Maximum A Posteriori Estimation of Distances Between Deep Features in Still-to-Video Face Recognition

Andrey V. Savchenko

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

avsavchenko@hse.ru

Natalya S. Belova

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

nbelova@hse.ru

August 29, 2017

Abstract

The paper deals with the still-to-video face recognition for the small sample size problem based on computation of distances between high-dimensional deep bottleneck features. We present the novel statistical recognition method, in which the still-to-video recognition task is casted into Maximum A Posteriori estimation. In this method we maximize the joint probabilistic density of the distances to all reference still images. It is shown that this likelihood can be estimated with the known asymptotically normal distribution of the Kullback-Leibler discriminations between nonnegative features. The experimental study with the LFW (Labeled Faces in the Wild), YTF (YouTube Faces) and IJB-A (IARPA Janus Benchmark A) datasets has been provided. We demonstrated, that the proposed approach can be applied with the state-of-the-art deep features and dissimilarity measures. Our algorithm achieves 3-5% higher accuracy when compared with conventional aggregation of decisions obtained for all frames.

Abbreviations

- CNN - Convolution Neural Network

- i.i.d. - independent identically distributed

- IJB-A - IARPA Janus Benchmark A

1

- KL - Kullback-Leibler divergence

- LFW - Labeled Faces in the Wild

- MAP - Maximum A Posteriori

- ML - Maximum Likelihood

- NN - Nearest Neighbor

- YTF - YouTube Faces

1 Introduction

The usage of deep learning technologies [1] instead of traditional computer vision methods [2] has recently made it possible to achieve a near human-level performance in various face recognition tasks, such as verification [3] or identification [4]. Moreover, contemporary techniques can deal even with well-known crucial issues appeared in practical applications of face recognition, e.g., unconstrained environment (various illumination and pose, partial occlusion) [5], or the small sample size problem [6], when usually only single facial image per person is available [7, 8]. The latter problem is solved using transfer learning methods [9], in which large external datasets of celebrities are used to train deep convolutional neural network (CNN) [10, 11]. This CNN is further applied to extract features of the training images from the limited sample of subjects of interest using the outputs of a bottleneck layer [12]. Such transfer learning technique makes it possible to create a classifier that ideally performs nearly as well as if rich dataset of photos of these individuals were present [9].

Though the algorithms that do recognition from a single test photo have its own practical applications, e.g., search for a person in a social network, one of the most significant use-cases is provided by surveillance and authentication systems. Hence, all the more attention is nowadays paid to video face recognition [13], in which a set or a sequence of observed images of the same individual are available for decision-making [14, 15, 16]. Despite the large number of recent papers devoted to video face recognition [17, 18, 19, 20], this task still remains very challenging. For example, most of the known algorithms suffered from heavy off-line training load [18], and they were experimentally studied only with databases, in which the still images are gathered under controlled environment. In this paper we proposed the novel still-to-video face recognition method suitable for application with the small sample of still photos per person, which is based on the probabilistic interpretation [6] without learning temporal coherence [4]. At first, the nearest neighbor (NN) reference images are examined for each video frame. Next, computed distances to all still images are used to weight the recognition results for each frame 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 frame and this subject. The likelihoods (joint probabilistic densities) are computed using the idea of the maximum-likelihood approximate NN algorithm [21] by assuming that the Kullback-Leibler (KL) divergence [22] is used

to compare distances. However, we demonstrate that our approach can be successfully applied with much more widely used dissimilarity measures.

The rest of the paper is organized as follows. Section 2 contains the literature survey of related papers. In Section 3 we present a simple statistical formulation of the still-to-video recognition task [23] using the KL minimum discrimination principle. In Section 4 we propose 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. In Section 5 we present the complete algorithm, in which the initial assumptions are facilitated and the computational complexity is taken into account. Section 6 presents the experimental results in recognition of videos from either IJB-A (IARPA Janus Benchmark A) [24] or YouTube Faces (YTF) datasets [25] with the still images from the Labeled Faces in the Wild (LFW) dataset [5] using such deep CNNs as VGGNet [10] and Lightened CNN [11]. Finally, concluding comments are given in Section 7.

2 Related works

Compared to still-to-still recognition, the requirement of real-time processing in video face recognition prevents an implementation of too sophisticated recognition algorithms [26], especially, when the number of individuals is rather large [21]. However, more information of the individuals can be exploited from several video frames [20], and the recognition results can be generally improved when using multiple images or video sequences [27]. Most of the papers devoted to person identification or retrieval in surveillance videos [14, 28, 15, 25, 20] address the video-to-video face recognition problem, in which query videos are matched against a set of reference video sets [18]. However, real-world applications usually involve the still-to-video face recognition scenario, when only very few still images per person are enrolled into the training set while the input sequence of video frames is captured [17, 19]. This problem is very difficult, because the testing videos have generally low quality and are captured under poor illumination and pose [29].

There are two types of still-to-video recognition methods [18]. First group of methods exploits the temporal information in a video sequence. Liu and Chen used the hidden Markov models in recognition of video frames [28]. The paper [30] describes successful application of a state space model [2] parametrized by a tracking state vector and a recognizing identity variable, simultaneously characterizing the kinematics and identity of humans. All such methods require large datasets of video clips to train the extraction of dynamic information. Moreover, their performance relied on strong temporal coherence in training and testing environment [18]. Hence, another group of methods with accumulation of multiple observations of the same individual has recently been studied. These methods are usually based on fusion of evidence from multiple measurements [31]. For example, the authors of the paper [27] extended the probabilistic appearance-based face recognition approach to work with multiple images and video sequences. Shakhnarovich et al [23] assumed that the multiple frames for each human subject followed the normal distribution, and they used the KL divergence [22] to measure the distance between distributions. Several important works

3

on still-to-video face recognition adopted a metric learning framework. For example, Zhu et al [32] extended the well-known Mahalonbis metric learning to point-to-set and set-to-set distance learning. Huang et al [33] applied Euclidean-to-Riemannian metric learning framework to image-to-set object classification and still-to-video face recognition tasks.

