Single Sample Face Recognition from Video via Stacked Supervised Auto-encoder

Pedro J. Soto Vega*, Raul Queiroz Feitosa*†, Victor H. Ayma Quirita*, Patrick Nigri Happ*
*Pontifical Catholic University of Rio de Janeiro, Brazil
†Rio de Janeiro State University, Brazil
{psoto, raul, vhaymaq, patrick}@ele.puc-rio.br

***Abstract*—This work proposes and evaluates strategies based on Stacked Supervised Auto-Encoders (SSAE) for face representation in video surveillance applications. The study focuses on the identification task with a single sample per person (SSPP) in the gallery. Variations in terms of pose, facial expression, illumination and occlusion are approached in two ways. First, the SSAE extracts features from face images, which are robust to such variations. Second, we propose methods to exploit the multiple samples per persons probes (MSPPP) that can be extracted from video sequences. Three variants of the proposed method are compared upon HONDA/UCSD and VIDTIMIT video datasets. The experimental results demonstrate that strategies combining SSAE and MSPPP are able to outperform other SSPP methods, such a local binary patterns, in face recognition from video.**

***Keywords*-Auto-encoder; Face Recognition; Surveillance.**

I. Introduction

Face recognition with a Single Sample per Person (SSPP) [1], [2] in the gallery is the most common scenario for facial recognition systems. The gallery is typically a set of good quality frontal images in neutral expressions. In contrast, probe images vary in a number of ways (pose, expression, occlusion, illumination, etc.). This is particularly challenging for face recognition from video. In surveillance applications for instance, face images are captured under uncontrolled conditions, mostly without the cooperation of the person being identified. In consequence, probe images are often affected by occlusions and variations in pose, illumination and facial expression combined with low video quality.

To face this challenge several mathematical models to represent compactly facial images under adverse conditions have been proposed. The commonly used subspace analysis methods for face representation, like Eigenfaces [3] and Fisherfaces [4], are not suitable for SSPP applications since they require a large number of training samples. However, the usage of manually generated virtual faces [5] or external datasets [1] makes possible the extension of the traditional face representation methods to the SSPP scenario. The performance of the face recognition systems based on these methods is still unsatisfactory for real data.

Recognition systems based on local feature representations like Local Binary Pattern (LBP) [6] and Local Phase Quantization (LPQ) [7] may achieve good results under certain adverse conditions like illumination variations. However, their performances decrease under occlusion and strong pose and facial expressions variations.

In the last years, Deep Neural Networks have been successfully applied for image representation [8] [9] [10] [11] [12]. Recently denoising auto-encoders [11] have been used to build this kind of neural networks.

Motivated by the success of denoising auto-encoders, Gao and co-workers. [13] introduced a novel solution for SSPP face recognition. They proposed the Stacked Supervised Auto-Encoder (SSAE) that treats facial images with pose, facial expressions, occlusions and lighting variations as noisy data, whereas the clean data is taken from frontal, occlusion free images, in neutral expression with uniform illumination.

The SSAE aims to produce face image descriptors that are robust against the aforementioned variations. The reported results indicate that this approach outperforms alternative image representation techniques for static images datasets with manually annotated landmarks and face alignment. However, if an automatic face detector, like Viola & Jones [14], is applied, the performance might decrease considerably [13].

In this paper we propose and evaluate SSAE based strategies for face recognition from video sequences. Usually, surveillance systems are able to track a face through the video. The set of facial image instances collected this way forms what we call multiple samples per person probes (MSPPP). In this study we investigate how MSPPP can be exploited in order to offset the negative effects of the low quality video images in surveillance applications.

The objective of this paper is threefold. First we evaluate SSAE upon still images collected from video sequences. Second, we evaluate SSAE for face images detected and annotated automatically, instead of manually as in [13]. Third, we propose and evaluate three strategies to exploit MSPPP for surveillance applications. We further report experiments carried out on two public video datasets to demonstrate that MSPPP can compensate for the deleterious effect of low quality video images in surveillance applications.

The rest of this paper is organized as follows. An overview of related works in the SSPP scenario is presented in Section I-A. Section II describes succinctly some techniques underlying our proposal. A detailed description of the proposed extensions of SSAE for video based face recognition is the subject of Section III. The experimental analysis, is reported

in Section IV. Section V summarizes the main conclusions and indicates future directions.

A. Related work

Face representation is an important issue in face recognition. A good face representation must preserve relevant information of facial images and be robust to changes in appearance.

Methods based on subspace analysis, like Eigenfaces [3] and 2DPCA [15], are well-known by their effective and efficient face representation. However, they are not suited to SSPP face representation due to the limited number of available training images to estimate the projection matrix.

In order to make such approaches more adequate for SSPP, Projection-Combined PCA [16] and Enhanced Projection-Combined PCA [17] have been proposed, taking advantage of some synthetic faces generated by different methods. Artificial images can be created by adding noise to clean images samples or by applying different geometrical transformations on the whole image or on its patches [4] [5] [18] [19] [20]. All these approaches slightly improve the original methods they derive from, but the gain is still modest.

Hand-crafted face descriptors like Local Binary Pattern (LBP) [6] get a feature descriptor that allows for good results under certain adverse conditions like illumination variations, but their performance decreases substantially for non-frontal images and under occlusion. Recently, with the emergence of the deep learning, a new approach called Deep Lambertian Networks, proposed by Tang et al. [21], presented a good performance in SSPP setting, by extracting illumination invariant features. However, the method does not handle other variations like expressions, pose and occlusions.

