On Robust Biometric Identity Verification via Sparse Encoding of Faces: Holistic vs Local Approaches

Yongkang Wong, Mehrtash T. Harandi, Conrad Sanderson, Brian C. Lovell

The University of Queensland, School of ITEE, QLD 4072, Australia

***Abstract*—In the field of face recognition, Sparse Representation (SR) has received considerable attention during the past few years. Most of the related literature focuses on holistic descriptors in closed-set identification applications. The underlying assumption in identification is that the gallery always has sufficient samples per subject to linearly reconstruct a query image. Unfortunately, such assumption is easily violated in the more challenging and realistic face verification scenario. A verification algorithm is required to determine if two faces (where one or both have not been seen before) belong to the same person, while explicitly taking into account the possibility of impostor attacks. In this paper, we first discuss why most of the SR literature is not applicable to verification problems. Motivated by the success of bag-of-words methods in the field of object recognition, which describe an image as a set of local patches or interest points, we then propose to tackle the verification problem by encoding each local face patch through SR. The locally encoded sparse vectors are pooled to form regional descriptors, where each descriptor covers a relatively large portion of the face. Experiments in various challenging conditions show that the proposed method achieves high and robust verification performance.**

I. Introduction

Face based identity inference (normally known by the all-encompassing term "face recognition"), can be generalised into two distinct configurations: identification and verification [1]. The task of identification is to classify a given face as belonging to one of K previously seen persons in a gallery. In such configuration, identification performance can be maximised by utilising class labels. For example, Linear Discriminant Analysis (LDA) [2] separates the gallery such that small within-class scatter and large between-class scatter are achieved. However, this identification task is a *closed-set* problem as it assumes impostor attacks do not exist (*i.e.* a probe face always matches one of the persons in the gallery). Therefore, algorithms relying on the closed-set assumption do not readily translate to real-world applications [3]. In contrast, the task of verification is to determine if two given faces (or two face sets) belong to the same person, which explicitly takes into account the possibility of impostor attacks. In challenging and realistic applications such as video surveillance [4], the verification system is able to make decisions for persons it has not seen in advance (*e.g.* person re-identification across multiple cameras [5]).

Wright *et al.* [6] recently proposed Sparse Representation based Classification (SRC) for face identification problems. The underlying idea is to represent a query sample $\boldsymbol{y}$ as a sparse linear combination of a dictionary $\boldsymbol{D}$, where the dictionary usually constitutes of holistic face descriptors. Moreover, it is assumed that each subject has sufficient samples in the dictionary to span over possible subspaces. Each probe image can be considered to be represented by a sparse code that is comprised of coefficients that linearly reconstruct the image via the dictionary. As such, it is expected that only those atoms in the dictionary that truly match the class query sample contribute to the sparse code. Wright *et al.* [6] exploited this by computing a class-specific similarity measure. More specifically, they computed the reconstruction error of a query image to class i by considering only the sparse codes associated with the atoms of the i-th class. The class that results in the minimum reconstruction error specifies the label of query. A more thorough discussion of class-based SRC can be found in [7].

A significant body of literature was proposed with the aim of improving the original SRC. For example, Yang and Zhang [8] extended the original approach to use a holistic representation derived from Gabor features. The Gabor-based SRC was shown to be relatively more robust against illumination changes as well as small degree of pose mismatches. Another example is the Robust Sparse Coding (RSC) scheme proposed by Yang *et al.* [9], where sparse coding is modelled as a sparsity-constrained robust regression problem. RSC was shown to outperform the original SRC and Gabor-based SRC, as well as being more effective in handling of face occlusions.

In spite of the recent success in face identification, SRC relies on the *sparsity* assumption. The assumption holds when each class in the gallery has sufficient samples and the query lies on the subspace spanned by the gallery of the same class. Shi *et al.* [10] questioned the validity of the sparsity assumption for face data and showed that the assumption may be violated even in the identification scenario. Since in a verification system there might not be any mutual overlap between the probe faces and the training data (*i.e.* the probe identities were never seen by the system during training), violation of the sparsity assumption is more likely to happen. In other words, a verification system needs to be capable of making decisions even for classes it has not seen before. This contradicts the sparsity assumption, and hence existing SRC approaches do not naturally extend to verification scenarios.