Nowadays, it is more typical to map each face image into a feature vector using a deep CNN. Next, a naive approach would be representing a video face as a set of frame-level face features from the CNN [34]. Such a representation obviously maintains all the information across all frames. To speed-up the recognition process, the features of all video frames are aggregated into a single vector (video face representation) using such pooling strategies as average and max pooling [35, 36]. The Eigen-PEP representations [36] integrates the visual information from all relevant video frames using part-based average pooling through the probabilistic elastic part model. Then the intermediate representation is compressed through principal component analysis, and only a number of principal eigen dimensions are kept. Yang et al [20] proposed the learning of the frames weights in the aggregation module, which consists of two attention blocks driven by a memory storing all the extracted features. Canziani and Culurciello proposed the CortexNet neural architecture [37] in order to obtain a robust and stable representation of temporal visual inputs. In the paper [16] two video sequences representing head motion and facial expressions are compared using a new positive definite kernel based on the concept of principal angles between two linear subspaces. Most of described methods learned the relationship between the still images and video frames but did not directly handle bad quality frames, which very likely make the recognition perform badly. To deal with this problem, another kind of methods such as [38] was proposed, which first select the best quality frames, perform facial alignments and then integrate the recognition results of selected frames [17]. Thus, one can notice from this brief survey, that the methods without learning temporal coherence are more widely used nowadays. Hence, in our paper we decided to focus on application of statistical recognition of individual frames [23] described by the modern deep CNN features [1].

3   Statistical still-to-video face recognition

The task of closed-set still-to-video face identification is to assign an observed sequence of T video frames to one of $C > 1$ classes (identities). The classes are specified by the training set of $R \geq C$ still images. We consider the supervised learning case, when the class label (subject id) $c(r) \in \{1, ..., C\}$ of the rth photo is known. For simplicity, we assume that only one image per person is available ($C = R, c(r) = r$), and all frames contain facial region of only one identity, i.e., the faces are detected, and the whole video clip is clustered to extract sequential frames containing the same subject. At first, each image should be described with a feature vector using the preliminarily trained deep CNN, as it was described in introduction. The outputs at the last layer of this CNN for all t frames and each r-th reference image are used as the D-dimensional feature vectors $\mathbf{x}(t) = [x_1(t), ..., x_D(t)]$ and $\mathbf{x}_r = [x_{r;1}, ..., x_{r;D}]$, respectively.

Let us assume that the normalized feature vector $\mathbf{x}(t)$ is the estimate of multinomial distribution of (hypothetical) random variable $X(t), t \in \{1, ..., T\}$. We also assume

4

that each rth instance represents the probabilistic distribution of the random variable $X_r, r \in \{1, ..., R\}$. The video face recognition task is reduced to a problem of statistical testing of simple hypothesis $W_r, r \in \{1, ..., R\}$ for all frames. Statistically optimal decision is defined as follows:

$$\max_{r \in \{1,...,R\}} P(W_r|\mathbf{x}(1), ..., \mathbf{x}(T)), \tag{1}$$

where the posterior probability is estimated using the Bayes rule:

$$P(W_r|\mathbf{x}(1), ..., \mathbf{x}(T)) = \frac{f(\mathbf{x}(1), ..., \mathbf{x}(T)|W_r)}{\sum\limits_{i=1}^{R} f(\mathbf{x}(1), ..., \mathbf{x}(T)|W_i)}, \tag{2}$$

In this paper we focus on the case of full prior uncertainty, hence, the prior probabilities of observing each subject here are equal, so the rule (1) represents the maximum likelihood (ML) criterion. Finally, it is necessary to perform aggregation of the frame recognition results [26]. Kittler et al [31] presented a statistical interpretation of a number of common methods for cross-modal fusion, such as the product, sum, maximum, and majority vote rules, which are also appropriate for late integration over a set of observations from a single modality. For example, it can be shown that under the assumption that the frames in the video sequence correspond to independent identically distributed (i.i.d) random variables, the likelihood is estimated using the product rule [23]

$$f(\mathbf{x}(1), ..., \mathbf{x}(T)|W_r) = \prod_{t=1}^{T} \prod_{d=1}^{D} (x_{r;d})^{n \cdot x_d(t)}, \tag{3}$$

where n is the sample size used to estimate the distribution $\mathbf{x}(t)$. In practice, this parameter can be chosen proportional to the total number of pixels in the input image. By substituting (3) in (2) and dividing its numerator and denominator to $\prod\limits_{t=1}^{T} \prod\limits_{d=1}^{D} (x_d(t))^{n \cdot x_d(t)}$, one can obtain the following estimate of the posterior probability:

$$P(W_r|\mathbf{x}(1), ..., \mathbf{x}(T)) = \frac{\exp(-n \sum\limits_{t=1}^{T} \sum\limits_{d=1}^{D} x_d(t) \ln(\frac{x_d(t)}{x_{r;d}}))}{\sum\limits_{i=1}^{R} \exp(-n \sum\limits_{t=1}^{T} \sum\limits_{d=1}^{D} x_d(t) \ln(\frac{x_d(t)}{x_{i;d}}))}, \tag{4}$$

or

$$P(W_r|\mathbf{x}(1), ..., \mathbf{x}(T)) = \frac{\exp\left(-n \sum\limits_{t=1}^{T} I(\mathbf{x}(t) : \mathbf{x}_r)\right)}{\sum\limits_{i=1}^{R} \exp\left(-n \sum\limits_{t=1}^{T} I(\mathbf{x}(t) : \mathbf{x}_i)\right)}, \tag{5}$$

where $I(\mathbf{x}(t) : \mathbf{x}_r))$ is the KL divergence between feature vectors $\mathbf{x}(t)$ and $\mathbf{x}_r$.

Hence, the ML criterion (1) can be written in the simplified form:

$$\min_{r \in \{1,...,R\}} \sum_{t=1}^{T} I(\mathbf{x}(t) : \mathbf{x}_r). \tag{6}$$

Thus, the ML decision (1) of the still-to-video face recognition task [23] and assumptions about i.i.d. frames is equivalent to the KL minimum information discrimination principle [22].