Continuing on the deep learning research, the auto-encoders [22] became a frequent used building block in deep neural networks and a range of different ways of employing them for image representation have emerged in the recent years. Vincent et al. [11] proposed the denoising auto-encoder, which enhances the original auto-encoder concept by training the network with manually corrupted inputs. Rafiai and co-authors [23] enhanced the auto-encoder robustness to noise using different loss functions [24]. Zou and co-workers [25] used the pooling operation after the reconstruction ICA [26] encoding process and enforced the pooled features to be similar for instances with the same class label. Although these works have brought some improvements, they still fail to handle all variations of face images, especially in surveillance applications.

This scenario motivated Gao and co-workers [13] to propose a supervised auto-encoder that treats images with pose, facial expressions, and lighting variations or occlusion as noisy data. As clean data (gallery images) a frontal single image with homogeneous lighting conditions, no occlusions and neutral facial expression is taken. For static images, where the faces were manually cropped and aligned, the method achieved high recognition rates, but the performance decreased for faces detected automatically. This is a major hindrance for recognition from video sequences, whose images typically present low quality.

In this context, many techniques have been proposed in order to take advantage of multiple image instances of the same person in videos. Hayat et al. [27] perform a majority voting scheme considering all the frames from a video. With the same purpose of exploiting MSPPP Xiaoming et al. [28] propose a method based on Hidden Markov Models (HMM), whereas Kuang-Chi et al. [29] rely on a Bayesian approach.

Driven by the relative success of the aforesaid proposals in the following we describe and evaluate SSAE based strategies to exploit MSPPPs available in surveillance applications.

II. Fundamentals

A. Auto-encoders

Basically, an auto-encoder is an unsupervised neural network that creates a compact data representation from which the original data can be accurately reconstructed. It usually has two parts: an encoder and a decoder [30], often implemented by a single hidden layer network.

The encoder, denoted as f, maps the input data $x \in \mathbb{R}^d$,to a compact representation $z \in \mathbb{R}^r$ through the activations of the r neurons in the hidden layer, whereby $r < d$. The function f has the form:

$$z = f(x) = s(Wx + b) \tag{1}$$

where $W \in \mathbb{R}^{rxd}$ is the matrix containing the learned coefficients of the non-linear transformation, $b \in \mathbb{R}^r$ denotes the bias and $s(\cdot)$ is the so-called 'element-wise activation function', which is usually non-linear functions, such as the sigmoid or the hyperbolic tangent.

The decoder, denoted as g, aims at mapping the representation z back to the input x, formally:

$$\hat{x} = g(z) = s(\hat{W}z + \hat{b}) \tag{2}$$

with $\hat{W} \in \mathbb{R}^{dxr}$ being the matrix of non-linear transformation coefficients and $\hat{b} \in \mathbb{R}^d$ the reconstruction bias.

The parameters $W, b, \hat{W}$ and $\hat{b}$ are determined by minimizing the loss function:

$$\left[W^*, b^*, \hat{W^*}, \hat{b^*}\right] = \min_{W,b,\hat{W},\hat{b}} \sum_{i=1}^{N} \|x_i - g(f(x_i))\|_2^2 \tag{3}$$

where x_i corresponds to the i^{th} out of N training samples. Equation 3 can be solved by gradient descent methods.

B. Face Representation using denoising auto-encoder

The so called denoising auto-encoder is a variant of auto-encoders with the ability to learn the distributions of the noise present in the input data, and to reproduce its corresponding uncorrupted version [11].The training set for a denoising auto-encoder is built by adding noise to a clean data. The auto-encoder is trained so as to produce for a noisy input data an output as close as possible to the corresponding noise free data. Denoting the clean data as x and the corrupted data as $\tilde{x}$ the eq. 3 can be rewritten as follows:

$$\left[W^*, b^*, \hat{W^*}, \hat{b^*}\right] = \min_{W,b,\hat{W},\hat{b}} \sum_{i=1}^{N} \|x_i - g(f(\tilde{x}_i))\|_2^2 \quad (4)$$

C. Face Representation using Stacked Supervised Auto-encoders

By and large, gallery image samples of real applications consist of frontal faces with uniform illumination and neutral expression with no occlusion. In contrast, these properties generally do not hold for probe images. Building upon the auto-encoders rationale Gao and co-authors proposed the Stacked Supervised Auto-encoders (SSAE) [13] for face representation. It treats a gallery image of a person as clean data and all its variants in terms of pose, illumination, expression, occlusion, etc., as corrupted data. Like the denoising auto-encoder, SSAE is trained so as to produce for any image from a person its clean counterpart.

Denoting each corrupted image as $\tilde{x}_i$ and its clean version as $x_i (i = 1, ...N)$, the network is trained to reproduce at the output an image $\hat{x}_i$ as close as possible to its clean version. Then the loss function can be defined as:

$$\begin{aligned}\left[W^*, b^*, \hat{W^*}, \hat{b^*}\right] = \min_{W,b,\hat{W},\hat{b}} \frac{1}{N} \sum_{i=1}^{N} (\|x_i - g(f(\tilde{x}_i))\|_2^2 + \\ \lambda \|f(x_i) - f(\tilde{x}_i)\|_2^2) + \dots \\ \dots + \lambda_{wd} J_{wd} + \lambda_{sp} J_{sp}\end{aligned} \quad (5)$$

The first term on the right hand side of eq. 5 is the reconstruction error. The second term is the similarity preservation term, which aims at forcing face representations of the same person, $f(x_i)$ and $f(\tilde{x}_i)$, to be similar to each other. The relative importance of these terms is tuned by λ. The third term where J_{wd} is the weight decay penalty term that favors small weights values, whereby given by λ_{wd} is a regularization coefficient and J_{wd} is the sum of the squared Frobenius norm of both weight matrices [27]:

$$J_{wd} = \|W\|_F^2 + \|\hat{W}\|_F^2 \quad (6)$$

It is usual to tie both weight matrices by imposing $\hat{W} = W^T$.

