right)^2}{4\rho_{c;i}+\frac{D-1}{WH}}. \quad (10)$$

As the number of outputs (features) in the last layer in modern DNNs is very high ($D\gg 1$) (Parkhi et al., 2015; Wu et al., 2015), it is possible to approximately rewrite Eq. (11) (Savchenko, 2017c):

$$\phi_{c;i}(\mathbf{x}) \approx \frac{(\rho_i(\mathbf{x})-\rho_{c;i}-\frac{D-1}{WH})^2}{4\rho_{c;i}+\frac{D-1}{WH}}. \quad (11)$$

The first multiplier in (6) cannot be computed similarly, because, $\rho_{KL}(\mathbf{x}_r:\mathbf{x}_r)=0$, and asymptotic distribution (7) does not hold in practice if $i=r$. However, in such a case this conditional density can be estimated using Eq. (2). Similarly to (4), (5) one can show that

$$f(\rho_c(\mathbf{x})|W_c)) \propto \exp\left(-WH\rho_c(\mathbf{x})\right). \quad (12)$$

By combining expressions (9), (11), (12), the unknown posterior probability is estimated using softmax operation (3) as follows

$$P(W_c|\rho_1(\mathbf{x}),\dots,\rho_C(\mathbf{x})) = \frac{\exp\left(-WH\left(\rho_c(\mathbf{x})+\sum_{i=1,i\neq c}^{C}\phi_{r;i}(\mathbf{x})\right)\right)}{\sum_{j=1}^{C}\exp\left(-WH\left(\rho_j(\mathbf{x})+\sum_{i=1,i\neq j}^{C}\phi_{j;i}(\mathbf{x})\right)\right)}. \quad (13)$$

It is possible to convert the MAP criterion into rather simple face recognition rule

$$\min_{c\in\{1,\dots,C\}}\left(\rho_c(\mathbf{x})+\sum_{i=1,i\neq c}^{C}\phi_{r;i}(\mathbf{x})\right), \quad (14)$$

which can be viewed as an extension of the objective function from conventional NN (5) using the new term $\sum_{i=1,i\neq c}^{C}\phi_{c;i}(\mathbf{x})$. This term aims to choose such individual r, distance to which from any ith individual is approximately equal to the distance between input image and the ith class (Savchenko, 2017a; 2017c): the closer are the distances $\rho_i(\mathbf{x})$ and $\rho_{r;\,i}$ and the higher is the distance between references $\mathbf{x}_i$ and $\mathbf{x}_r$, the lower is the $\phi_{r;\,i}(\mathbf{x})$ in (14). As a result, our approach penalizes the unreliable nearest neighbor instances (5), which feature vectors do not behave like the features of observed image in the space of dissimilarities of the gallery images.

It is necessary to emphasize, that the resulted criterion (14) contains only computations of the KL divergence between off-the-shelf features of the input image and every instances from the gallery set. Moreover, conclusion about asymptotical distribution (7) of the KL divergence is in agreement with the well-known assumption about Gaussian distribution of dissimilarity measures between high-dimensional feature vectors, which is supported by many experiments (Burghouts, Smeulders, & Geusebroek, 2008). Thus, it is possible to replace the KL divergence $\rho_{KL}(\mathbf{x}, \mathbf{x}_r)$ with an arbitrary dissimilarity measure $\rho(\mathbf{x}, \mathbf{x}_r)$ between deep features $\mathbf{x}$ and $\mathbf{x}_r$. In this case it is necessary to multiply our summand in (14) using the smoothing parameter λ/C in order to reflect different scale of various dissimilarity measures.

Unfortunately, the run-time complexity of the proposed criterion (14) $O(RD+C^2)$ is much higher than the complexity of the baseline NN rule (5). Hence, our approach cannot be applied in practical tasks when the number of classes is rather

Algorithm 1 Maximum likelihood of distances in unconstrained face recognition.

Require: observed image, gallery set of facial feature vectors $\{\mathbf{x}_r\}$
Ensure: subject id (class label) corresponded to the objective function (15)
1: Initialize arrays of distances $\rho[c] := 0, c \in \{1, \dots, C\}$
2: Detect facial region in the observed image, resize it and extract deep off-the-shelf features $\mathbf{x}$ from one of the last layers of the CNN
3: **for** each class label $c \in \{1, \dots, C\}$ **do**
4: Assign $\rho[c] = \rho_c(\mathbf{x})$ (4)
5: **end for**
6: Obtain labels $\{c_1, \dots, c_M\}$ using a selection algorithm for finding M smallest numbers in the array $\{\rho[c] | c \in \{1, \dots, C\}\}$
7: Initialize posterior probabilities $L[m] := 0, m \in \{1, \dots, M\}$
8: **for** each top class index $m \in \{1, \dots, M\}$ **do**
9: Assign $\Omega := 0$
10: **for** each class label $i \in \{1, \dots, (c_m - 1), (c_m + 1), \dots, C\}$ **do**
11: Assign $\Omega := \Omega + (\rho[i] - \rho_{c_m;i} - \frac{D-1}{WH})^2 / (\rho_{c_m;i} + \frac{D-1}{4WH})$
12: **end for**
13: Assign $L[m] := \exp(-WH(\rho[c_m] + \lambda\Omega/C))$
14: **end for**
15: **return** class label c_{m^*}, where $m^* = \arg\max_{m \in \{1,\dots,M\}} L[m]$

large (Savchenko, 2017c). To speed-up the decision process, we propose to modify criterion (14) by examining only $M \ll C$ candidate classes. Namely, the distances between an observed image and each class are computed identically to the conventional approach (5), and M candidate classes $\{c_1, \dots, c_M\}$ with the lowest distances are chosen. This Mth element search (selection algorithm) is known to have linear average complexity $O(C)$. In the final decision only these M candidates are checked, so the criterion (14) is modified as follows:

$$\min_{c \in \{c_1, \dots, c_M\}} \left(\rho_c(\mathbf{x}) + \frac{\lambda}{C} \sum_{i=1, i \neq c}^{C} \frac{(\rho_i(\mathbf{x}) - \rho_{c;i} - \frac{D-1}{WH})^2}{\rho_{c;i} + \frac{D-1}{4WH}} \right). \quad (15)$$

Our complete procedure is presented in Algorithm 1. Its runtime complexity is equal to $O(RD + MC)$. This algorithm can be used to increase the accuracy of the NN rule if the training sample is rather small. The next section experimentally supports this fact.

4. Experimental results

In this section we consider one of acute face recognition tasks, namely, image-set based face identification (Mian, Hu, Hartley, & Owens, 2013), in which a set or a sequence of observed images of the same individual are available for decision-making (Cevikalp & Triggs, 2010; Wolf & Shashua, 2003). This task appears in, e.g. still-to-video face recognition (Huang, Zhao, Shan, Wang, & Chen, 2013; Savchenko et al., 2018; Zhu et al., 2015). It is very difficult, because the testing data have generally low quality and are captured under poor illumination and pose (Huang et al., 2012). Most of the known algorithms for this task suffered from heavy off-line training load (Liu, Zhang, Liu, & Yan, 2014), and they were experimentally studied only with databases, in which the training set is gathered under controlled environment, i.e. reference photos are of high quality and resolution, in frontal view, with normal lighting and neutral expression (Huang et al., 2012; Zhu et al., 2015). However, in this paper we decided to consider the most challenging experimental setup, in which the training set contains photos gathered in uncontrolled environment (Learned-Miller et al., 2016).

4.1. Experimental setup

The proposed Algorithm 1 is implemented in the publicly available stand-alone C++ application[1] with Visual C++ 2013 Express Edition using OpenCV library. All experiments were carried out on a HP Pavilion 14-ba020ur laptop (Intel Core i5 7200U 2.5 GHz, 6Gb RAM). The Caffe framework is applied to extract deep off-the-shelf features using five publicly available pre-trained CNN models for unconstrained face recognition, namely, the VGG-Face (VGG-16 Network) (Parkhi et al., 2015), ResFace-101 (ResNet-101 for face recognition) (Masi et al., 2016b), CenterFace (discriminative features with center loss) (Wen et al., 2016), Light CNN (Wu et al., 2015) and two models trained on recently introduced VGGFace2 dataset (Cao, Shen, Xie, Parkhi, & Zisserman, 2017), namely, ResNet-50 and SENet (Hu, Shen, & Sun, 2017) (hereinafter "VGG2 (ResNet)" and "VGG2 (SENet)", respectively). We downloaded the already trained models from official web sites of these networks. The "fc8" layer of the VGG-Face extracts $D = 4096$ non-negative features from 224×224 RGB image. The same image format is used to extract $D = 2048$ features from "pool5" layer of ResFace-101 and "pool5/7x7_s1" layer of VGG2 (ResNet/SENet). The CenterFace descriptor contains $D = 512$ features, which can be obtained at the "fc5" output of the network feeding 96x112 RGB image. The Light CNN (version C) extracts $D = 256$ (possibly negative) features at "etlwise_fc2" layer from 128x128 grayscale image. The outputs of the latter two CNNs were L_2 normalized to form the final feature vectors, which are matched using L_2 (Euclidean) distance. The features in the VGG-Face, VGG2 (both ResNet and SENet) and ResFace are positive, hence we also perform L_1 normalization to treat them as the probability distributions and match them using the symmetrized KL divergence. All facial images are resized without maintaining the aspect ratio according to the parameters specified in the architecture details of a particular CNN.