4 Maximum a posteriori estimation of distances between deep bottleneck features

In this paper we use the idea of the maximal-likelihood approximate NN method from the paper [12, 21]. Namely, we exploit 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 [22]. Hence, another recognition criterion is used, in which the maximum posterior probability based on the joint distribution of all R-dimensional random vectors of *distances* $\mathbf{I}(t) = [I(\mathbf{x}(t) : \mathbf{x}_1), ..., I(\mathbf{x}(t) : \mathbf{x}_R)]$ is obtained:

$$\max_{r \in \{1,...,R\}} f(W_r | \mathbf{I}(1), ..., \mathbf{I}(T)). \tag{7}$$

To deal with this conditional density of distances let us firstly consider much more simple marginal conditional distribution $f(\mathbf{I}(t)|W_r), t \in \{1, ..., T\}$. By using a natural assumption about independence of the still photos of all identities from the training set, we estimate the joint density of distances between reference images and the tth frame $f(\mathbf{I}(t)|W_r)$ as follows:

$$f(\mathbf{I}(t)|W_r) = f(I(\mathbf{x}(t) : \mathbf{x}_r)|W_r) \cdot \prod_{i=1, i \neq r}^{R} f(I(\mathbf{x}(t) : \mathbf{x}_i)|W_r). \tag{8}$$

At first, let us consider each term in the second multiplier. We propose to estimate the conditional density function $f(I(\mathbf{x}(t) : \mathbf{x}_i)|W_r)$ of the distance between the observed object from the rth class and the ith instance using the known asymptotic properties of the KL divergence. It is known [22] that if $i \neq r$, then $2n$-times KL divergence $I(\mathbf{x}(t) : \mathbf{x}_r)$ is asymptotically distributed as the non-central chi-squared with $D-1$ degrees of freedom and the non-centrality parameter $2n \cdot I(\mathbf{x}_r : \mathbf{x}_i)$. This asymptotic distribution is also known to be achieved for such probabilistic dissimilarities as the chi-squared distance, the Jenson-Shannon divergence, etc. [12, 22]. As the number of features D is usually large, it is possible to approximate the non-central chi-squared distribution with the Gaussian distribution [21]. Hence, we use the asymptotically normal distribution of the distance between the facial image from the rth class and the ith reference photo:

$$\mathcal{N}\left(I(\mathbf{x}_r : \mathbf{x}_i) + \frac{D-1}{n}, \frac{4n \cdot I(\mathbf{x}_r : \mathbf{x}_i) + (D-1)}{2n^2}\right) \tag{9}$$

6

Thus, the conditional distributions in the second multiplier in (8) are estimated as follows:

$$f(I(\mathbf{x}(t) : \mathbf{x}_i)|W_r) = \exp(-n \cdot \phi_{r;i}(\mathbf{x}(t))), \tag{10}$$

where we denote $\phi_{r;i}(\mathbf{x}(t))$ for

$$\phi_{r;i}(\mathbf{x}(t)) = \frac{1}{2n} \ln \left(4I(\mathbf{x}_r : \mathbf{x}_i) + \frac{\pi + D - 1}{n} \right) +$$
$$+ \frac{(I(\mathbf{x}(t) : \mathbf{x}_i) - I(\mathbf{x}_r : \mathbf{x}_i) - (D-1)/n)^2}{4I(\mathbf{x}_r : \mathbf{x}_i) + (D-1)/n}. \tag{11}$$

On the one hand, the number of outputs (features) in the last layer in modern DNNs is very high ($D \gg 1$) [10, 11]. On the other hand, the feature vector dimensionality D should be much less than the number of samples n to estimate the conditional densities (3) reliably ($D \ll n$). Hence, it is possible to approximately rewrite eq. (11) [21]:

$$\phi_{r;i}(\mathbf{x}(t)) \approx \frac{(I(\mathbf{x}(t) : \mathbf{x}_i) - I(\mathbf{x}_r : \mathbf{x}_i))^2}{4I(\mathbf{x}_r : \mathbf{x}_i)}. \tag{12}$$

The first multiplier in (8) cannot be computed similarly, because, if $i = r$, then $I(\mathbf{x}_r : \mathbf{x}_i) = 0$, and asymptotic distribution (9) does not hold in practice. However, in such a case this conditional density can be estimated using eq. (3). Similarly to (4), (5) one can show that

$$f(I(\mathbf{x}(t) : \mathbf{x}_r)|W_r)) \propto \exp\left(-nI(\mathbf{x}(t) : \mathbf{x}_r)\right). \tag{13}$$

By combining expressions (10), (12), (13), the marginal distribution (8) is written using softmax operation as follows

$$f(\mathbf{I}(t)|W_r) = \frac{\exp\left(-n \cdot (I(\mathbf{x}(t) : \mathbf{x}_r) + \sum\limits_{i=1}^{R} \phi_{r;i}(\mathbf{x}(t)))\right)}{\sum\limits_{j=1}^{R} \exp\left(-n \cdot (I(\mathbf{x}(t) : \mathbf{x}_j) + \sum\limits_{i=1}^{R} \phi_{j;i}(\mathbf{x}(t)))\right)}. \tag{14}$$

After that it is possible to apply some known rules [31] of combining statistical classifiers (14). For example, in case of the i.i.d. frames assumptions the criterion (7) is equivalent to rather simple video face recognition rule

$$\min_{r \in \{1,...,R\}} \sum_{t=1}^{T} \left(I(\mathbf{x}(t) : \mathbf{x}_r) + \sum_{i=1}^{R} \phi_{r;i}(\mathbf{x}(t)) \right), \tag{15}$$

which can be viewed as an extension of conventional statistical classification (6) using the new regularization term $\sum\limits_{i=1}^{R} \phi_{r;i}(\mathbf{x}(t))$. However, Kittler et al [31] argued, that the sum rule is much more robust to the errors in unknown density estimation. In our task this rule can be formulated in the MAP way:

$$\max_{r \in \{1,...,R\}} \sum_{t=1}^{T} f(W_r|\mathbf{I}(t)), \tag{16}$$

where posterior probability $f(W_r|\mathbf{I}(t))$ is defined using the Bayes rule (2) for the proposed ML of distances estimate (14).