The fourth term is the sparsity constraint defined by the product of the coefficient λ_{sp} and the Kullback-Leiber divergence (KL divergence) [13] given by:

$$J_{sp} = KL(\rho_x || \rho_0) + (\rho_{\tilde{x}} || \rho_0) \quad (7)$$

where:

$$\rho_x = \frac{1}{N} \sum_i f(x_i)$$

$$\rho_{\tilde{x}} = \frac{1}{N} \sum_i f(\tilde{x}_i)$$

$$KL(\rho || \rho_0) = \sum_j \left(\rho_0 \log\left(\frac{\rho_0}{\rho_j}\right) + (1 - \rho_0) \log\left(\frac{1 - \rho_0}{1 - \rho_j}\right) \right) \quad (8)$$

In order to build a deep neural network, Gao and co-authors [13] proposed to replicate this scheme in multiple layers. Training is carried out in a layer wise manner. Once a network has been trained according to eq. 5, the feature it produces for the same training data serve as clean and corrupted data to train the next layer. This is repeated until the final layer has been trained. The authors further determined empirically that no substantial performance gain is attained by more than two layers.

Fig. 1 helps understanding the SSAE architecture, the training and feature extraction procedures. The images on the left represent the training images. The gallery sample (clean data), on the top-left, is frontal, occlusion free, well illuminated and has neutral expression. The test samples (corrupted data) might contain occlusions, lighting variations and non-neutral facial expressions. A compact representation of the training process is exhibited in the middle of Fig. 1. It is done in two sequential steps, as mentioned before. The symbols f_1 and f_2 refer to the encoder functions of the first and second layer, respectively. Similarly, g_1 and g_2 refer to the decoder functions of the first and second layer.

It should be noted that eq. 5 includes the weight decay following [27], thought it was not present in the original SSAE proposal [13].

A detailed description of the training procedure is given in Algorithm 1. The training set is divided into clean and corrupted face images, each labeled according to the person represented in the image. The loss function in step 2 is essentially the same given in eq. 5. It merely allows for multiple corrupted versions of the same clean image.

Once the network has been trained, it can be used to compute the features, or representation, of any test image, as depicted on the right side of Fig. 1.

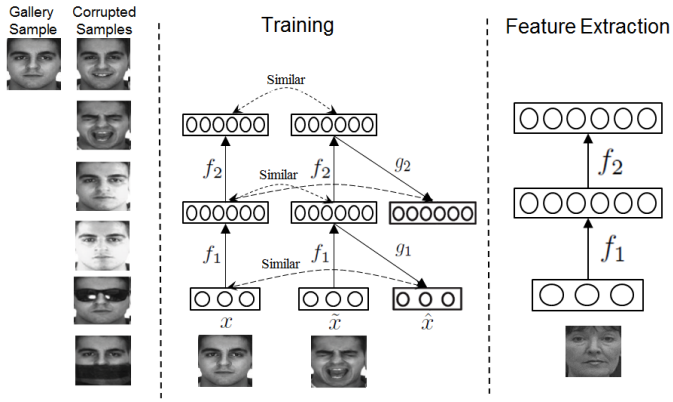


Fig. 1. The SSAE Achitecture

III. Video Based Stacked Supervised Auto-encoder

Face images collected from video often present low quality, which conspires against the recognition. In addition, still due

Algorithm 1 Learning SSAE

Input: Gallery images $X_{train} = \{x_{train}^{(i)}\}$ for $i = 1, ..., G_{train}$
Corrupted images $\tilde{X}_{train} = \{\tilde{x}_{train}^{(j)}\}$ for $j = 1, ..., N$
Corrupted image labels $L_{train} = \{l_{train}^{(j)}\}$ for $j = 1, ..., N$ where $l_{train}^{(j)}$ may take values in the set $\{1, ..., G_{train}\}$
Number of hidden layer H
Reconstruction/Preservation coefficient λ
Regularization coefficient λ_{wd}
Initial weight matrices $W^{(h)}, \hat{W}^{(h)}$ for $h = 1, ..., H$
Initial biases $b^{(h)}, \hat{b}^{(h)}$ for $h = 1, ..., H$

For $h = 1, ..., H$ **do**
  /train the network
  $\left[W^*, b^*, \hat{W^*}, \hat{b^*}\right] \leftarrow \min_{W,b,\hat{W},\hat{b}} \Psi$

$$where: \Psi = \frac{1}{N}\sum_{i=1}^{G}\sum_{j|l^{(j)}=i}\left(\left\|x_{train}^{(i)} - g(f(\tilde{x}_{train}^{(j)}))\right\|_2^2 + \lambda||f(x_{train}^{(i)}) - f(\tilde{x}_{train}^{(j)})||_2^2\right) + \lambda_{wd}J_{wd} + \lambda_{sp}J_{sp}$$

  /save the weights and bias of current layer
  $\left\{W^{(h)}, b^{(h)}\right\} \leftarrow \{W^*, b^*\}$
  If $h < H$ **do**
    /compute the representations at the current layer
    For $i = 1, ..., G_{train}$ **do**
      $x_{train}^{(i)} \leftarrow f(x_{train}^{(i)})$
    end For
    For $i = 1, ..., N$ **do**
      $\tilde{x}_{train}^{(i)} \leftarrow f(\tilde{x}_{train}^{(i)})$
    end For
  end If
end For

Output:SSAE Model $\theta = \left\{W^{(1)}, b^{(1)}, ..., W^{(H)}, b^{(H)}\right\}$

to the low image quality, detection algorithms, such as [14], frequently fail to locate face fiducial points accurately bringing about badly framed face images and degrading the recognition rates even further.