In addition, we decided to evaluate the proposed approach by using different feature sets rather than the deep off-the-shelf features. We adapted the Matlab implementation (Ortiz & Becker, 2014) of traditional HOGs (Dalal & Triggs, 2005) and LBP (Local Binary Patterns) histograms (Ahonen, Hadid, & Pietikainen, 2006). In order to compute $D = 1800$ HOG features, facial images are divided into 10x10 regular grid, and the range of gradient orientation in each cell was split into 18 equal parts. Similarly, histogram of LBP with 59 bins were computed in each block (in total, 8x8 non-overlapped blocks) in order to obtain $D = 3776$ features. All the histograms were L_1 normalized in each block in order to treat the resulted descriptors as a sequence of probability densities. These parameters were chosen by using recommendations of Ortiz and Becker (2014) and allowed us obtaining the highest accuracy.

In order to make a final decision of an image-set identification for the whole set of testing photos, it is necessary to aggregate decisions for individual photos. In the proposed approach (Algorithm 1) we accumulate the posterior probabilities (13) for all M best classes $\{c_1, \dots, c_M\}$. Our algorithm is compared with conventional classifiers, namely, 1) the minimum average distance criterion similar to the k-NN rule (5) with $k = 1$ and $k = 3$ neighbours and parameter $C = 1$; 2) average scores estimated by support vector machine (SVM) from LIBSVM with radial basis function kernel; and 2) average outputs of softmax layer of multi-layered perceptron (MLP) from OpenCV with one hidden layer (128 neurons) and sigmoid activations trained using resilient propagation (RPROP) during 1000 epochs and learning rate 0.01. SVM and MLP showed the highest accuracy for VGG-Face, ResFace and CenterFace features if 256 main components are extracted with the PCA.

[1] https://github.com/HSE-asavchenko/HSE_FaceRec/tree/master/windows

The Light CNN features for these classifiers did not require PCA decomposition. The parameters of all classifiers were tuned using the LFW dataset (Learned-Miller et al., 2016).

4.2. Experiments on IJB-A

In the first experiment we examine IJB-A dataset (Klare et al., 2015). We followed the IJB-A 1:N closed-set protocol, which provides three sets of data for each of its 10 splits. All templates from the probe set of 112 or 113 persons (without 55 distractors) are recognized using the given gallery set, i.e. an external train set of other 333 persons was completely ignored.

As this paper is primarily focused on small training samples, we tried to implement some key approaches described in Section 2, namely, synthetic generation of facial images and combination of multiple representations of facial images. In the former case, we transformed pose of existing gallery images and enhanced their illumination with the source code from the book by Baggio et al. (2012). In the latter case, we simply combined the various feature sets (CNN outputs, HOGs and LBP histograms). Unfortunately, none of these techniques outperformed classification of the deep CNN features. The usage of several representations did not provide significant difference in accuracies with the top classifier for particular dataset with only one CNN in the feature extractor. Generation of synthetic images even *decreases* the accuracy in 1–2%. It seems that all above-mentioned CNNs were trained using datasets with rich variations, so that additional information in the given small gallery set cannot provide higher quality. At the same time, we observed an expected gain in accuracy for traditional HOG and LBP features. Hence, we decided to report results obtained for generation of synthetic images for these feature sets. It was observed that significantly better results can be achieved with only two synthetic images per each reference photo. These images are synthesized using: 1) mirroring, and 2) translation using eye detection, smoothed equalization of left and right parts, and denoising with the median filter with 3x3 kernel (Baggio et al., 2012).

It is known that the usage of face detection techniques and choice of different bounding box positioning for the best configuration (Ferrari, Lisanti, Berretti, & Del Bimbo, 2017) is a keystone in many face recognition methods. Hence, we implemented two approaches to extract facial regions. In the first case, the bounding boxes provided in the IJB-A metadata were used to crop out faces from the images. In the second case, we detected the faces using the Tensorflow's implementation of the MTCNN (Zhang et al., 2016) to keep the cropping consistent between training and evaluation (Cao et al., 2017). Such preprocessing is widely applied in other papers to achieve near state-of-the-art results. For example, Yang et al. (2017) detected the faces with landmarks using STN face detector (Chen, Hua et al., 2016), and then aligned the face image with similarity transformation. As most of photos in the IJB-A dataset contain several faces of different subjects, we choose the facial region, which intersects with the bounding box provided in the IJB-A metadata. Only 1% of faces have not been detected by the MTCNN. As all of them are non-frontal and highly occluded, we decided to ignore these samples in the gallery sets. Moreover, we experimentally noticed that the highest accuracy for the CNN-based features is achieved when the facial region detected by the MTCNN is made squared with the side equal to the maximum of height and width of this facial region.

We use the following parameters of our Algorithm 1: $\lambda = 8$ and $M = 64$. The average rank-1 accuracies for all methods described above are shown in Table 1. In addition, we report average time to recognize a single image in Table 2. Here we do not include performance of feature extraction (including inference in CNNs) in order to assess the computation complexities of compared classifiers.

As expected, the computation complexity linearly depends on the feature vector dimensionality D. The error rate of traditional features (LBP/HOG) is much higher when compared to modern CNN-based approach. Generating synthetic images makes it possible to increase accuracy in 2–4%.

Secondly, preliminary face detection with the MTCNN significantly outperforms recognition with provided bounding boxes. For example, accuracy for CenterFace, Light CNN and ResFace descriptors increased at approximately 13%, 20% and 8%, respectively. The error rate with face detection becomes 2–4% lower even for traditional LBP/HOG features. In our experiments we noticed that the high resolution of initial images fed to the CNN is extremely important for this dataset. This empirical fact explains insufficient quality of the CenterFace and Light CNN models. The ResFace model is not the best one because it has been pre-trained using synthetically generated images (Masi et al., 2016b) from the CASIA WebFace dataset, which contains 4–6-times less images than the VGG-Face and VGGFace2 datasets. As a result, the features from the VGG-Face (Parkhi et al., 2015) with provided bounding boxes allows classifying images 35%, 25%, and 10–15% more accurately when compared to the CenterFace, Light CNN and ResFace-101, respectively. However, the difference in error rates is not so obvious if facial regions are detected with the MTCNN. Nevertheless, we believe that the best facial descriptors should not drastically depend on the face detection procedure. Such requirement of recognition stability holds for the VGG-Face features and for the recently released VGG2 (ResNet/SENet) models (Cao et al., 2017), which outperform the accuracy of the VGG-Face in approximately 8% and makes it possible to achieve near state-of-the-art rank-1 error rates. Our experiments support the statement of Cao et al. (2017) that SENet has a consistently superior (0.1–0.7%) accuracy when compared to ResNet-50.

Thirdly, the KL divergence increases the accuracy in 1–2% when compared to the Euclidean distance in practically all cases including traditional HOG/LBP features. However, the usage of L_2 metric is 1.5–2.5-times faster (Table 2). Finally, though SVM is slightly more accurate than the NN rule (5), the proposed Algorithm 1 is the most accurate classifier for all distances and feature sets. Its error rate is 0.3–5.5% lower due to the usage of additional information about pair-wise distances between instances from the training set (15). Though the *absolute* difference in error rates is rather low (0.2–0.4%) for the most accurate VGG2 (ResNet/SENet) features, the *relative* improvements ($\frac{96.8-95.9}{100-95.9} \approx 22\%$ for the L_2 distance and IJB-A bounding boxes and $\frac{98.2-98.0}{100-98.0} \approx 10\%$ for the KL divergence and MTCNN detectors) remains noticeable. For instance, the proposed approach decreased error rates *relatively* to the NN rule in 11–12%, 9–16%, 8–11% and 5–10% for VGG-Face, ResFace, Light CNN and CenterFace descriptors, respectively. Moreover, very high (6-8%) absolute decrease of error rate for the proposed method and HOG features with Euclidean distance corresponds to only 8–10% relative improvements. Complexity of introduced regularization term (15) is much lower when compared to the complexity of the distance computations in an exhaustive search. As a result, our algorithm requires only 0.3–0.8 ms to improve the accuracy of the NN decision.