5 Proposed algorithm

Though the previous sections contain many unrealistic assumptions about independence of sequential frames and representation of feature vectors as the estimates of the distributions of hypothetical random variables, the resulted criteria (15), (16) looks rather straightforward. In this section we demonstrate the possibility to implement this approach in more realistic scenarios. At first, it is necessary to emphasize, that the criteria discussed in the previous section contain only computations of the KL divergence between features of all frames and every instances from the database (12). Thus, it is possible to replace the KL divergence $I(\mathbf{x}(t) : \mathbf{x}_r)$ in (12), (13) with an arbitrary dissimilarity measure $\rho(\mathbf{x}(t), \mathbf{x}_r)$ between deep bottleneck features $\mathbf{x}(t)$ and $\mathbf{x}_r$. For instance, the KL minimum information discrimination criterion (6) is equivalent to the NN rule with the simple accumulation [31]:

$$\min_{r \in \{1,...,R\}} \sum_{t=1}^{T} \rho(\mathbf{x}(t), \mathbf{x}_r). \tag{17}$$

The sum rule [31] implements the MAP estimate of the the class, which corresponds to the maximal average posterior probability [26]:

$$\max_{c \in \{1,...,C\}} \sum_{t=1}^{T} \frac{\exp(-n \cdot \rho_c(\mathbf{x}(t))}{\sum\limits_{i=1}^{C} \exp(-n \cdot \rho_i(\mathbf{x}(t))}, \tag{18}$$

It is important to emphasize that our conclusion about asymptotical distribution (10) 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 [39].

Secondly, though the small sample size problem is usually appeared in practical applications of video face recognition, it is not typical to have only one instance per class in the available training set. Hence, we propose to extend criterion (15) to the case of $C \leq R$:

$$\max_{c \in \{1,...,C\}} \sum_{t=1}^{T} \frac{\exp\left(-n \cdot (\rho_c(\mathbf{x}(t)) + \frac{\lambda}{C} \sum\limits_{i=1}^{C} \frac{(\rho_i(\mathbf{x}(t)) - \rho_{c;i})^2}{\rho_{c;i}})\right)}{\sum\limits_{j=1}^{C} \exp\left(-n \cdot (\rho_j(\mathbf{x}(t)) + \frac{\lambda}{C} \sum\limits_{i=1}^{C} \frac{(\rho_i(\mathbf{x}(t)) - \rho_{j;i})^2}{\rho_{j;i}})\right)}. \tag{19}$$

Here we added the smoothing parameter $\lambda > 0$ to tune the influence of our regularization. The distance between feature vector $\mathbf{x}(t)$ and the cth identity in this criterion can be defined using the idea of the single-linkage clustering:

$$\rho_c(\mathbf{x}(t)) = \min_{r \in \{1,...,R\},c(r)=c} \rho(\mathbf{x}(t), \mathbf{x}_r). \tag{20}$$

The distance between images from cth and ith identities can be computed as the average distance between instances from these classes:

$$\rho_{c;i} = \frac{1}{R_c R_i} \sum_{r=1}^{R} \sum_{r_i=1}^{R} \delta(c - c(r))\delta(i - c(r_i))\rho(\mathbf{x}_r, \mathbf{x}_{r_i}), \tag{21}$$

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 still photos of the cth subject.

Finally, it is necessary to notice that the run-time complexity of the proposed criterion (19) $O(T(R + C^2))$ is more than C-times higher than the complexity of the baseline ML rule (17). Hence, our approach cannot be applied in practical tasks when the number of classes is rather large [21]. To speed-up the decision process, we propose to modify criterion (19) by examining only $M \ll C$ candidate classes. Namely, the sum of distances between all frames and each class are computed identically to the conventional approach (17):

$$\rho_c(\mathbf{x}(1), ..., \mathbf{x}(T)) = \frac{1}{T} \sum_{t=1}^{T} \rho_c(\mathbf{x}(t)), \tag{22}$$

and M candidate classes $\{c_1, ..., c_M\}$ with the lowest distances (18) 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 (19) is modified as follows:

$$\max_{c \in \{c_1, ..., c_M\}} \sum_{t=1}^{T} \frac{\exp\left(-n \cdot \left(\rho_c(\mathbf{x}(t)) + \frac{\lambda}{C} \sum_{i=1}^{C} \frac{(\rho_i(\mathbf{x}(t)) - \rho_{c;i})^2}{\rho_{c;i}}\right)\right)}{\sum_{m=1}^{M} \exp\left(-n \cdot \left(\rho_{c_m}(\mathbf{x}(t)) + \frac{\lambda}{C} \sum_{i=1}^{C} \frac{(\rho_i(\mathbf{x}(t)) - \rho_{c_m;i})^2}{\rho_{c_m;i}}\right)\right)}. \tag{23}$$

Our complete procedure is presented in Algorithm 1. Its runtime complexity is equal to $O(T(R + MC))$. Here the decision is made only after observation of all T frames. However, it is not difficult to implement online recognition in our approach by looking for the class label c^* right after the next video frame with the same subject is available.

6 Experimental results

In most experiments of still-to-video face recognition methods it is usually assumed that the still photos are of high quality and resolution, in frontal view, with normal lighting and neutral expression [29, 19]. However, in this paper we decided to consider the most challenging experimental setup, in which the training set contains photos gathered in uncontrolled environment [5]. Namely, we choose all photos and videos of $C = 1589$ classes as the intersection of people from the well-known LFW and YTF

Algorithm 1 Maximum likelihood of distances in still-to-video recognition

---

Require: observed sequence of T frames, database of facial feature vectors $\{\mathbf{x}_r\}$

Ensure: subject id (class label) corresponded to the objective function (23)

1: Initialize arrays of distances $\mathrm{P}[c] := 0, \rho[c][t] := 0, c \in \{1, ..., C\}, t \in \{1, ..., T\}$
2: **for** each frame $t \in \{1, ..., T\}$ **do**
3:     Detect facial region in the tth frame and extract deep features $\mathbf{x}(t)$ from bottle-neck layer
4:     **for** each class label $c \in \{1, ..., C\}$ **do**
5:         Assign $\rho[c][t] = \rho_c(\mathbf{x}(t))$ (20)
6:         Assign $\mathrm{P}[c] := \mathrm{P}[c] + \rho[c][t]$
7:     **end for**
8: **end for**
9: Obtain labels $\{c_1, ..., c_M\}$ using a selection algorithm for finding M smallest numbers in the array $\{\mathrm{P}[c] | c \in \{1, ..., C\}\}$
10: Initialize posterior probabilities $L[m] := 0, m \in \{1, ..., M\}$
11: **for** each frame $t \in \{1, ..., T\}$ **do**
12:     Assign $s := 0$, initialize posterior probabilities $L_t[m] := 0, m \in \{1, ..., M\}$
13:     **for** each top class index $m \in \{1, ..., M\}$ **do**
14:         Assign $\Omega := 0$
15:         **for** each class label $i \in \{1, ..., C\}$ **do**
16:             Assign $\Omega := \Omega + (\rho_i(\mathbf{x}(t)) - \rho_{c_j;i})^2 / \rho_{c_j;i}$
17:         **end for**
18:         Assign $L_t[m] := \exp(-n \cdot (\rho[c_j][t] + \lambda\Omega/C))$
19:         Assign $s := s + L_t[m]$
20:     **end for**
21:     **for** each top class index $m \in \{1, ..., M\}$ **do**
22:         Assign $L[m] := L[m] + L_t[m]/s$
23:     **end for**
24: **end for**
25: **return** class label c_{m^*}, where $m^* = \underset{m \in \{1,...,M\}}{\mathrm{argmax}}\ L[m]$