The current trend of SRC can be further criticised by observing that the majority of works represent faces in a rigid and holistic manner [7], [8] (*i.e.* holistic descriptors). That is, each face is represented by one feature vector that

describes the entire face, which implicitly embeds rigid spatial constraints between face components [1], [11], [12]. Examples of such representation include classic techniques such as PCA-based feature extraction [12]. Such treatment implies ideal image acquisition (*e.g.* perfect image alignment). In reality, especially for fully automated systems, attaining ideal images is very challenging (if not impossible) for low resolution moving objects [13]. The empirical studies in [1], [14] verified the adverse impacts of imperfect face acquisition on systems that utilise holistic face descriptors.

One way of addressing this problem is through an SR-based face alignment algorithm [15]. Given a set of frontal training images and a query face image, $\boldsymbol{x}_{\text{auto}}$, extracted using an automatic face locator (detector), the algorithm finds the image transformation parameters which transform $\boldsymbol{x}_{\text{auto}}$ for the best reconstruction error. This approach can be criticised as being a computationally intensive method for correcting rigid face descriptors, rather than tackling the source of the problem: rigid descriptors are inherently not robust to in-class face variations.

In contrast to such rigid representations, a face can also be represented by a set of local features with relaxed spatial constraints[1]. This allows for some movement and/or deformations of face components [1], [17], [18], and in turn leads to a degree of inherent robustness to expression and pose changes [18], as well as robustness to misalignment (where the misalignment is a byproduct of automatic face locators/detectors [1]). Recently, Aharon *et al.* [19] showed that local features satisfy the sparsity assumption when an overcomplete dictionary (trained from sufficient amount of samples) is presented. Given the above discussions, in this paper we focus on the use of local feature-based approaches to handle the problem of imperfect image acquisition.

A. Contributions

There are three main contributions in this paper. **(1)** We briefly discuss why current SRC is not applicable for verification problems and present a natural extension of SR with holistic representation to verification problems. **(2)** Motivated by the benefits of local feature-based face representation, we propose a new face descriptor, namely Local Sparse Encoded Descriptor (LSED), which is inspired by the well-established bag-of-words literature [20], [21], [22]. In LSED, sparse codes are obtained on local image patches using a learned dictionary. The local sparse codes from each image patch are then pooled together to form the face descriptor. **(3)** The LSED approach shows great robustness against environmental variations such as pose mismatches, imperfect face alignment, blurring, and variations caused by capturing images in various environmental conditions.

[1]However, it must be noted that not all local feature-based face representations automatically have relaxed spatial constraints. For example, in [16] local feature extraction is followed by concatenation of the local feature vectors into one long vector. The concatenation, in this case, effectively enforces rigid spatial constraints.

We continue the paper as follows. We first describe the background theory of sparse encoding in Section II. In Section III, we discuss how can holistic SR approaches be applied for face verification. In Section IV, we present and discuss the proposed LSED. Experiments on various identity inference experiments are shown in Section V. Section VI provides the main findings and suggests future directions.

II. Background Theory

In this section, we describe the background theory of two sparse encoding approaches, namely **(1)** l_1-minimisation, and **(2)** Sparse Autoencoder Neural Network (SANN). Consider a training set $\boldsymbol{Y} = [\, \boldsymbol{y}_1,\, \boldsymbol{y}_2,\, \cdots,\, \boldsymbol{y}_M \,] \in \mathbb{R}^{d \times M}$. Each sparse encoding approach learns a dictionary (or model), $\boldsymbol{D} \in \mathbb{R}^{d \times N}$, using an unsupervised learning algorithm, where each column $\boldsymbol{d}_i \in \mathbb{R}^d$ of the dictionary is an atom. Given the learned dictionary $\boldsymbol{D}$, a probe vector $\boldsymbol{x}$ is then encoded as a sparse code $\widehat{\boldsymbol{\alpha}}$ either by l_1-minimisation or SANN.