In order to compare our results with the state-of-the-art works for this dataset, Table 3 summarizes the rank-1 accuracies of 1:N identification for IJB-A dataset reported in existing literature.

In fact, our results for VGG2 ResNet/SENet descriptors (Table 1) look rather competitive when compared to the state-of-the-art methods. Our approach combined with the symmetric KL divergence is as accurate as the state-of-the-art L_2-regularized L_2-loss primal SVM with class weighted squared hinge loss objective (Crosswhite et al., 2017) for SENet features learned on the VGGFace2 dataset (Cao et al., 2017). However, in contrast to the latter

Table 1
Rank-1 accuracy (%) in 1:N identification for IJB-A dataset.

Features	Distance	IJB-A bounding boxes			MTCNN face detection		
		SVM	NN	Proposed	SVM	NN	Proposed
CenterFace	L_2	52.6 ± 1.8	49.5 ± 1.9	54.6 ± 1.7	64.6 ± 1.0	64.9 ± 1.1	66.8 ± 1.1
Light CNN	L_2	57.2 ± 0.5	58.3 ± 0.7	61.8 ± 0.6	79.9 ± 0.4	81.3 ± 0.4	83.3 ± 0.3
ResFace	L_2	79.8 ± 0.7	77.9 ± 0.5	80.7 ± 0.7	86.4 ± 0.5	85.6 ± 0.6	87.6 ± 0.6
ResFace	KL	–	79.2 ± 0.6	81.3 ± 0.7	–	86.0 ± 0.5	88.3 ± 0.6
VGG-Face	L_2	87.7 ± 0.6	87.0 ± 0.5	88.5 ± 0.6	89.2 ± 0.4	89.0 ± 0.4	90.3 ± 0.4
VGG-Face	KL	–	87.5 ± 0.6	88.9 ± 0.6	–	90.0 ± 0.3	91.5 ± 0.4
VGG2 (ResNet)	L_2	95.7 ± 0.3	95.7 ± 0.5	96.1 ± 0.4	96.8 ± 0.4	97.7 ± 0.4	98.0 ± 0.3
VGG2 (ResNet)	KL	–	96.4 ± 0.4	96.7 ± 0.3	–	97.5 ± 0.4	98.0 ± 0.5
VGG2 (SENet)	L_2	95.3 ± 0.3	95.9 ± 0.5	96.8 ± 0.5	97.3 ± 0.5	97.8 ± 0.5	98.1 ± 0.4
VGG2 (SENet)	KL	–	96.0 ± 0.5	96.7 ± 0.3	–	98.0 ± 0.4	98.2 ± 0.6
LBP histograms	L_2	25.8 ± 1.2	24.2 ± 1.1	26.2 ± 1.3	29.0 ± 1.0	28.1 ± 1.2	29.2 ± 1.2
LBP histograms (synthetic images)	L_2	28.2 ± 1.5	27.3 ± 1.2	28.9 ± 1.4	32.3 ± 1.6	31.0 ± 1.3	32.8 ± 1.4
LBP histograms	KL	–	31.0 ± 1.3	33.7 ± 1.5	–	33.6 ± 1.1	34.7 ± 1.4
LBP histograms (synthetic images)	KL	–	33.7 ± 1.4	35.2 ± 1.6	–	35.4 ± 1.2	36.3 ± 1.4
HOG	L_2	11.7 ± 3.4	19.7 ± 2.0	27.9 ± 2.1	12.1 ± 3.3	23.0 ± 1.8	29.1 ± 2.0
HOG (synthetic images)	L_2	12.3 ± 3.5	23.9 ± 3.1	31.9 ± 3.2	12.9 ± 3.1	26.6 ± 2.6	32.3 ± 2.8
HOG	KL	–	29.4 ± 1.9	32.3 ± 2.0	–	31.5 ± 1.6	33.3 ± 1.8
HOG (synthetic images)	KL	–	32.9 ± 1.4	34.4 ± 1.4	–	35.1 ± 1.6	37.2 ± 1.2

Table 2
Average time to classify an image (ms) in 1:N identification for IJB-A dataset.

Features	Distance	SVM	NN	Proposed
CenterFace	L_2	2.8 ± 0.1	2.7 ± 0.1	3.6 ± 0.2
Light CNN	L_2	1.1 ± 0.1	1.6 ± 0.1	1.9 ± 0.1
ResFace	L_2	6.1 ± 0.2	9.4 ± 0.1	9.9 ± 0.2
ResFace	KL	–	15.2 ± 0.2	16.2 ± 0.3
VGG-Face	L_2	12.7 ± 0.3	18.2 ± 0.2	18.8 ± 0.2
VGG-Face	KL	–	39.6 ± 0.3	40.5 ± 0.5
VGG2 (ResNet/SENet)	L_2	5.8 ± 0.1	9.4 ± 0.2	9.9 ± 0.3
VGG2 (ResNet/SENet)	KL	–	16.6 ± 0.3	17.6 ± 0.4
LBP histograms	L_2	11.5 ± 0.2	17.6 ± 0.1	18.4 ± 0.3
LBP histograms (synthetic images)	L_2	37.3 ± 0.3	52.0 ± 0.2	52.4 ± 0.2
LBP histograms	KL	–	51.9 ± 0.1	52.6 ± 0.3
LBP histograms (synthetic images)	KL	–	157.7 ± 0.3	159.9 ± 0.6
HOG	L_2	5.4 ± 0.1	9.8 ± 0.2	10.5 ± 0.2
HOG (synthetic images)	L_2	7.3 ± 0.2	20.2 ± 0.3	21.1 ± 0.3
HOG	KL	–	19.2 ± 0.2	19.8 ± 0.2
HOG (synthetic images)	KL	–	28.6 ± 0.2	29.4 ± 0.3

Table 3
Rank-1 accuracy (%) of the best known methods in 1:N identification for IJB-A dataset.

Classifier	Rank-1 accuracy
OpenBR (Klontz, Klare, Klum, Jain, & Burge, 2013)	24.6 ± 1.1
Fisher Vector (Simonyan, Parkhi, Vedaldi, & Zisserman, 2013)	38.1 ± 1.8
GOTS (Klare et al., 2015)	44.3 ± 2.1
Face-Search (Wang, Otto, & Jain, 2015)	82.0 ± 2.4
Pose-aware model (Masi, Rawls, Medioni, & Natarajan, 2016a)	84.0 ± 1.2
Deep Multi-Pose (AbdAlmageed et al., 2016)	84.6
Triplet Similarity (Sankaranarayanan, Alavi, Castillo, & Chellappa, 2016)	88.0 ± 1.5
Bilinear CNN + VGGNet (Chowdhury, Lin, Maji, & Learned-Miller, 2016)	89.5 ± 1.1
Domain adaptation (Sohn et al., 2017)	89.5 ± 0.3
DCNN fusion (Chen, Patel, & Chellappa, 2016)	90.3 ± 1.2
ResFace, Combined images (Masi et al., 2016b)	90.6
Template adaptation (Crosswhite et al., 2017)	92.8 ± 1.0
Triplet Embedding (Sankaranarayanan et al., 2016)	93.2 ± 0.1
Neural aggregation network (Yang et al., 2017)	95.8 ± 0.5
VGGFace2_ft (SENet) (Cao et al., 2017)	98.2 ± 0.4

paper, we did not used the media id provided for each image (one probe template may contain multiple media). The aggregation of the frames in each media with further aggregating the media features to generate the final template feature is crucial for contemporary classifiers like SVM, but is not important for the NN-based techniques. Thus, our approach is more appropriate for practical systems, in which the media id information may not be always available (Yang et al., 2017).

4.3. Practical example

Let us demonstrate how the approach increases the face recognition accuracy by using a simple example. We recognized one of the probe images (Fig. 1a) of subject with id 942 given the gallery images from split 2 of IJB-A dataset. The features were extracted with the VGG2 (ResNet) CNN and matched using the standard L_2 metric.

(a) (b) (c) (d)

Fig. 1. Images from face recognition example: (a) Probe (SubjectId 942), (b) first NN (SubjectId 1001), (c) second NN (SubjectId 1303), (d) Result of the proposed algorithm (seventh NN gallery image, SubjectId 942).

Table 4
Main results for recognition of probe image (Fig. 1a), correct subject id 942.