---

datasets. It is important to mention that though the YTF dataset contains the videos 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 [5]. Our testing set contains 3353 videoclips from the YTF [25] (in average, 183 frames per person). 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.

The Caffe framework is applied to extract deep bottleneck features using two publicly available CNN models for unconstrained face recognition, namely, the well-known VGG-13 Network (VGGNet) [10], and the Lightened CNN [11]. The fc8 layer of the VGGNet extracts $D = 4096$ non-negative features from 224x224 RGB image. The Lightened CNN (version C) extracts $D = 256$ (possibly negative) features from 128x128 grayscale image. The outputs of these CNNs were L_2 normalized to form the final feature vectors, which are matched using Euclidean distance. The features in the VGGNet are positive, hence we also perform L_1 normalization to treat them as the probability distributions and match them using the KL divergence.

The proposed Algorithm 1 is implemented in a stand-alone C++ application with Qt 5 framework. In the first experiments we examined the dependence of the top-1 accuracy on the most important parameters of our algorithm, namely, the importance λ of regularization term $\sum_{t=1}^{T} \sum_{i=1}^{C} (\rho_i(\mathbf{x}(t)) - \rho_{c;i})^2 / \rho_{c;i}$ (23), and the number of candidate classes M. The results are presented in Fig. 1 and Fig. 2, respectively.

Here, firstly, error rates drastically depend on the deep CNN used for feature extraction. The accuracy of the Lightened CNN [11] is approximately 20% higher than the accuracy of much widely used VGGNet [10]. It is worth noting that treating the features from the latter network as probability distributions and matching them with the KL divergence (4) allows decreasing error rate at 1.4% in the best case. Secondly, the U-curves in Fig. 1 prove that the proper choice of the parameter λ can significantly influence the recognition accuracy. It is especially true for the VGGNet features, in which the difference in error rates achieves 2.5%. However, the optimal value of this parameter is practically identical ($\lambda = 6..8$) for all three combination of the CNN and dissimilarity measure. At the same time, the curves in Fig. 2 reaches stability very fast with an increase of the number of candidate classes M. For example, the accuracy for the VGGNet features does not increase further if $M \geq 16$. Though the examination of all $M = C$ classes causes the lowest error rates, this value is not recommended for practical usage due to the very low recognition speed. Hence, in the further part of this paper we use near-optimal values of these parameters: $\lambda = 7$ and $M = 64$.

In the next experiments 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 frame of all testing videoclips, where $X_{max} \geq 0$ determines the noise level. Original videos from the YTF dataset are recognized, when $X_{max} = 0$. The proposed approach (23) is compared with conventional video face recognition techniques, namely, 1) the ML rule (17) applied to all frames [26, 23]; 2) selection of one representative frame with the k-medoids clustering method ($k = 1$) and the NN matching (17) of this frame [35, 36] (hereinafter "ML with clustering");
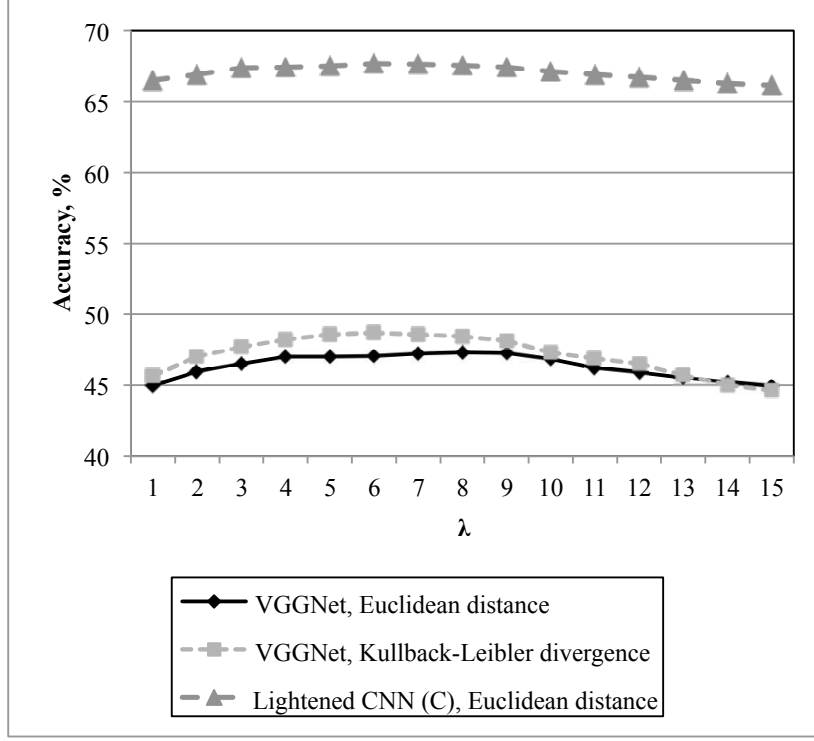
Figure 1: Dependence of the recognition accuracy (%) on the importance λ of regularization term (23), $M = C$.

and 3) estimation of the MAP class for each frame (18) (hereinafter "MAP"). The parameter n of the proposed criterion and MAP rule is tuned in order to provide the highest validation accuracy. The main results of these experiments are shown in Fig. 3, Fig. 4 and Fig. 5 for VGGNet features matched with the Euclidean distance and the KL divergence, together with the Lightened CNN features, respectively.