A. Sparse Encoding via l_1-minimisation

Given the trained overcomplete dictionary $\boldsymbol{D}$ and a probe vector $\boldsymbol{x}$, the corresponding sparse code $\widehat{\boldsymbol{\alpha}}$ can be estimated by solving the following l_1-minimisation problem:

$$\widehat{\boldsymbol{\alpha}} = \min \|\boldsymbol{\alpha}\|_1 \text{ subject to } \|\boldsymbol{D}\boldsymbol{\alpha} - \boldsymbol{x}\|_2^2 \leq \epsilon \qquad (1)$$

where $\|\cdot\|_p$ denotes the l_p-norm and ϵ is the threshold for the reconstruction error $\|\boldsymbol{D}\boldsymbol{\alpha} - \boldsymbol{y}\|_2^2$.

As discussed in [23], the choice of dictionary learning algorithm has little contribution to the performance of a selected sparse encoding algorithm. Therefore, the aforementioned l_1-minimisation problem can be coupled with any dictionary learning algorithm. In this paper, we train the dictionary $\boldsymbol{D}$ using the K-SVD algorithm [19]. The algorithm first initialises a random dictionary $\boldsymbol{D}$ with l_2 normalised atoms and performs an iterative two stage process until convergence. The objective function is to minimise the following cost function:

$$\min_{\boldsymbol{D},\boldsymbol{\alpha}} \|\boldsymbol{Y} - \boldsymbol{D}\boldsymbol{\alpha}^{\text{train}}\|_F^2 \text{ subject to } \forall i,\ \|\boldsymbol{\alpha}_i^{\text{train}}\|_0 \leq T_0 \quad (2)$$

The first stage (sparse coding stage), with dictionary $\boldsymbol{D}$, the representation vectors $\boldsymbol{\alpha}_i^{\text{train}}$ in Eqn. (2) are obtained using any pursuit algorithm [24]. In the second stage (dictionary update stage), the algorithm updates each atom, $\boldsymbol{d}_i$, by first computing the overall representation error matrix, $\boldsymbol{E}_i$, using:

$$\boldsymbol{E}_i = \boldsymbol{Y} - \sum\nolimits_{j \neq i} \boldsymbol{d}_j \boldsymbol{\alpha}_j^{\text{train}} \qquad (3)$$

By restricting to use a subset of $\boldsymbol{E}_i$, which corresponds to the training vectors that use the atom $\boldsymbol{d}_i$, we obtain $\boldsymbol{E}_i^R$. Let $\boldsymbol{U}\Delta\boldsymbol{V}^T$ represent the singular value decomposition of $\boldsymbol{E}_i^R$. The updated version of atom $\boldsymbol{d}_i$ is then obtained as the first column of $\boldsymbol{U}$.

B. Sparse Encoding via Sparse Autoencoder Neural Network

Under the framework of SANN [25], [26], we employ unsupervised model training, which trains a single hidden layer such that each training sample is reconstructed with a small subset of the hidden nodes. The objective is to learn a hidden layer that consists of N nodes, which is parameterised with a weight $\boldsymbol{W} \in \mathbb{R}^{d \times N}$ and bias $\boldsymbol{b} \in \mathbb{R}^N$. The back-propagation algorithm can be used for training by minimising the following cost function:

$$J(\boldsymbol{W}, \boldsymbol{b}) = J_{\text{error}} + J_{\text{weight}} + \beta J_{\text{sparsity}} \tag{4}$$

where

$$J_{\text{error}} = \frac{1}{M} \sum_{i=1}^{M} \left(\frac{1}{2} \|\widehat{\boldsymbol{x}}_i - \boldsymbol{x}_i\|^2 \right) \tag{5}$$

$$J_{\text{weight}} = \frac{\lambda}{2} \|\boldsymbol{W}\|^2 \tag{6}$$

$$J_{\text{sparsity}} = \sum_{i=1}^{N} \left[\rho \log \left(\frac{\rho}{\widehat{\rho}_i} \right) + (1-\rho) \log \left(\frac{1-\rho}{1-\widehat{\rho}_i} \right) \right] \tag{7}$$