SubjectId	L_2 distance	Regularization term	Total objective (15)
1001	0.0236	0.0152	0.0247
1303	0.0238	0.0152	0.0249
314	0.0239	0.0113	0.0247
1980	0.0240	0.0166	0.0252
387	0.0241	0.0300	0.0262
1289	0.0241	0.0140	0.0251
942	**0.0242**	**0.0059**	**0.0246**
1287	0.0242	0.0174	0.0258

The NN classifier (5) obtained incorrect decision (Fig. 1b). Correct photo from the gallery set (Fig. 1d) is only the seventh closest neighbor. Nevertheless, the proposed criterion (15) with $\lambda = 8$ and $M = 16$ successfully classifies this probe image (Table 4). In fact, though correct instance from the gallery set (Fig. 1b) is rather far from the recognized face, our regularization term $\sum_{i=1,i\neq c}^{C} \frac{(\rho_i(\mathbf{x})-\rho_{c;i}-(D-1)/(WH))^2}{\rho_{c;i}+(D-1)/(4WH)}$ is very low (0.0059). At the same time, other instances are penalized by this term due to significant differences in the pair-wise instance distances $\rho_{c;\,i}$ and the distances to the probe image $\rho_i(\mathbf{x})$ for most subjects i. Hence, even after multiplying this term to $\lambda/C \approx 0.071$ the resulted sum (total objective (15)) becomes slightly higher than the objective (15) of correct subject 942.

4.4. Experiments on LFW/YTF

In this subsection we discuss face identification results for LFW (Learned-Miller et al., 2016) and YTF (Wolf et al., 2011) datasets. The LFW protocols are focused on face verification task (Taigman et al., 2014). In order to introduce new challenges in face *identification* task, the novel protocol was presented by Best-Rowden et al. (2014). The $C = 596$ subjects who have at least two images in the LFW database and at least one video in the YTF database (subjects in YTF are a subset of those in LFW) are used to evaluate the performance of face identification. Training set contains exactly one facial image, all other images from LFW were put into the testing set. The rank-1 accuracy of random 10-times repeated cross-validation for the proposed method in comparison with the state-of-the-art results is shown in Table 5.

The facial regions were selected by either center crop of each image or MTCNN face detector (Zhang et al., 2016). We experimentally found that in this setup the rank-1 accuracy of the CenterFace and LightCNN features decreases at 13% and 35%, respectively, for the faces detected with the MTCNN. In the latter case the error rates decreased only for ResFace and VGGFace descriptors. Similarly to the previous experiment, the results for VGG2 (ResNet/SENet) are rather stable, though the accuracy with the MTCNN decreased at 1%. The center crop in this case with the proposed approach and the VGG2 (SENet) features made it possible to obtain the new state-of-the-art error rate for the LFW face identification setup, which is 0.56% lower (28% in relative terms) than the best known result obtained with an ensemble of very large CNNs trained by Baidu researchers (Liu et al., 2015).

We should emphasize that the accuracy of the face identification with contemporary facial descriptors is rather high due to the simplicity of the task. The situation changes dramatically if the training and testing images are significantly different. To support this statement, we followed a special protocol (Savchenko et al., 2018) for the still-to-video face identification originally described by Best-Rowden et al. (2014). Namely, we choose all photos and videos of $C = 1589$ classes as the intersection of people from the LFW and YTF datasets. It is important to mention that though the YTF dataset contains the images of celebrities from the LFW, the quality of the images in these datasets are completely different. The training set is filled by all $R = 4732$ photos of these $C = 1589$ identities taken from the LFW dataset. Our testing set contains 3353 clips from the YTF (in average, 183 frames per person). Center crop was used to select the facial regions. In our experiments we discovered that it is not necessary to process all frames in each video, hence, only each fifth frame in every video is presented in the testing set, given $T = 36$ frames per video in average.

At first, we examine the influence of additive noise on the face recognition quality. The uniform random number from the range $[-X_{\max}; X_{\max}]$ was added to every pixel of each testing images, where $X_{\max} \geq 0$ determines the noise level. Note that if $X_{\max} = 0$, original data from the YTF dataset are recognized. The main results of these experiments are shown in Figs. 2–5 for VGG-Face, ResFace, CenterFace and Light CNN, respectively.

Here, firstly, error rates drastically depend on the deep CNN used for feature extraction. In contrast to the previous experiments, the accuracy of the Light CNN (Wu et al., 2015) is approximately 20% higher than the accuracy of much widely used VGG-Face (Parkhi et al., 2015). Center loss (Wen et al., 2016) also looks rather promising.

Secondly, though the additive noise leads to degraded recognition accuracy in most cases, the increase of the error rate is rather low even if up to $X_{\max} = 10$ number is added to the value of each pixel. It is worth noting that treating the VGG-Face and ResFace features from the latter network as probability distributions and matching them with the KL divergence allows decreasing error rate at 1.4% in the best case for original YTF dataset ($X_{\max} = 0$). However, the Euclidean metric is more robust to the presence of noise. In fact, it is better to apply more robust probabilistic dissimilarities, which are based on testing for statistical homogeneity of the feature vectors (Savchenko, 2016; Savchenko & Belova, 2015). Nevertheless, the Light CNN significantly outperforms the VGG-Face and ResFace even for very high noise levels.

Thirdly, MLP significantly outperforms SVM in this task due to the usage of RPROP training procedure. As a matter of fact, the

Table 5
Rank-1 accuracy (%) in face identification for LFW dataset.

Classifier	Rank-1 accuracy
COTS-s1+s4 (Best-Rowden et al., 2014)	66.5
DeepFace (Taigman et al., 2014)	64.9
Web-Scale (Taigman, Yang, Ranzato, & Wolf, 2015)	82.5
VIPLFaceNetFull (Liu, Kan, Wu, Shan, & Chen, 2017)	92.79
Light CNN (C) (Wu et al., 2015)	93.8
DeepID2+ (Sun, Wang, & Tang, 2015)	95.0
DeepID3 (Sun, Liang, Wang, & Tang, 2015)	96.0
Light CNN-29 (Wu et al., 2015)	97.33
IDL ensemble (Liu et al., 2015)	98.03
Proposed approach, VGGFace, L_2, MTCNN face detection	87.35
Proposed approach, VGGFace, KL, MTCNN face detection	87.92
Proposed approach, VGGFace, L_2, center crop	84.61
Proposed approach, VGGFace, KL, center crop	85.22
Proposed approach, VGG2 (ResNet), L_2, MTCNN face detection	97.36
Proposed approach, VGG2 (ResNet), KL, MTCNN face detection	97.62
Proposed approach, VGG2 (ResNet), L_2, center crop	98.20
Proposed approach, VGG2 (ResNet), KL, center crop	98.43
Proposed approach, VGG2 (SENet), L_2, MTCNN face detection	97.55
Proposed approach, VGG2 (SENet), KL, MTCNN face detection	97.96
Proposed approach, VGG2 (SENet), L_2, center crop	98.51
Proposed approach, VGG2 (SENet), KL, center crop	98.59

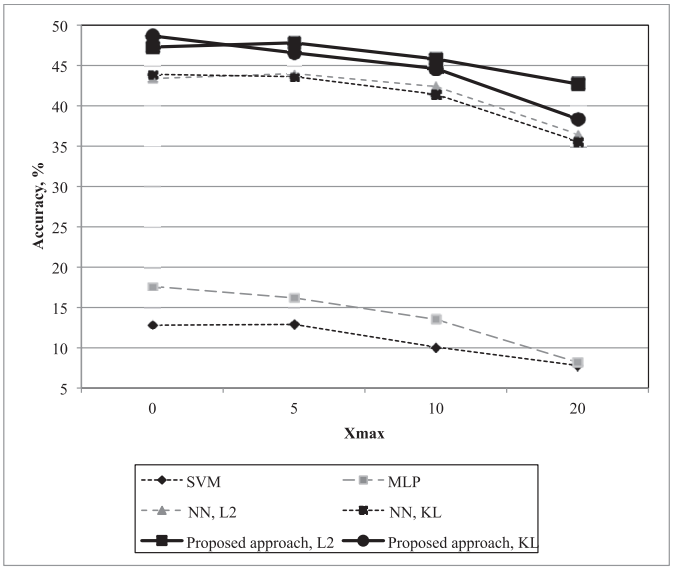

Fig. 2. Dependence of the recognition accuracy (%) on the noise level X_{max}, VGG-Face, YTF/LFW still-to-video identification.

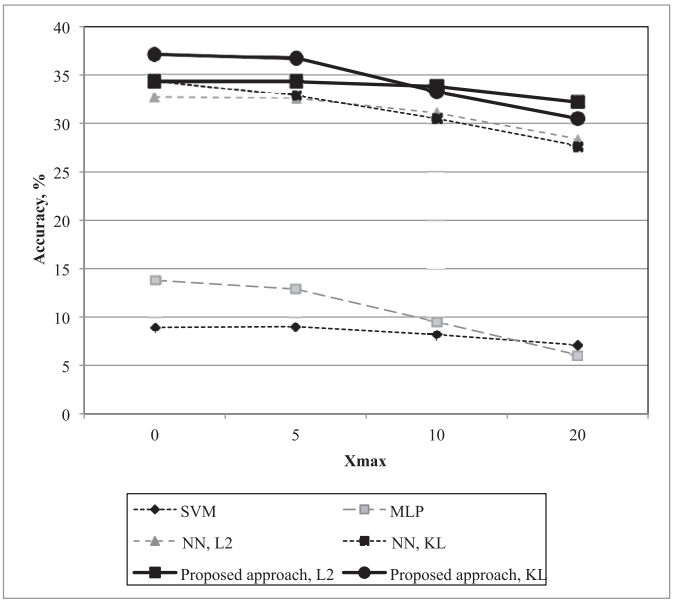

Fig. 3. Dependence of the recognition accuracy (%) on the noise level X_{max}, ResFace, YTF/LFW still-to-video identification.