Based on these results one can draw the following conclusions. Firstly, 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. Secondly, though the KL divergence outperforms the conventional Euclidean distance for original YTF dataset ($X_{\max} = 0$), the latter dissimilarity 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 [6, 7]. Nevertheless, the Lightened CNN again significantly outperforms the VGGNet even for very high noise levels. Thirdly, though the rather popular accumulation of all frames in one centroid (video face representation [35, 36, 20]) in the "ML with clustering" method makes it possible to increase performance in T times, its error rate is 2-5% and 2-6% higher than the error rate of the baseline ML (17) and the aver-
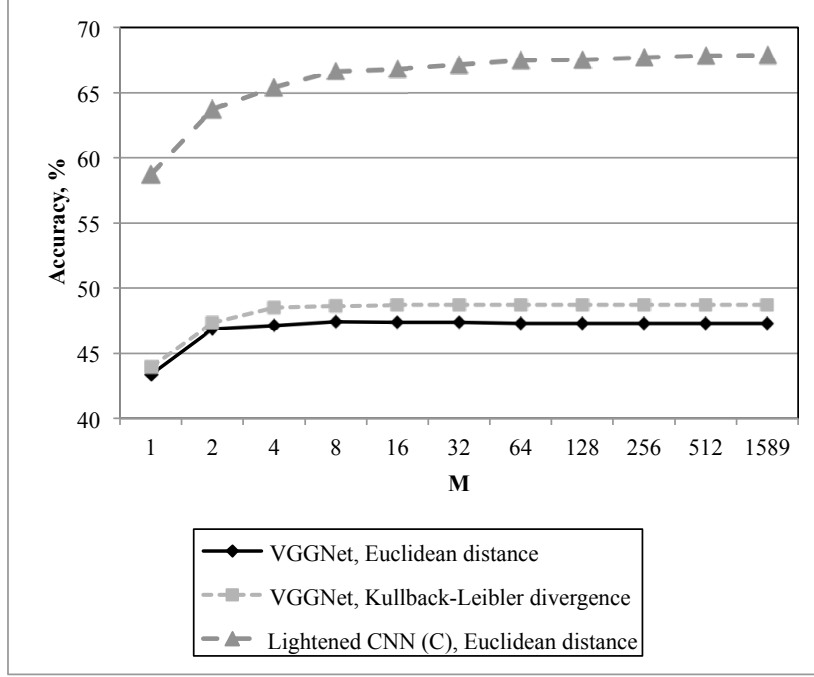
12

Figure 2: Dependence of the recognition accuracy (%) on the number of candidate classes M, $\lambda = 7$.

age posterior probability pooling (18) [31, 26]. Moreover, if this approach is applied in online video recognition right after new tth frame is observed, the gain in performance over other methods is not so obvious. Really, it is easy to implement criteria (17), (23) in adaptive mode [26] so that the processing of the tth frame is aggregated with the results obtained for recognition of previous $(t-1)$ frames. It is not surprising that the frame selection methods [17] have more potential in video-to-video recognition or video verification tasks [20], in which the matching of all pairs of frames from two video clips is very expensive. Finally, the most important conclusion here is the highest recognition accuracy of the proposed algorithm 1 in all cases. Our approach is 3-4% more accurate when compared to other methods. Moreover, though our idea is based on looking for correspondence of the distances $\rho_i(\mathbf{x}(t))$ (17) and $\rho_{c;i}$ (18), (19), the gain in error rate does not significantly degrade even with the presence of noise in the testing videos. 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 part of this section we provide preliminary results for IJB-A dataset [24]. The IJB-A 1:N protocol is primarily used for testing of the video-to-video recognition. Moreover, each split is used to recognize only 33% subjects, the rest part can be used
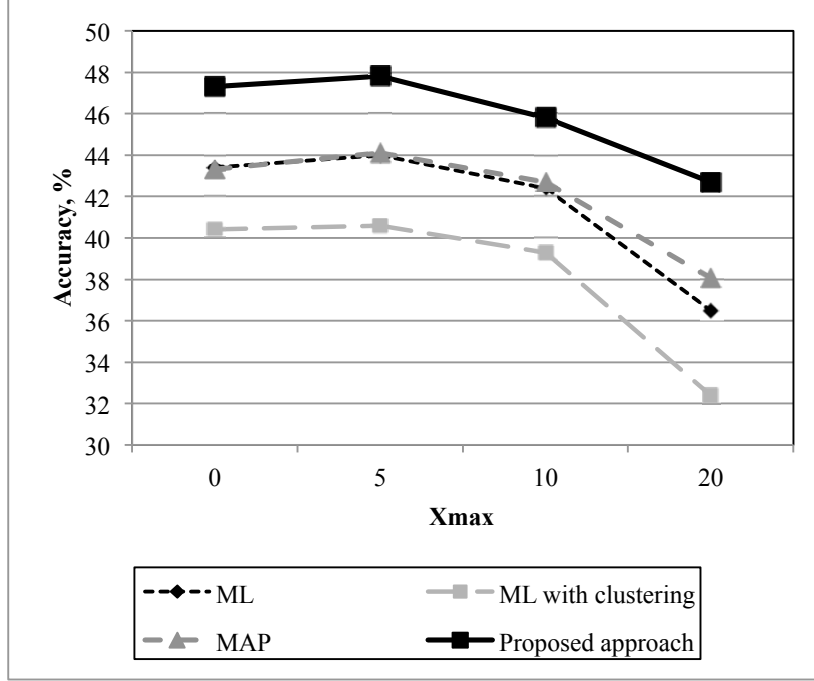
Figure 3: Dependence of the recognition accuracy (%) on the noise level X_{max}, VG-GNet, Euclidean distance.

to train an algorithm under the strict condition that no such imagery contain the same subjects that are in the test split. Hence, we decided to implement more difficult scenario of the still-to-video recognition with all $C = 500$ subjects. We put all $R = 5712$ still photos of these subjects to the training set. The testing set contains all 2043 videos from this dataset with approximately 8 frames per video. The average accuracies for all methods described above are shown in Table 1.