The cost functions J_{error}, J_{weight}, and J_{sparsity} are respectively the *square reconstruction error* term, *weight decay* term and *sparsity penalty* term. J_{error} minimises the overall reconstruction error, with $\widehat{\boldsymbol{x}}_i$ denoting the reconstructed version of $\boldsymbol{x}_i$ [26]. The regularisation term J_{weight} decreases the magnitude of the weights to prevent overfitting. J_{sparsity} constrains the network to achieve low "activation", where ρ controls the degree of sparsity and $\widehat{\rho}_i$ is the average activation of hidden node i. The parameter β in Eqn. (4) controls the contribution of J_{sparsity}.

Given the trained SANN and a probe vector $\boldsymbol{x}$, the corresponding sparse code $\widehat{\boldsymbol{\alpha}} = \left[\widehat{\alpha}^{[1]}, \widehat{\alpha}^{[2]}, \cdots, \widehat{\alpha}^{[N]}\right]$ is calculated using:

$$\widehat{\alpha}^{[i]} = \text{sig}(\boldsymbol{w}_i^T \boldsymbol{x} + b_i) \tag{8}$$

where $\widehat{\alpha}^{[i]}$ is the i-th dimension of sparse code $\widehat{\boldsymbol{\alpha}}$, while $\boldsymbol{w}_i$ and b_i are the i-th weight and bias respectively. The logistic sigmoid function $\text{sig}(t) = 1/(1+\exp(-t))$ maps the output to the range of $[0, 1]$. In contrast to the l_1-minimisation approach, SANN has the advantage of avoiding the minimisation problem during the sparse encoding stage, resulting in a lower computational cost.

III. SR: Identification vs. Verification

In this section, we first briefly review how SR is applied to face identification problems. We then discuss why such methodology is not suitable for face verification problems and describe a natural extension to allow the use of SR with holistic descriptors in such problems.

A. Sparse Representation for Face Identification

Consider a closed-set face identification problem with a gallery comprised of N atoms. Let $\boldsymbol{D} \in \mathbb{R}^{d \times N}$ be the dictionary comprising all samples in the gallery. Given a query $\boldsymbol{x} \in \mathbb{R}^d$, the sparse solution $\widehat{\boldsymbol{\alpha}}$ can be estimated by solving Eqn. (1). Using only the coefficients associated with the i-th class, Wright *et al.* [6] computed the residual, $r_i(\boldsymbol{x})$, using:

$$r_i(\boldsymbol{x}) = \|\boldsymbol{x} - \boldsymbol{D}\delta_i(\widehat{\boldsymbol{\alpha}})\|_2^2 \tag{9}$$

where δ_i is denoted as a vector with the non-zero entries being the association to class i. The identity of query $\boldsymbol{x}$ is assigned using the rule: $\text{identity}(\boldsymbol{x}) = \arg\min_i r_i(\boldsymbol{x})$. This classification methodology is also used in the Gabor-based SRC [8] and RSC [9].

B. Extension of SR to Face Verification

In the context of face verification, the identities of probe faces may not be present in the gallery. As such, the sparsity assumption is likely to be violated, making the classification methodology described above not applicable to verification problems.

An alternative way to incorporate SR in verification problems is to use the sparse code (*i.e.* $\widehat{\boldsymbol{\alpha}}$) as a face descriptor. Given a dictionary $\boldsymbol{D}$ and two faces $\boldsymbol{x}_a, \boldsymbol{x}_b \in \mathbb{R}^d$, we first generate their respective sparse solutions $\widehat{\boldsymbol{\alpha}}_a$ and $\widehat{\boldsymbol{\alpha}}_b$ using Eqn. (1). The similarity score between these descriptors can be calculated using:

$$s_{\text{SR}}(\boldsymbol{x}_a, \boldsymbol{x}_b | \boldsymbol{D}) = \|\widehat{\boldsymbol{\alpha}}_a - \widehat{\boldsymbol{\alpha}}_b\|_2 \tag{10}$$

where a smaller value means a higher similarity between $\boldsymbol{x}_a$ and $\boldsymbol{x}_b$. The classification decision (*i.e.* whether $\boldsymbol{x}_a$ and $\boldsymbol{x}_b$ represent the same person) is obtained by comparing s_{SR} to a threshold value.