This work investigates if these problems can be mitigated by exploiting multiple image samples that can be extracted from a person being tracked on a video sequence.

The recognition procedure is described in details in Algorithm 2. Initially, the SSAE model comprising the weight matrices and bias vectors estimated in the training phase is loaded.

For a two layers SSAE the representation $\alpha_{gall}^{(i)}$ of a gallery face image $x_{gall}^{(i)}$,is computed by

$$\alpha_{gall}^{(i)} \leftarrow s(W^{(2)}s(W^{(1)}x_{gall}^{(i)} + b^{(1)}) + b^{(2)}) \quad (9)$$

for $i = 1, ..., G_{gall}$,where G_{gall} is the number of subjects in the gallery.

The first time a person is detected in a video sequence, a set $\delta_{probe}^{(i)}$ is created for each gallery entry (i) These sets will accumulate the dissimilarities among that entry and the upcoming face image probes.

The recognition itself can be executed at each new video frame or only after the entire video sequence containing the tracked person have been collected. In any case, if $\tilde{x}_{probe}$ is a probe face captured at any frame, its representation α_{probe} can be calculated by the same function on the right hand side of eq. 9, specifically

$$\alpha_{probe} \leftarrow s(W^{(2)}s(W^{(1)}\tilde{x}_{probe} + b^{(1)}) + b^{(2)}) \quad (10)$$

The *dissimilarity* between the probe image and each gallery entry is calculated by a suitable function $D(\cdot)$, whose arguments are the representations of the probe (α_{probe}) and the gallery $\alpha_{gall}^{(i)}$ images. The values produced this way are accumulated in $\delta_{probe}^{(i)}$.

Based on the dissimilarity values accumulated along the video sequence up to that point a function $\phi\left(\left\{\delta_{probe}^{(i)}\right\}\right)$ determines the identity L_{probe} of the person in the scene. In this work we consider three formulations for $\phi(\cdot)$:

Majority voting

For each frame the most similar gallery entry is determined. The final identification falls upon the subject, who has been considered the most similar one in most of the frames upto the current frame.

Best score

For each gallery entry the lowest dissimilarity value is determined. The identity is assigned to the gallery entry having the lowest among the lowest dissimilarities.

Median score

For each entry, the median of the dissimilarity values is computed. The probe is assigned to the gallery entry with the lowest median dissimilarity.

IV. Experiments

Two series of experiments were conducted in order to assess the performance of the SSAE method on video sequences. We took as baseline the method based on Local Binary Patterns - LBP [6], which is a well-known model for SSPP scenarios. In the following we describe the datasets used in the analysis, the experimental protocol and finally the results.

A. Datasets

The experiments were executed using two video datasets (Honda/UCSD [29] and VIDTIMIT [31]) and two static images databases (CMU-PIE [32] and Extended Yale B [33].

The Honda/UCSD[1] dataset [29] contains 59 videos from 19 subjects. The number of frames per video sequence varies from

[1]http://vision.ucsd.edu/∼leekc/HondaUCSDVideoDatabase/HondaUCSD.html

Algorithm 2 Face Recognition from Video (SSAE+T)

Input:Gallery images set $X_{gall} = \left\{x_{gall}^{(i)}\right\}$, for $i = 1, ..., G_{gall}$
SSAE Model $\theta = \left\{W^{(1)}, b^{(1)}, W^{(2)}, b^{(2)}\right\}$

For i **from** 1 **to** G_{gall} **do**
  /compute the SSAE representation
  $\alpha_{test}^{(i)} \leftarrow s(W^{(2)}s(W^{(1)}x_{gall}^{(i)} + b^{(1)}) + b^{(2)})$
  /initialize dissimilarity sets
  $\delta_{probe}^{(i)} = \emptyset$
end For
For each new probe image $\tilde{x}_{probe}$ **do**
  /compute the SSAE representation
  $\alpha_{probe} \leftarrow s(W^{(2)}s(W^{(1)}\tilde{x}_{probe} + b^{(1)}) + b^{(2)})$
  For i **from** 1 **to** G_{gall} **do**
    /compute the dissimilarity of the $i-th$ gallery entry
    $\delta_{probe}^{(i)} \leftarrow \delta_{probe}^{(i)} \cup D(\alpha_{probe}, \alpha_{gall}^{(i)})$
  end For
  /select the identity
  $L_{probe} = \phi\left(\left\{\delta_{probe}^{(i)}\right\}\right)$
  Output:L_{probe}
end For

92 to 645 with a resolution of 640x480 pixels. The number of face images per subject varies from 92 to 1,149. All videos contain significant head rotations, large scale changes, partial occlusions and considerable lighting variations.