In contrast to the previous experiments, here the features from the VGGNet [10] allows classifying videos 25% more accurately when compared to the Lightened CNN [11], which can be probably caused by the presence of many non-frontal photos in this dataset. However, all the previous conclusions about the still-to-video recognition methods remain the same. The selection of the most representative frame in the ML with clustering is 0.6-0.9% less accurate that the matching of all frames in criterion (17). The pooling of posterior probabilities (18) here is much preferable than the ML rule (17). At the same time, the error rate of the proposed algorithm 1 is 4.5-5.5% and 1.1-2.5% lower, than the error rates of the widely-used criteria (17) and (18), respectively.
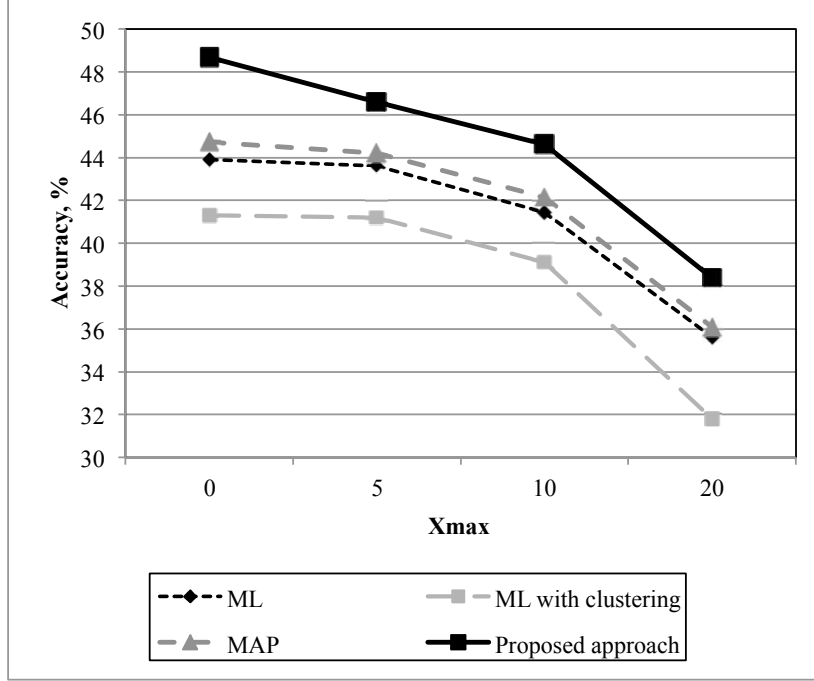
Figure 4: Dependence of the recognition accuracy (%) on the noise level $X_{\max}$, VG-GNet, Kullback-Leibler divergence.

7   Conclusion

In this paper we proposed the novel statistical approach to still-to-video 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 introduction of a special regularization term (23) in the nearest neighbor matching (17) of high-dimensional features from the outputs of the deep CNNs. In this regularization it is assumed that the resulted class $c^*(t)$ for recognition of individual tth video frame is reliable only if the distances between the features of this frame and the rth still image are approximately equal to the distances between the reference images from the $c^*(t)$ and $c(r)$ classes for all $r \in \{1, ..., R\}$. This assumption is known to be asymptotically correct for the KL divergence and rather simple probabilistic model from Section 3 [22]. However, it was experimentally demonstrated that the proposed approach can be combined even with the state-of-the-art Euclidean distance (Fig. 4). Moreover, our algorithm allows to increase accuracy even for general feature vectors with possible negative values (Fig. 5). We demonstrated how to tune the parameters of the proposed algorithm (Fig. 2) in order to drastically reduce its computation complexity by proper choice of M candidate classes (compare the criterion (23) with the original one (19)).

The main direction for further research of the proposed algorithm is its applica-
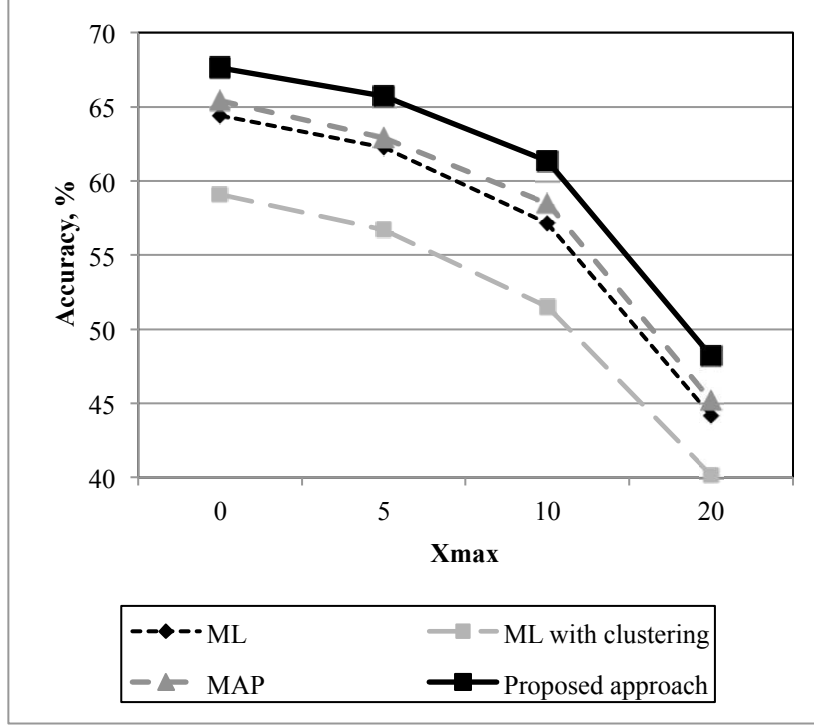
Figure 5: Dependence of the recognition accuracy (%) on the noise level $X_{\max}$, Lightened CNN (version C), Euclidean distance.

tions with more accurate approximation of the distance probability distributions, e.g., the usage of the more appropriate Weibull distribution [39]. Secondly, it is important to make our regularization smoother by taking into account the temporal coherence of sequential frames [28]. Thirdly, it is necessary to combine our approach with the modern techniques with frames weighting in the aggregation module (22) [20] to make distinction between frames with different quality [17]. Finally, though the deep bottleneck 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 [33] or geometry aware feature matching [40] can also be discussed in the future experiments.

8 Acknowledgement

Method	VGGNet, Euclidean distance	VGGNet, KL divergence	Lightened CNN, Euclidean distance
ML	68.3	67.5	42.2
ML with clustering	67.4	66.9	41.7
MAP	71.7	69.2	46.5
Proposed approach	72.8	71.6	47.8

Table 1: Recognition accuracy (%) for IJB-A dataset.

References

References

[1] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (7553) (2015) 436–444.