MLP trained with traditional stochastic gradient descent remains 1–4% *less* accurate than our best SVM. However, both conventional classifiers (MLP, SVM) remain significantly worth than the NN-based criteria in all our experiments. It seems that the difference in training and testing sets make the complex classifier impossible to train, especially when the training sample is rather small.

Finally, the most important conclusion here is the highest accuracy of the proposed Algorithm 1 in all cases. Our approach is 3–5%, 20–40% and 25–45% more accurate when compared to NN, MLP and SVM, respectively. Moreover, though our idea is based on looking for correspondence of the distances $\rho_i(\mathbf{x})$ and $\rho_{c;\,i}$ (8), the gain in error rate does not significantly degrade even with the presence of noise in the testing images. It is important to highlight that though the original version of the proposed approach (14) was based on the properties of the KL divergence between positive features, our algorithm can be successfully used with the state-of-the-art distances and arbitrary feature vectors.

In the last experiments we demonstrate how to tune the parameters of our approach, namely, the importance λ of the introduced term (15), and the number of candidate classes M. The average accuracy and recognition time are presented in Figs. 6–8. For the two latter figures we use the best parameter λ obtained at Fig. 6.

Based on these results one can draw the following conclusions. Firstly, the U-curves in Fig. 6 prove that the proper choice of the parameter λ can significantly influence the recognition accuracy. It is especially true for the VGG-Face and ResFace features, in which the difference in error rates achieves 2.5%. However, the optimal value of this parameter is practically identical ($\lambda = 20..40$) for most combinations of the CNN and dissimilarity measure. The only one exception is the VGG2 (ResNet/SENet) with Euclidean distance, for which the top accuracy is achieved with the large values of λ. At the same time, the curves in Fig. 7 reach stability very fast with an increase of the number of candidate classes M. For example,

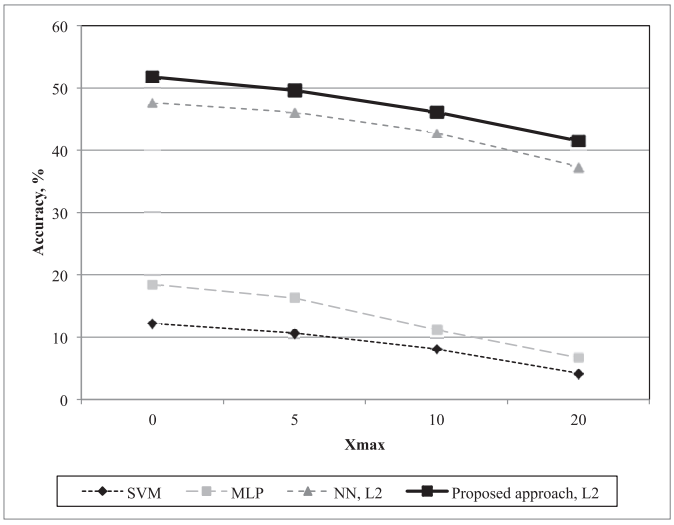

Fig. 4. Dependence of the recognition accuracy (%) on the noise level X_{max}, CenterFace, YTF/LFW still-to-video identification.

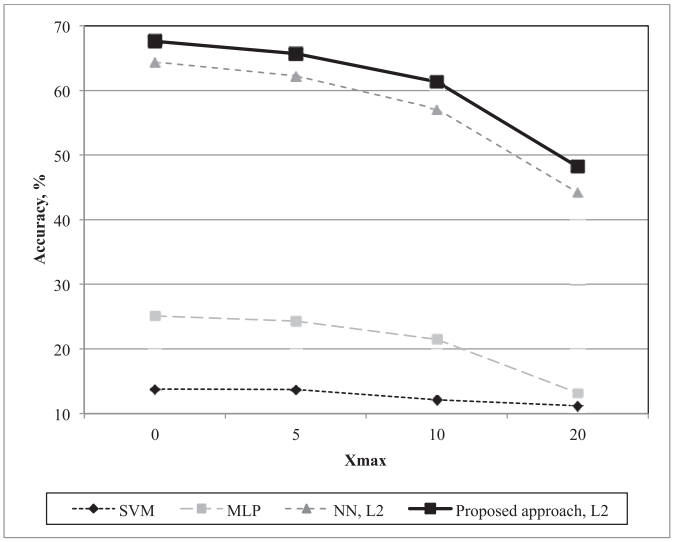

Fig. 5. Dependence of the recognition accuracy (%) on the noise level X_{max}, Light CNN (C), YTF/LFW still-to-video identification.

the accuracy for the VGG-Face features does not increase further if $M \geq 16$. Computation complexity linearly depends on the number of candidate classes M (Fig. 8). Though the examination of $M = 256$ classes causes the lowest error rates (Fig. 7), this value is not recommended for practical usage due to the very low recognition speed.

It is important to emphasize that in our experiments (Sokolova, Kharchevnikova, & Savchenko, 2017) all discovered feature extraction methods showed rather high quality of *face verification* using standard YTF protocol. In particular, we obtained the following estimates of AUC (Area Under Curve): 98.2%, 96.8%, 97.7% and 98.7% for VGG, ResFace, CenterFace and Light CNN, respectively. As one can notice from our results (Figs. 2–5), the situation becomes much more difficult for *face identification* task in our setup of image-set recognition. For example, the highest identification accuracy (Algorithm 1 with VGG2 (SENet) features) is less than 80% (Fig. 4). That is why we believe that it is very important to touch the limits of current deep learning methods in such challenging conditions (Zhou et al., 2015). However, from the best of our knowledge, this experimental setup

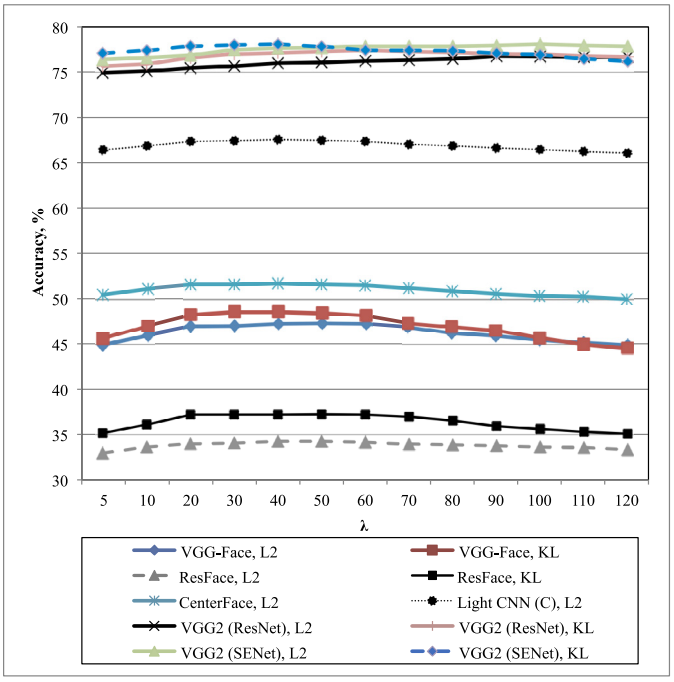

Fig. 6. Dependence of the recognition accuracy (%) on the importance λ of the introduced term (15), $M = C$, YTF/LFW still-to-video identification.