IV. Local Sparse Encoded Descriptor

In previous section, we have shown a natural extension of SR to verification problems. However, as shown in Section V, this holistic SR descriptor provides low discriminative information as well as having high sensitivity to out-of-plane rotations and imperfect face alignment. In this section, we present an alternative way to utilise sparse coding in verification problems. The proposed framework can be considered as a member of the family of bag-of-words methods, but being specially designed for face images. As such, the advantages of the bag-of-words methods like robustness against misalignment are inherited. We continue this section by first describing the proposed LSED algorithm, followed by elaborating on how the descriptor can be used for discriminating faces.

A. Framework

The overall idea of LSED is to describe a face image by sparse encoding of its local patches. A conceptual diagram of the framework is shown in Figure 1. A given face image is first split into R fixed size regions, where each region covers a relatively large portion of the face image. For region r, a set of low-dimensional feature vectors, $\boldsymbol{X}_r = \{\boldsymbol{x}_{r,1}, \boldsymbol{x}_{r,2}, \ldots, \boldsymbol{x}_{r,n}\}$, is attained by dividing the region into smaller patches $\boldsymbol{p}_{r,1}, \boldsymbol{p}_{r,2}, \ldots, \boldsymbol{p}_{r,n}$. To account for varying contrast caused by illumination changes, each patch is normalised to have zero mean and unit variance.

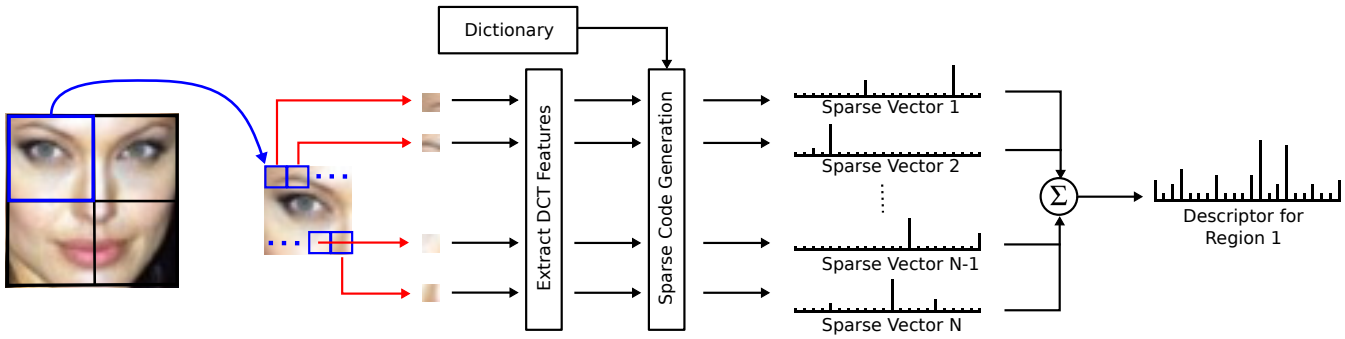

Fig. 1. Conceptual graph of the LSED framework. Face image is divided into regions followed by breaking each region into smaller patches. For each patch, a sparse vector is obtained by a sparse encoder using a learned dictionary. Each regional face descriptor is computed by pooling the sparse vectors from the corresponding region.

From each normalised patch $\widehat{\boldsymbol{p}}_{r,i}$, a low dimensional feature vector, $\boldsymbol{x}_{r,i}$, is obtained via 2D DCT decomposition [27]. Preliminary experiments suggest that patches of size 8×8 pixels with 75% overlap (*i.e.* adjacent patches are overlapped by either 6×8 or 8×6 pixels) lead to good performance. Moreover, we selected the 15 lowest frequency components of 2D DCT coefficients, with the DC coefficient discarded (as it is zero due to the aforementioned normalisation step).