[2] S. J. Prince, Computer vision: models, learning, and inference, Cambridge University Press, 2012.

[3] Y. Taigman, M. Yang, M. Ranzato, L. Wolf, Deepface: Closing the gap to human-level performance in face verification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1701–1708.

[4] J. Liu, Y. Deng, T. Bai, Z. Wei, C. Huang, Targeting ultimate accuracy: Face recognition via deep embedding, arXiv preprint arXiv:1506.07310.

[5] E. Learned-Miller, G. B. Huang, A. RoyChowdhury, H. Li, G. Hua, Labeled faces in the wild: A survey, in: Advances in Face Detection and Facial Image Analysis, Springer, 2016, pp. 189–248.

[6] A. V. Savchenko, Search techniques in intelligent classification systems, Springer, 2016.

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

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

[9] X. Cao, D. Wipf, F. Wen, G. Duan, J. Sun, A practical transfer learning algorithm for face verification, in: Proceedings of the IEEE International Conference on Computer Vision, 2013, pp. 3208–3215.

[10] O. M. Parkhi, A. Vedaldi, A. Zisserman, Deep face recognition, in: BMVC, Vol. 1, 2015, p. 6.

[11] X. Wu, R. He, Z. Sun, T. Tan, A light cnn for deep face representation with noisy labels, arXiv preprint arXiv:1511.02683.

[12] A. V. Savchenko, Deep convolutional neural networks and maximum-likelihood principle in approximate nearest neighbor search, in: 8th Iberian Conference on Pattern Recognition and Image Analysis (IbPRIA 2017), Springer, 2017, pp. 42–49.

[13] Z. Cui, W. Li, D. Xu, S. Shan, X. Chen, Fusing robust face region descriptors via multiple metric learning for face recognition in the wild, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 3554–3561.

[14] H. Cevikalp, B. Triggs, Face recognition based on image sets, in: Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), 2010, pp. 2567–2573.

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

[16] L. Wolf, A. Shashua, Kernel principal angles for classification machines with applications to image sequence interpretation, in: Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), Vol. 1, 2003.

[17] Z. Huang, X. Zhao, S. Shan, R. Wang, X. Chen, Coupling alignments with recognition for still-to-video face recognition, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 3296–3303.

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

[19] Y. Zhu, Z. Zheng, Y. Li, G. Mu, S. Shan, G. Guo, Still to video face recognition using a heterogeneous matching approach, in: Proceedings of the IEEE International Conference on Biometrics Theory, Applications and Systems (BTAS), 2015, pp. 1–6.

[20] J. Yang, P. Ren, D. Chen, F. Wen, H. Li, G. Hua, Neural aggregation network for video face recognition, arXiv preprint arXiv:1603.05474 (accepted for CVPR17).

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

[22] S. Kullback, Information theory and statistics, Courier Corporation, 1997.

[23] G. Shakhnarovich, J. W. Fisher, T. Darrell, Face recognition from long-term observations, in: European Conference on Computer Vision, Springer, 2002, pp. 851–865.

[24] B. F. Klare, B. Klein, E. Taborsky, A. Blanton, J. Cheney, K. Allen, P. Grother, A. Mah, A. K. Jain, Pushing the frontiers of unconstrained face detection and recognition: Iarpa janus benchmark a, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1931–1939.

[25] L. Wolf, T. Hassner, I. Maoz, Face recognition in unconstrained videos with matched background similarity, in: Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), 2011, pp. 529–534.

[26] A. V. Savchenko, Adaptive video image recognition system using a committee machine, Optical Memory and Neural Networks 21 (4) (2012) 219–226.

[27] Y. Zhang, A. M. Martínez, A weighted probabilistic approach to face recognition from multiple images and video sequences, Image and Vision Computing 24 (6) (2006) 626–638.

[28] X. Liu, T. Chen, Video-based face recognition using adaptive hidden markov models, in: Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), Vol. 1, 2003.

[29] Z. Huang, S. Shan, H. Zhang, S. Lao, A. Kuerban, X. Chen, Benchmarking still-to-video face recognition via partial and local linear discriminant analysis on cox-s2v dataset, in: Asian Conference on Computer Vision, Springer, 2012, pp. 589–600.

[30] S. Zhou, V. Krueger, R. Chellappa, Face recognition from video: A condensation approach, in: Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition, 2002, pp. 221–226.

[31] J. Kittler, M. Hatef, R. P. Duin, J. Matas, On combining classifiers, IEEE transactions on pattern analysis and machine intelligence 20 (3) (1998) 226–239.

[32] P. Zhu, L. Zhang, W. Zuo, D. Zhang, From point to set: Extend the learning of distance metrics, in: Proceedings of the IEEE International Conference on Computer Vision, 2013, pp. 2664–2671.

[33] Z. Huang, R. Wang, S. Shan, X. Chen, Learning euclidean-to-riemannian metric for point-to-set classification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1677–1684.

[34] F. Schroff, D. Kalenichenko, J. Philbin, Facenet: A unified embedding for face recognition and clustering, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 815–823.

[35] J.-C. Chen, R. Ranjan, A. Kumar, C.-H. Chen, V. M. Patel, R. Chellappa, An end-to-end system for unconstrained face verification with deep convolutional neural networks, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV) Workshops, 2015, pp. 118–126.

[36] H. Li, G. Hua, X. Shen, Z. Lin, J. Brandt, Eigen-pep for video face recognition, in: Proceedings of the Asian Conference on Computer Vision, Springer, 2014, pp. 17–33.

[37] A. Canziani, E. Culurciello, Cortexnet: a generic network family for robust visual temporal representations, arXiv preprint arXiv:1706.02735.

[38] Y. Wong, S. Chen, S. Mau, C. Sanderson, B. C. Lovell, Patch-based probabilistic image quality assessment for face selection and improved video-based face recognition, in: Proceedings of the IEEE Computer Vision and Pattern Recognition Workshops (CVPRW), 2011, pp. 74–81.

[39] G. Burghouts, A. Smeulders, J.-M. Geusebroek, The distribution family of similarity distances, in: Proceedings of the International Conference on Advances in neural information processing systems (NIPS), 2008, pp. 201–208.

[40] M. T. Harandi, M. Salzmann, R. Hartley, From manifold to manifold: Geometry-aware dimensionality reduction for spd matrices, in: European conference on computer vision, Springer, 2014, pp. 17–32.