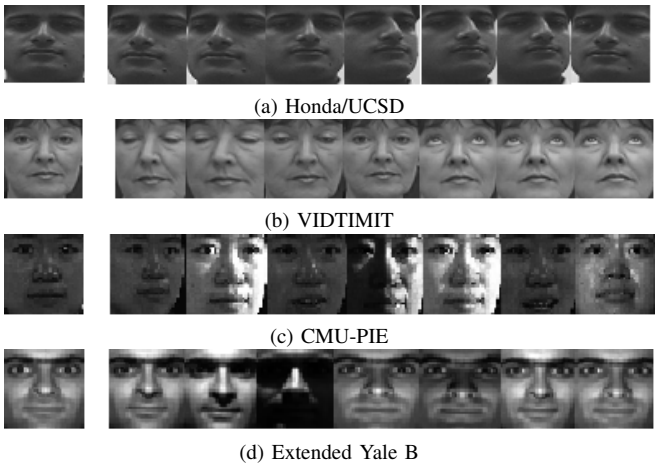
The VIDTIMIT[2]dataset [31] comprises sequences with 43 subjects, each subject containing 3 video sequences recorded in different moments. The average number of frames is approximately 100 with a resolution of 512x384 pixels. The average number of face images per subject is 300. In each video, the person moves his head to the left, right, back to the center, up, then down and finally to the center again. In VIDTIMIT pose does not vary as much as in the HONDA dataset, expression and lighting condition are roughly constant and there is no occlusion.

The CMU-PIE[3] dataset [32] contains 41,368 images of 68 subjects. For each subject, the images were taken under 13 different poses, 43 different illumination conditions and 4 different expressions. Each image has a resolution of 640x486 pixels.

The Extended Yale B[4] [33] dataset contains 16,128 images from 28 subjects under 9 poses and 64 illumination conditions.

Fig. 2 presents sample images of all four datasets. The samples in Fig. 2 show that the Honda/UCSD is clearly more challenging than VIDTIMIT in terms of pose and facial expressions. Regarding the static image datasets, CMU-PIE contains more pose variation than the Extended Yale.

[2]http://conradsanderson.id.au/vidtimit/
[3]http://www.ri.cmu.edu/
[4]http://vision.ucsd.edu/∼leekc/ExtYaleDatabase/ExtYaleB.html



(a) Honda/UCSD

(b) VIDTIMIT

(c) CMU-PIE

(d) Extended Yale B

Fig. 2. Samples of the gallery images(first column)and probe images from Honda/UCSD, VIDTIMIT, CMU-PIE and Extended Yale B datasets

B. Experimental Setup

Each layer was trained using backpropagation [34] and the Polack-Ribiere conjugate gradient method [35]. The latter replaced the Limited Memory Broyden Fletcher Goldfarb Shanno (L-BFGS) algorithm used in [13]. The reason for this choice is the comparatively lower complexity concerning the number of parameters. In our experiments we used the MATLAB implementation available at http://www.cs.toronto.edu/∼hinton/MatlabForSciencePaper.html.

Faces were automatically detected using the Viola & Jones algorithm [14] at each video frame. Each detected face was geometrically normalized to 32x32 pixels, keeping the detected centers of the eyes at fixed coordinates. For each subject, we selected a frontal image with good lighting conditions and no occlusion for the gallery. This image was selected from the subject‘s detected faces and removed from the test set.

In accordance with [13], we used an architecture containing 2 hidden layers with 1024 nodes each. The similarity parameter λ was set to 10^{-1}, and the regularization coefficient λ_{wd} to 10^{-7}. The sparsity parameters λ_{sp} and ρ_0 both related to the KL divergence, were set to 10^{-4} and 10^{-3} respectively.

We used *sigmoid* instead of *tanh* as activation function. In our experiments training with the sigmoid converged faster and the resulting networks achieved slightly higher recognition accuracies. Following [13], [36] and [37], the initial weight‘s values were randomly sampled between $\left[-\sqrt{\frac{6}{d_x+d_h}}, \sqrt{\frac{6}{d_x+d_h}}\right]$, where d_x and d_h are the dimensions of the input data and the neural network hidden layer respectively. For the bias b and $\hat{b}$ zero vectors were taken as initial values.

Four SSAE models were trained using face images collected from each data set individually. The evaluation was conducted only upon the two video datasets using the models trained on the other three datasets. In all cases the gallery comprised 100 subjects, being 19 from Honda/UCSD, 43 from VIDTIMIT and 18 randomly selected from CMU-PIE and Extended Yale B. Dissimilarity between SSAE representations were computed by four distances metrics: Euclidean [38], Cosine

TABLE I
RECOGNITION RATES FOR FRAME WISE RECOGNITION (%) WITH SSAE TRAINED ON CMU-PIE

Method	Honda/UCSD	VIDTIMIT
SSAE+$Euc.$	34	62
SSAE+$Cos.$	34	62
SSAE+$\chi^2.$	34	62
SSAE+$SRC.$	30	59

TABLE II
RECOGNITION RATES FOR FRAME WISE RECOGNITION (%) WITH SSAE TRAINED ON EXTENDED YALE B

Method	Honda/UCSD	VIDTIMIT
SSAE+$Euc.$	33	58
SSAE+$Cos.$	33	58
SSAE+$\chi^2.$	34	59
SSAE+$SRC.$	28	56

[39], χ^2 [6] and SRC [13].

The LBP codes for each pixel were generated from 8 samples over a circle with the center at that pixel and radius equal to 2 pixels. Histograms for each 8x8 pixel non-overlapping block of the resulting LBP image representation were the basis for recognition using the χ^2 distance metric.

C. Results and Analysis

Frame wise recognition

The first experiment had the objective to assess the recognition accuracy for individual face images collected from video for the four dissimilarity metrics. In this experiment we treated each face image extracted from the videos as a single probe to be recognized. Referring to Algorithm 2, a video sequence comprising n frames was processed as n video sequences composed by 1 frame.