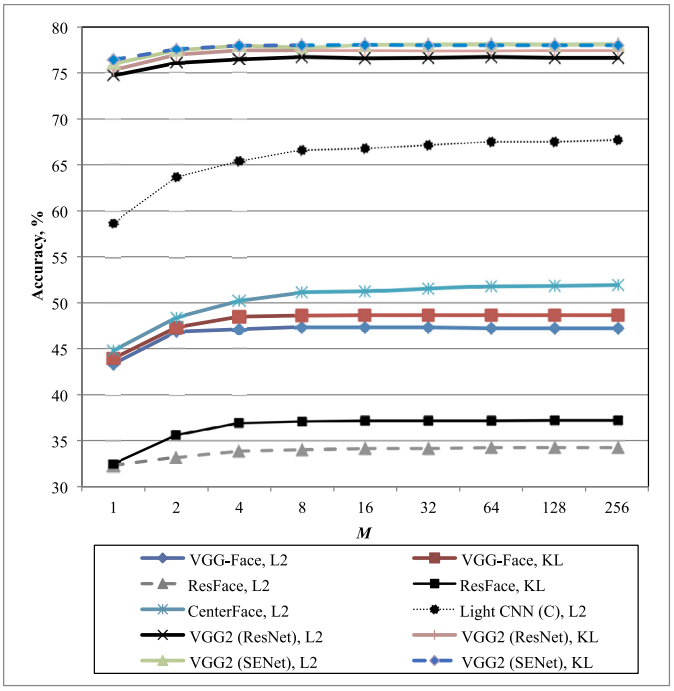

Fig. 7. Dependence of the recognition accuracy (%) on the number of candidate classes M, YTF/LFW still-to-video identification.

was used only in the original paper (Best-Rowden et al., 2014), in which an ensemble COTS-s1+s4 achieved 38.8% rank-1 accuracy. Thus, our approach made it possible to improve this accuracy at 39%, 30%, 23% and 9% for VGG2 (ResNet), Light CNN, CenterFace and VGG-Face descriptors, respectively. It is remarkable that, in contrast to all previous experiments, our approach with Euclidean distance between SENet features is sometimes *more* accurate here (Figs. 6 and 7) when compared to the matching of facial descriptors with the KL divergence. The usage of the proposed

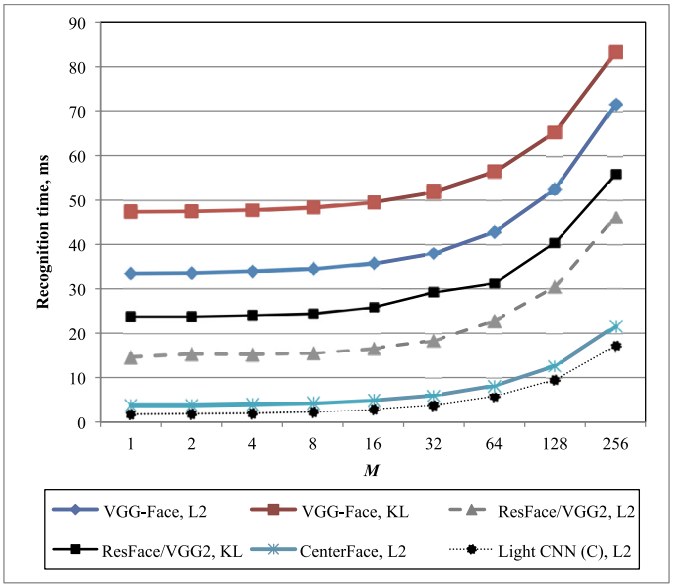

Fig. 8. Dependence of the image recognition time (ms) on the number of candidate classes M, YTF/LFW still-to-video identification.

Algorithm 1 with parameters $M = 64$, $\lambda = 100$ and the VGG2 (SENet) features matched with the L_2 distance made it possible to achieve new state-of-the-art (78.14% rank-1 accuracy) in this very complex experimental setup.

5. Conclusion

In this paper we proposed the novel MAP-based statistical approach to unconstrained face recognition (Algorithm 1), in which the joint density of distances to all reference images is maximized (7), (8). We have shown that this approach is implemented by introducing a special summand (15) in the nearest neighbor matching of high-dimensional off-the-shelf features from the outputs of the deep CNNs. This summand penalizes unreliable decisions obtained with the NN rule (Table 4). Here the resulted class c^* is defined as reliable only if the distances between the features of this image and the r-th instance are approximately equal to the distances between the reference images from the c^* and $c(r)$ classes for all $r \in \{1, \ldots, R\}$. Asymptotically normal distribution (7) of the KL divergence (Kullback, 1997) makes this definition correct for rather simple probabilistic model from Section 3.

The most important advantage of the proposed approach is the possibility to decrease the error rate for various high-dimensional feature sets and dissimilarity measures. Though our algorithm is based on asymptotical distribution (7) of the KL divergence, it can be successfully combined with the traditional Euclidean distance (Figs. 2 and 3). Moreover, our algorithm allows to increase accuracy even for general feature vectors with possible negative values (Figs. 4 and 5). Finally, we demonstrated the possibility to tune the parameters of the proposed algorithm (Fig. 7) in order to drastically reduce its computation complexity (Fig. 8) by proper choice of M candidate classes (compare the criterion (15) with the original one (14)). As a result, the proposed method repeated the state-of-the-art for IJB-A closed-set identification task (Table 1) and obtained the state-of-the-art rank-1 accuracies in the LFW face identification (Table 5) and the YTF/LFW still-to-video recognition (Figs. 6 and 7) using the protocols described by Best-Rowden et al. (2014).

It is necessary to mention the following disadvantages of our algorithm. Firstly, its computational complexity is still higher when compared to either baseline NN (5) or traditional SVM (Table 2). Performance degradation is especially noticeable if the number of classes is very large (Fig. 8) and the parameter M was set to achieve the highest accuracy. Secondly, proposed approach introduces new parameter λ, which should be carefully tuned (Fig. 6). In fact, as for the known similar regularization techniques, e.g. weight decay (Goodfellow et al., 2016), this parameter significantly influence the recognition accuracy. Hence, its incorrect choice can even increase the error rate.

The main direction for further research of the proposed algorithm is its applications with more accurate approximation of the distance probability distributions, e.g., the usage of the more appropriate Weibull distribution (Burghouts et al., 2008). Secondly, one important aspects with deep learning is that it treats feature learning and classifier learning together (LeCun et al., 2015). However in this paper the two steps are still separated: we proposed the novel classifier (Algorithm 1) for the off-the-shelf CNN features (Sharif Razavian et al., 2014). It is important to examine the possibility to incorporate the proposed summand term (15) into existing deep metric learning methods (Hu, Lu, & Tan, 2014; Lu, Wang, Deng, Moulin, & Zhou, 2015) in order to obtain the end-to-end classifier. Thirdly, it is important to examine the application of our method in still-to-video recognition task, in which the difference of training and testing data is acute. In particular, it is important to make our regularization smoother by taking into account the temporal coherence of sequential frames (Liu & Chen, 2003). Moreover, it is possible to combine our approach with the modern techniques with frames weighting in the aggregation module (Yang et al., 2017) to make distinction between frames with different quality (Huang et al., 2013). As it was shown in our experiments with IJB-A (Table 1) and LFW (Table 5) datasets, though the difference in accuracy of the proposed approach and conventional methods remains stable for different face detection techniques, they have significant impact on identification performance. Hence, it is also very important to examine facial descriptors, which are robust to the changes of bounding boxes (Ferrari et al., 2017). Finally, though the deep features are known to be highly embedded in the manifold space, non-Euclidean metrics, e.g, such techniques as the point-to-set metric learning (Huang, Wang, Shan, & Chen, 2014; Lu et al., 2015) or geometry aware feature matching (Harandi, Salzmann, & Hartley, 2014) can also be discussed in the future experiments.

Acknowledgement

The paper is supported by Russian Federation President grant no. MD-306.2017.9. The work in Section 3 was conducted by A.V. Savchenko at Laboratory of Algorithms and Technologies for Network Analysis, National Research University Higher School of Economics and supported by RSF grant 14-41-00039.

References

AbdAlmageed, W., Wu, Y., Rawls, S., Harel, S., Hassner, T., Masi, I., et al. (2016). Face recognition using deep multi-pose representations. In *Proceedings of the winter conference on applications of computer vision (WACV)* (pp. 1–9). IEEE.

Ahonen, T., Hadid, A., & Pietikainen, M. (2006). Face description with local binary patterns: Application to face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 28*(12), 2037–2041.

Baggio, D. L., Emami, S., Escrivá, D. M., Ievgen, K., Mahmood, N., Saragih, J., et al. (2012). *Mastering OpenCV with practical computer vision projects*. Packt Publishing Ltd.

Bashbaghi, S., Granger, E., Sabourin, R., & Bilodeau, G.-A. (2014). Watch-list screening using ensembles based on multiple face representations. In *Proceedings of the 22nd international conference on pattern recognition (ICPR)* (pp. 4489–4494). IEEE.

Best-Rowden, L., Han, H., Otto, C., Klare, B. F., & Jain, A. K. (2014). Unconstrained face recognition: Identifying a person of interest from a media collection. *IEEE Transactions on Information Forensics and Security, 9*(12), 2144–2157.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.

Burghouts, G., Smeulders, A., & Geusebroek, J.-M. (2008). The distribution family of similarity distances. In *Proceedings of the international conference on advances in neural information processing systems (NIPS)* (pp. 201–208).

Cao, Q., Shen, L., Xie, W., Parkhi, O. M., & Zisserman, A. (2017). VGGFace2: A dataset for recognising faces across pose and age. arXiv preprint: 1710.08092.