The recognition rate was computed for each subject separately considering all available images of his/her face. The overall recognition rate was computed by averaging the subject specific rates. The results are shown in Table I,II and III.

The reported rates for VIDTIMIT are very low, whereas the rates for Honda/UCSD are even worse. Such disappointing rates have at least two main reasons: first, the typically low quality of images collected from video, and second, the usage of automatic detection and framing procedures. It is worth mentioning that the accuracies reported in [13] were measured on still manually annotated face images. In fact, the authors of [13] shortly mention that SSAE performs poorly when working on images annotated by state of the art detection algorithms. In this sense, the poor rates recorded in these experiments are no surprise.

TABLE III
RECOGNITION RATES FOR FRAME WISE RECOGNITION (%) WITH SSAE TRAINED ON VIDTIMIT(FOR HONDA/UCSD TEST) AND HONDA/UCSD(FOR VIDTIMIT TEST)

Method	Honda/UCSD	VIDTIMIT
SSAE+$Euc.$	22	58
SSAE+$Cos.$	22	58
SSAE+$\chi^2.$	21	58
SSAE+$SRC.$	30	59

The rates achieved on Honda/UCSD were much lower than VIDTIMIT because the former entails the more variations in lighting, pose, and occlusion than latter (see Fig. 2).

Regarding the dissimilarity metrics no substantial difference could be observed among results obtained with the Euclidian, cosine and χ^2 distances. SRC performed similarly to the other metrics for VIDTIMIT, but attained either the best or the worst rates for Honda/UCSD, still at very lower values.

Sequence wise recognition

The second experiment series aimed at assessing the improvements that could be achieved by exploiting the MSPPP collected from video sequences.

The function $\phi(\cdot)$ was first applied to all subsequences of length $\mathbb{L}$ of the same person using Algorithm 2. Then, the average recognition rate for length $\mathbb{L}$ was computed for that person. This procedure was executed for each subject. The mean recognition rate for length $\mathbb{L}$ was computed by averaging the per subject rates. This procedure was carried out for all possible values of $\mathbb{L}$. Fig.3 shows the recognition rate for different sequence lengths of three decision functions for sequence wise recognition using the cosine distance as dissimilarity metric in all cases.

The plots in Fig. 3a to 3c refer to tests on Honda/UCSD, whereas Fig. 3d to 3f related to tests performed on VIDTIMIT. The SSAE models were trained on CMU-PIE for the plots on the left (Fig. 3a and 3d) and on Extended Yale B for the plots in the middle (Fig.3b and 3e). Fig. 3c and 3f refer to models trained with images collected from VIDTIMIT and Honda/UCSD respectively.

The first observation from the plots is that sequence recognition is generally superior to frame wise recognition. This shows that the poor image quality of most face recognition applications from videos can be partially alleviated by exploiting the multiple face image samples collected along video sequences.

The accuracy increases in all cases with the sequence length, i.e., as more samples are added to the probe up to rates that are typical of recognition from photography taken under controlled conditions. Nevertheless the curves are not perfectly smooth; most of them present abrupt changes for some sequence lengths. This occurs because the sequences of face images in our video databases have different length depending on the subject. This means that the rates for shorter sequences are computed upon more subjects than for longer ones. The steps on the curves occur at lengths where the number of subjects available to compute the recognition rates changes, and are more abrupt for Honda/USCD than for VIDTIMIT, because the former has less subjects in total (19) than the latter (43). Yet the increasing profile is clearly recognizable in all curves, especially for shorter sequences.

A further observation is that *best score* outperformed the *median* and *majority voting* strategies with no exception.

Furthermore, SSAE was much more sensitive to image quality variations than LBP. This is clearly seen in the tests on Honda/USCD that embodies more variations than VIDTIMIT. This might be due to the intrinsic robustness of LBP to different illumination patterns. However, for well-behaved videos, as in VIDTIMIT, SSAE was able to outperform LBP, but only for sequences longer than 120 image samples.

V. Conclusions and Future Work

In this paper, we have investigated the performance of Stacked Supervised Auto-Encoders (SSAE) for single sample per person (SSPP) identification from video sequences. In this scenario each enrolled person is represented by a single image sample or by its descriptor in the gallery. Conversely, the probe may comprise multiple samples per person (MSPP) collected along the video sequence. Local Binary Patterns (LBP) was taken as baseline for the SSPP scenario.

Three strategies were tested experimentally upon two video datasets with different characteristics, Honda/USCD and VIDTIMIT. The experiments demonstrated that both LBP and SSAE can greatly benefit from probes comprising multiple samples.

Among the tested strategies for MSPP, the one that relies on the best match produced the best results among the strategies considered in this analysis.

In the tests conducted on Honda/USCD that contains much lighting variation LBP was superior to SSAE. However, SSAE outperformed LBP for VIDTIMIT in our experiments for probes comprising more than 120 image samples.

Our experiments were conducted on two datasets containing relatively few subjects. Experiments on larger datasets are planned for the continuation of this research.

Acknowledgment

This work is supported by CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior).

References

[1] Y. Su, S. Shan, X. Chen, and W. Gao, "Adaptive generic learning for face recognition from a single sample per person," 2010.

[2] X. Tan, S. Chen, Z.-H. Zhou, and F. Zhang, "Face recognition from a single image per person: A survey," *Pattern recognition*, vol. 39, no. 9, pp. 1725–1745, 2006.

[3] M. A. Turk and A. P. Pentland, "Face recognition using eigenfaces," in *Computer Vision and Pattern Recognition, 1991. Proceedings CVPR'91., IEEE Computer Society Conference on*. IEEE, 1991, pp. 586–591.

[4] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs. fisherfaces: Recognition using class specific linear projection," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 19, no. 7, pp. 711–720, 1997.

[5] Q.-x. Gao, L. Zhang, and D. Zhang, "Face recognition using flda with single training image per person," *Applied Mathematics and Computation*, vol. 205, no. 2, pp. 726–734, 2008.

[6] T. Ojala, M. Pietikäinen, and T. Mäenpää, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 24, no. 7, pp. 971–987, 2002.

[7] V. Ojansivu and J. Heikkilä, "Blur insensitive texture classification using local phase quantization," in *Image and signal processing*. Springer, 2008, pp. 236–243.

[8] Q. V. Le, "Building high-level features using large scale unsupervised learning," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 8595–8598.

[9] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[10] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural computation*, vol. 18, no. 7, pp. 1527–1554, 2006.

[11] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol, "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion," *The Journal of Machine Learning Research*, vol. 11, pp. 3371–3408, 2010.

[12] D. Erhan, Y. Bengio, A. Courville, P.-A. Manzagol, P. Vincent, and S. Bengio, "Why does unsupervised pre-training help deep learning?" *The Journal of Machine Learning Research*, vol. 11, pp. 625–660, 2010.

[13] S. Gao, Y. Zhang, K. Jia, J. Lu, and Y. Zhang, "Single sample face recognition via learning deep supervised autoencoders," *Information Forensics and Security, IEEE Transactions on*, vol. 10, no. 10, pp. 2108–2118, 2015.

[14] P. Viola and M. J. Jones, "Robust real-time face detection," *International journal of computer vision*, vol. 57, no. 2, pp. 137–154, 2004.

[15] J. Yang, D. Zhang, A. F. Frangi, and J.-y. Yang, "Two-dimensional pca: a new approach to appearance-based face representation and recognition," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 26, no. 1, pp. 131–137, 2004.

[16] J. Wu and Z.-H. Zhou, "Face recognition with one training image per person," *Pattern Recognition Letters*, vol. 23, no. 14, pp. 1711–1719, 2002.

[17] S. Chen, D. Zhang, and Z.-H. Zhou, "Enhanced (pc) 2 a for face recognition with one training image per person," *Pattern Recognition Letters*, vol. 25, no. 10, pp. 1173–1181, 2004.

[18] A. M. Martínez, "Recognizing imprecisely localized, partially occluded, and expression variant faces from a single sample per class," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 24, no. 6, pp. 748–763, 2002.

[19] S. Shan, B. Cao, W. Gao, and D. Zhao, "Extended fisherface for face recognition from a single example image per person," in *Circuits and Systems, 2002. ISCAS 2002. IEEE International Symposium on*, vol. 2. IEEE, 2002, pp. II–81.

[20] S. Chen, J. Liu, and Z.-H. Zhou, "Making flda applicable to face recognition with one sample per person," *Pattern recognition*, vol. 37, no. 7, pp. 1553–1555, 2004.

[21] Y. Tang, R. Salakhutdinov, and G. Hinton, "Deep lambertian networks," *arXiv preprint arXiv:1206.6445*, 2012.

[22] Y. Bengio, "Learning deep architectures for ai," *Foundations and trends® in Machine Learning*, vol. 2, no. 1, pp. 1–127, 2009.

[23] S. Rifai, P. Vincent, X. Muller, X. Glorot, and Y. Bengio, "Contractive auto-encoders: Explicit invariance during feature extraction," in *Proceedings of the 28th international conference on machine learning (ICML-11)*, 2011, pp. 833–840.

[24] S. Rifai, G. Mesnil, P. Vincent, X. Muller, Y. Bengio, Y. Dauphin, and X. Glorot, "Higher order contractive auto-encoder," in *Machine Learning and Knowledge Discovery in Databases*. Springer, 2011, pp. 645–660.

[25] W. Zou, S. Zhu, K. Yu, and A. Y. Ng, "Deep learning of invariant features via simulated fixations in video," in *Advances in neural information processing systems*, 2012, pp. 3212–3220.

[26] Q. V. Le, A. Karpenko, J. Ngiam, and A. Y. Ng, "Ica with reconstruction cost for efficient overcomplete feature learning," in *Advances in Neural Information Processing Systems*, 2011, pp. 1017–1025.

[27] M. Hayat, M. Bennamoun, and S. An, "Deep reconstruction models for image set classification," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 37, no. 4, pp. 713–727, 2015.

[28] X. Liu and T. Chen, "Video-based face recognition using adaptive hidden markov models," in *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on*, vol. 1. IEEE, 2003, pp. I–340.

[29] K.-C. Lee, J. Ho, M.-H. Yang, and D. Kriegman, "Video-based face recognition using probabilistic appearance manifolds," in *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on*, vol. 1. IEEE, 2003, pp. I–313.

[30] M. Kan, S. Shan, H. Chang, and X. Chen, "Stacked progressive auto-encoders (spae) for face recognition across poses," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1883–1890.