Cao, X., Wipf, D., Wen, F., Duan, G., & Sun, J. (2013). A practical transfer learning algorithm for face verification. In *Proceedings of the international conference on computer vision (ICCV)* (pp. 3208–3215). IEEE.

Cevikalp, H., & Triggs, B. (2010). Face recognition based on image sets. In *Proceedings of the international conference on computer vision and pattern recognition (CVPR)* (pp. 2567–2573). IEEE.

Chen, D., Hua, G., Wen, F., & Sun, J. (2016). Supervised transformer network for efficient face detection. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 122–138). Springer.

Chen, J.-C., Patel, V. M., & Chellappa, R. (2016). Unconstrained face verification using deep CNN features. In *Proceedings of the winter conference on applications of computer vision (WACV)* (pp. 1–9). IEEE.

Chen, L.-F., Liao, H.-Y. M., Ko, M.-T., Lin, J.-C., & Yu, G.-J. (2000). A new LDA-based face recognition system which can solve the small sample size problem. *Pattern Recognition, 33*(10), 1713–1726.

Chowdhury, A. R., Lin, T.-Y., Maji, S., & Learned-Miller, E. (2016). One-to-many face recognition with bilinear CNNs. In *Proceedings of the winter conference on applications of computer vision (WACV)* (pp. 1–9). IEEE.

Crosswhite, N., Byrne, J., Stauffer, C., Parkhi, O., Cao, Q., & Zisserman, A. (2017). Template adaptation for face verification and identification. In *Proceedings of the 12th international conference on automatic face & gesture recognition (FG)* (pp. 1–8). IEEE.

Dalal, N., & Triggs, B. (2005). Histograms of oriented gradients for human detection. In *Proceedings of the international conference on computer vision and pattern recognition (CVPR): 1* (pp. 886–893). IEEE.

Dewan, M. A. A., Granger, E., Marcialis, G.-L., Sabourin, R., & Roli, F. (2016). Adaptive appearance model tracking for still-to-video face recognition. *Pattern Recognition, 49*, 129–151.

Ding, C., & Tao, D. (2018). Trunk-branch ensemble convolutional neural networks for video-based face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Ferrari, C., Lisanti, G., Berretti, S., & Del Bimbo, A. (2017). Investigating nuisance factors in face recognition with dcnn representation. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops (CVPRW)* (pp. 81–89).

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT press.

Guo, Y., & Zhang, L. (2017). One-shot face recognition by promoting underrepresented classes. arXiv preprint: 1707.05574.

Harandi, M. T., Salzmann, M., & Hartley, R. (2014). From manifold to manifold: Geometry-aware dimensionality reduction for SPD matrices. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 17–32). Springer.

Hong, S., Im, W., Ryu, J., & Yang, H. S. (2017). SSPP-DAN: Deep domain adaptation network for face recognition with single sample per person. arXiv preprint: 1702.04069.

Hu, J., Lu, J., & Tan, Y.-P. (2014). Discriminative deep metric learning for face verification in the wild. In *Proceedings of the international conference on computer vision and pattern recognition (CVPR)* (pp. 1875–1882). IEEE.

Hu, J., Shen, L., & Sun, G. (2017). Squeeze-and-excitation networks. arXiv preprint: 1709.01507.

Hu, P., & Ramanan, D. (2017). Finding tiny faces. In *Proceedings of the international conference on computer vision and pattern recognition (CVPR)* (pp. 1522–1530). IEEE.

Huang, Z., Shan, S., Zhang, H., Lao, S., Kuerban, A., & Chen, X. (2012). Benchmarking still-to-video face recognition via partial and local linear discriminant analysis on COX-S2V dataset. In *Proceedings of the Asian conference on computer vision (ACCV)* (pp. 589–600). Springer.

Huang, Z., Wang, R., Shan, S., & Chen, X. (2014). Learning Euclidean-to-Riemannian metric for point-to-set classification. In *Proceedings of the conference on computer vision and pattern recognition (CVPR)* (pp. 1677–1684). IEEE.

Huang, Z., Zhao, X., Shan, S., Wang, R., & Chen, X. (2013). Coupling alignments with recognition for still-to-video face recognition. In *Proceedings of the international conference on computer vision (ICCV)* (pp. 3296–3303). IEEE.

Klare, B. F., Klein, B., Taborsky, E., Blanton, A., Cheney, J., Allen, K., et al. (2015). Pushing the frontiers of unconstrained face detection and recognition: IARPA Janus Benchmark A. In *Proceedings of the international conference on computer vision and pattern recognition (CVPR)* (pp. 1931–1939). IEEE.

Klontz, J. C., Klare, B. F., Klum, S., Jain, A. K., & Burge, M. J. (2013). Open source biometric recognition. In *Proceedings on the international conference on biometrics: Theory, applications and systems (BTAS)* (pp. 1–8). IEEE.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. In *Proceedings of the international conference on advances in neural information processing systems (NIPS)* (pp. 1097–1105).

Kullback, S. (1997). Information theory and statistics. *Dover books on mathematics*. Dover Publications.

Learned-Miller, E., Huang, G. B., RoyChowdhury, A., Li, H., & Hua, G. (2016). Labeled faces in the wild: A survey. In *Proceedings of the international conference on advances in face detection and facial image analysis* (pp. 189–248). Springer.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521*(7553), 436–444.

Leonidou, M., Tsapatsoulis, N., & Kollias, S. (1999). Face recognition based on multiple representations: Splitting the error space. In *Advances in intelligent systems* (pp. 345–356). Springer.

Li, Q., Wang, H. J., You, J., Li, Z. M., & Li, J. X. (2013). Enlarge the training set based on inter-class relationship for face recognition from one image per person. *PloS one, 8*(7), e68539.

Liu, J., Deng, Y., Bai, T., Wei, Z., & Huang, C. (2015). Targeting ultimate accuracy: Face recognition via deep embedding. arXiv preprint: 1506.07310.

Liu, L., Zhang, L., Liu, H., & Yan, S. (2014). Toward large-population face identification in unconstrained videos. *IEEE Transactions on Circuits and Systems for Video Technology, 24*(11), 1874–1884.

Liu, X., & Chen, T. (2003). Video-based face recognition using adaptive hidden Markov models. In *Proceedings of the international conference on computer vision and pattern recognition (CVPR): 1*. IEEE.

Liu, X., Kan, M., Wu, W., Shan, S., & Chen, X. (2017). VIPLFaceNet: An open source deep face recognition SDK. *Frontiers of Computer Science, 11*(2), 208–218.

Lu, J., Tan, Y.-P., & Wang, G. (2013). Discriminative multimanifold analysis for face recognition from a single training sample per person. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 35*(1), 39–51.

Lu, J., Wang, G., Deng, W., Moulin, P., & Zhou, J. (2015). Multi-manifold deep metric learning for image set classification. In *Proceedings of the international conference on computer vision and pattern recognition (CVPR)* (pp. 1137–1145). IEEE.

Masi, I., Rawls, S., Medioni, G., & Natarajan, P. (2016a). Pose-aware face recognition in the wild. In *Proceedings of the international conference on computer vision and pattern recognition (CVPR)* (pp. 4838–4846). IEEE.

Masi, I., Tran, A., Hassner, T., Leksut, J. T., & Medioni, G. (2016b). Do we really need to collect millions of faces for effective face recognition?. In *Proceedings of the european conference on computer vision (ECCV)*.

Mian, A., Hu, Y., Hartley, R., & Owens, R. (2013). Image set based face recognition using self-regularized non-negative coding and adaptive distance metric learning. *IEEE Transactions on Image Processing, 22*(12), 5252–5262.

Micó, M. L., Oncina, J., & Vidal, E. (1994). A new version of the nearest-neighbour approximating and eliminating search algorithm (AESA) with linear preprocessing time and memory requirements. *Pattern Recognition Letters, 15*(1), 9–17.

Mokhayeri, F., Granger, E., & Bilodeau, G.-A. (2015). Synthetic face generation under various operational conditions in video surveillance. In *Proceedings of the international conference on image processing (ICIP)* (pp. 4052–4056). IEEE.

Ortiz, E. G., & Becker, B. C. (2014). Face recognition for web-scale datasets. *Computer Vision and Image Understanding, 118*, 153–170.

Parchami, M., Bashbaghi, S., & Granger, E. (2017). CNNs with cross-correlation matching for face recognition in video surveillance using a single training sample per person. In *Proceedings of the international conference on advanced video and signal based surveillance (AVSS)* (pp. 1–6). IEEE.

Parkhi, O. M., Vedaldi, A., & Zisserman, A. (2015). Deep face recognition. In *Proceedings of the British conference on machine vision (BMVC): 1* (p. 6).