[31] C. Sanderson and B. C. Lovell, "Multi-region probabilistic histograms

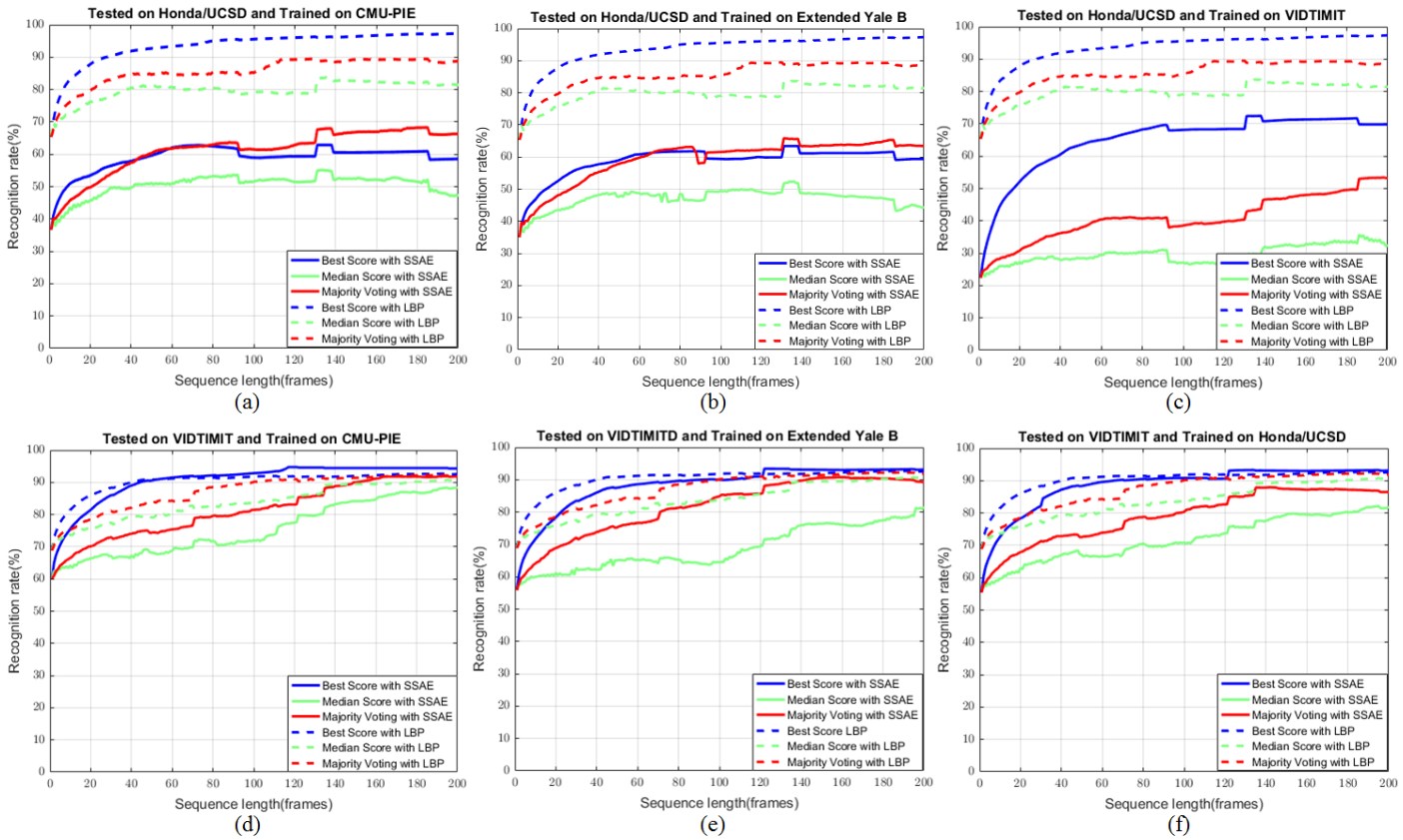


Fig. 3. Performance of the strategies along the video sequence

for robust and scalable identity inference," in *Advances in Biometrics*. Springer, 2009, pp. 199–208.

[32] T. Sim, S. Baker, and M. Bsat, "The cmu pose, illumination, and expression (pie) database," in *Automatic Face and Gesture Recognition, 2002. Proceedings. Fifth IEEE International Conference on*. IEEE, 2002, pp. 46–51.

[33] A. S. Georghiades, P. N. Belhumeur, and D. J. Kriegman, "From few to many: Illumination cone models for face recognition under variable lighting and pose," *IEEE transactions on pattern analysis and machine intelligence*, vol. 23, no. 6, pp. 643–660, 2001.

[34] R. Salakhutdinov, A. Mnih, and G. Hinton, "Restricted boltzmann machines for collaborative filtering," in *Proceedings of the 24th international conference on Machine learning*. ACM, 2007, pp. 791–798.

[35] A. H. Kramer and A. Sangiovanni-Vincentelli, "Efficient parallel learning algorithms for neural networks," in *Advances in neural information processing systems*, 1989, pp. 40–48.

[36] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *International conference on artificial intelligence and statistics*, 2010, pp. 249–256.

[37] Y. Bengio, "Practical recommendations for gradient-based training of deep architectures," in *Neural Networks: Tricks of the Trade*. Springer, 2012, pp. 437–478.

[38] C.-H. Chan, J. Kittler, and M. A. Tahir, "Kernel fusion of multiple histogram descriptors for robust face recognition," in *Structural, Syntactic, and Statistical Pattern Recognition*. Springer, 2010, pp. 718–727.

[39] M. S. Bartlett, J. R. Movellan, and T. J. Sejnowski, "Face recognition by independent component analysis," *Neural Networks, IEEE Transactions on*, vol. 13, no. 6, pp. 1450–1464, 2002.