Phillips, P. J., Grother, P., Micheals, R., Blackburn, D. M., Tabassi, E., & Bone, M. (2003). Face recognition vendor test 2002. In *Proceedings of the ieee international workshop on analysis and modeling of faces and gestures (AMFG)* (p. 44).

Prince, S. J. (2012). *Computer vision: Models, learning, and inference*. Cambridge University Press.

Prince, S. J., Elder, J. H., Warrell, J., & Felisberti, F. M. (2008). Tied factor analysis for face recognition across large pose differences. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 30*(6), 970–984.

Raudys, S. J., Jain, A. K., et al. (1991). Small sample size effects in statistical pattern recognition: Recommendations for practitioners. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 13*(3), 252–264.

Ross, G. (1969). Algorithm as 15: Single linkage cluster analysis. *Journal of the Royal Statistical Society. Series C (Applied Statistics), 18*(1), 106–110.

Sankaranarayanan, S., Alavi, A., Castillo, C. D., & Chellappa, R. (2016). Triplet probabilistic embedding for face verification and clustering. In *Proceedings of the 8th international conference on biometrics theory, applications and systems (BTAS)* (pp. 1–8). IEEE.

Savchenko, A. (2017a). Maximum-likelihood dissimilarities in image recognition with deep neural networks. *Computer Optics, 41*(3), 422–430.

Savchenko, A. V. (2012). Directed enumeration method in image recognition. *Pattern Recognition, 45*(8), 2952–2961.

Savchenko, A. V. (2016). *Search techniques in intelligent classification systems*. Springer.

Savchenko, A. V. (2017b). Deep convolutional neural networks and maximum-likelihood principle in approximate nearest neighbor search. In *Proceedings of the international conference on pattern recognition and image analysis (IBPRIA 2017). Lecture notes in computer science: 10255* (pp. 42–49). Springer, Cham.

Savchenko, A. V. (2017c). Maximum-likelihood approximate nearest neighbor method in real-time image recognition. *Pattern Recognition, 61*, 459–469.

Savchenko, A. V., & Belova, N. S. (2015). Statistical testing of segment homogeneity in classification of piecewise–regular objects. *International Journal of Applied Mathematics and Computer Science, 25*(4), 915–925.

Savchenko, A. V., Belova, N. S., & Savchenko, L. V. (2018). Fuzzy analysis and deep convolution neural networks in still-to-video recognition. *Optical Memory and Neural Networks (Information Optics), 27*(1), 23–31.

Schroff, F., Kalenichenko, D., & Philbin, J. (2015). FaceNet: A unified embedding for face recognition and clustering. In *Proceedings of the international conference on computer vision and pattern recognition (CVPR)* (pp. 815–823). IEEE.

Shakhnarovich, G., Fisher, J. W., & Darrell, T. (2002). Face recognition from long-term observations. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 851–865). Springer.

Sharif Razavian, A., Azizpour, H., Sullivan, J., & Carlsson, S. (2014). CNN features off-the-shelf: An astounding baseline for recognition. In *Proceedings of the international conference on computer vision and pattern recognition workshops (CVPRW)* (pp. 806–813). IEEE.

Simonyan, K., Parkhi, O. M., Vedaldi, A., & Zisserman, A. (2013). Fisher vector faces in the wild.. In *Proceedings of the british machine vision conference (BMVC): 2* (p. 4).

Sohn, K., Liu, S., Zhong, G., Yu, X., Yang, M.-H., & Chandraker, M. (2017). Unsupervised domain adaptation for face recognition in unlabeled videos. In *Proceedings of the international conference on computer vision (ICCV)* (pp. 3210–3218). IEEE.

Sokolova, A. D., Kharchevnikova, A. S., & Savchenko, A. V. (2017). Organizing multimedia data in video surveillance systems based on face verification with convolutional neural networks. In *Proceedings of the international conference on analysis of images, social networks and texts (AIST). Lecture notes in computer science: 10716* (pp. 223–230). Springer.

Sun, Y., Chen, Y., Wang, X., & Tang, X. (2014). Deep learning face representation by joint identification-verification. In *Proceedings of the international conference on advances in neural information processing systems (NIPS)* (pp. 1988–1996).

Sun, Y., Liang, D., Wang, X., & Tang, X. (2015). DeepID3: Face recognition with very deep neural networks. arXiv preprint: 1502.00873.

Sun, Y., Wang, X., & Tang, X. (2015). Deeply learned face representations are sparse, selective, and robust. In *Proceedings of the conference on computer vision and pattern recognition (CVPR)*. IEEE.

Szeliski, R. (2010). *Computer vision: Algorithms and applications*. Springer Science & Business Media.

Taigman, Y., Yang, M., Ranzato, M., & Wolf, L. (2014). DeepFace: Closing the gap to human-level performance in face verification. In *Proceedings of the conference on computer vision and pattern recognition (CVPR)* (pp. 1701–1708). IEEE.

Taigman, Y., Yang, M., Ranzato, M., & Wolf, L. (2015). Web-scale training for face identification. In *Proceedings of the conference on computer vision and pattern recognition (CVPR)* (pp. 2746–2754). IEEE.

Tan, X., Chen, S., Zhou, Z.-H., & Zhang, F. (2006). Face recognition from a single image per person: A survey. *Pattern Recognition, 39*(9), 1725–1745.

Viola, P., & Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. In *Proceedings of the international conference on computer vision and pattern recognition (CVPR): 1*. IEEE. I–I

Wang, D., Otto, C., & Jain, A. K. (2015). Face search at scale: 80 million gallery. arXiv preprint: 1507.07242.

Wen, Y., Zhang, K., Li, Z., & Qiao, Y. (2016). A discriminative feature learning approach for deep face recognition. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 499–515). Springer.

Wolf, L., Hassner, T., & Maoz, I. (2011). Face recognition in unconstrained videos with matched background similarity. In *Proceedings of the international conference on computer vision and pattern recognition (CVPR)* (pp. 529–534). IEEE.

Wolf, L., & Shashua, A. (2003). Kernel principal angles for classification machines with applications to image sequence interpretation. In *Proceedings of the international conference on computer vision and pattern recognition (CVPR): 1*. IEEE.

Wu, X., He, R., Sun, Z., & Tan, T. (2015). A light CNN for deep face representation with noisy labels. arXiv preprint: 1511.02683.

Yang, J., Ren, P., Chen, D., Wen, F., Li, H., & Hua, G. (2017). Neural aggregation network for video face recognition. In *Proceedings of the international conference on computer vision and pattern recognition (CVPR)* (pp. 4362–4371). IEEE.

Zeng, J.-Y., Zhao, X.-X., Zhai, Y.-K., Gan, J.-Y., Lin, Z.-Y., & Qin, C.-B. (2017). A novel expanding sample method for single training sample face recognition. In *Proceedings of the international conference on wavelet analysis and pattern recognition (ICWAPR)* (pp. 33–37). IEEE.

Zhang, D., Chen, S., & Zhou, Z.-H. (2005). A new face recognition method based on svd perturbation for single example image per person. *Applied Mathematics and computation, 163*(2), 895–907.

Zhang, G., Hu, J., Xiang, H., & Zhao, Y. (2017). Multiple representation based sample diversity for face recognition. *Optik-International Journal for Light and Electron Optics, 138*, 529–534.

Zhang, G., Huang, X., Li, S. Z., Wang, Y., & Wu, X. (2004). Boosting local binary pattern (LBP)-based face recognition.. In *Sinobiometrics* (pp. 179–186). Springer.

Zhang, J., Yan, Y., & Lades, M. (1997). Face recognition: Eigenface, elastic matching, and neural nets. *Proceedings of the IEEE, 85*(9), 1423–1435.

Zhang, K., Zhang, Z., Li, Z., & Qiao, Y. (2016). Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters, 23*(10), 1499–1503.

Zhang, N., Yang, J., & Qian, J.-j. (2012). Component-based global k-NN classifier for small sample size problems. *Pattern Recognition Letters, 33*(13), 1689–1694.

Zhao, Y., Liu, Y., Liu, Y., Zhong, S., & Hua, K. A. (2015). Face recognition from a single registered image for conference socializing. *Expert Systems with Applications, 42*(3), 973–979.

Zhou, E., Cao, Z., & Yin, Q. (2015). Naive-deep face recognition: Touching the limit of LFW benchmark or not? arXiv preprint: 1501.04690.

Zhu, P., Yang, M., Zhang, L., & Lee, I.-Y. (2014). Local generic representation for face recognition with single sample per person. In *Proceedings of the Asian conference on computer vision (ACCV)* (pp. 34–50). Springer.

Zhu, Y., Zheng, Z., Li, Y., Mu, G., Shan, S., & Guo, G. (2015). Still to video face recognition using a heterogeneous matching approach. In *Proceedings of the international conference on biometrics theory, applications and systems (BTAS)* (pp. 1–6). IEEE.