Each i-th patch from region r, $\boldsymbol{p}_{r,i}$, is then described by a sparse code $\widehat{\boldsymbol{\alpha}}_{r,i}$. The sparse code is generated via l_1-minimisation (using Eqn. 1) or SANN (using Eqn. 8). Having each patch represented by a sparse code, each region r is then described via:

$$\boldsymbol{h}_r = \frac{1}{N_{\text{patch}}} \sum\nolimits_{i=1}^{N_{\text{patch}}} \left[\left|\widehat{\alpha}_{r,i}^{[1]}\right|, \left|\widehat{\alpha}_{r,i}^{[2]}\right|, \cdots, \left|\widehat{\alpha}_{r,i}^{[N]}\right| \right] \quad (11)$$

where N is the dimensionality of LSED (*i.e.* size of the sparse code) and N_{patch} is the number of patches in region r.

B. Similarity-Based Classification

Comparison between two faces is accomplished by comparing their corresponding regional descriptors. Using the method from [22], the matching score between face A and B can be calculated via:

$$s_{\texttt{raw}}(A,B) = \frac{1}{R} \sum\nolimits_{r=1}^{R} \left\| \boldsymbol{h}_r^{[A]} - \boldsymbol{h}_r^{[B]} \right\|_1 \quad (12)$$

where R is the number of regions. To account for uncontrolled image conditions not already handled by the patch-based analysis, a cohort normalisation [3], [22] based distance is employed:

$$s_{\texttt{norm}}(A,B) = \frac{s_{\texttt{raw}}(A,B)}{\sum_{i=1}^{N_C} s_{\texttt{raw}}(A,C_i) + \sum_{i=1}^{N_C} s_{\texttt{raw}}(B,C_i)} \quad (13)$$

where the cohort faces C_i are assumed to be reference faces that are different from images of persons A or B. To reach a decision as to whether face A and B belonging to the same person, $s_{\texttt{norm}}(A,B)$ can be compared to a decision threshold.

V. Experiments

In this section, we examine the performance of the proposed method on several identity inference configurations: **(1)** verification with various face alignment errors and sharpness variations, **(2)** verification with pose mismatches, and **(3)** verification with controlled and uncontrolled images. Experiments were conducted on four datasets: FERET [28], AR [29], BANCA [30], and ChokePoint [31]. Figure 2 shows example raw images. In all experiments, we used closely cropped face images with a size of 64×64 pixels. Except for experiments with simulated image variations, each image was manually aligned so that the eyes were at fixed positions. See Figure 3 for examples.

In the following subsections, we denote the LSED with Sparse Autoencoder Neural Network as LSED + SANN and the l_1-minimisation based approach as LSED + l_1. The proposed method has a number of parameters that affect performance. Based on preliminary experiments, we selected 3×3 regions and 32 cohorts for the distance normalisation in Eqn. (13). LSED + SANN has 512 hidden units, where the parameters of the cost function (Eqns. (4) and (7)) were set as $\beta = 3$, $\rho = 0.1$, and $\lambda = 0.01$. LSED + l_1 has a dictionary with 1024 atoms and the threshold for reconstruction error, ϵ, in Eqn. (1) was set to 0.1. These parameters were kept unchanged for all experiments.

In each of the following verification experiments, the face images were divided into three sets: (1) training set, (2) development set, and (3) evaluation set. For all experiments, except the verification experiment on the BANCA dataset, we exclusively used the CAS-PEAL dataset [32] as the training set. The CAS-PEAL dataset provides 1200 face images from 1200 unique individuals. The development and evaluation sets have a balanced number of matched and mismatched pairs.

Using the development set, we obtained a decision threshold, τ_D, which was then used on the evaluation set for assessing the final accuracy. Specifically, the threshold was adjusted such that the False Acceptance Rate (FAR) and False Rejection Rate (FRR) on the development set were equal. The threshold was then applied on the evaluation set, with the final accuracy defined as $1 - \frac{1}{2}(\text{FAR} + \text{FRR})$.

Fig. 2. Example raw images of the datasets used in this paper. (a) AR dataset contains 14 images per subject with various expressions and lighting conditions. (b) BANCA dataset: each subject was recorded under 3 scenarios: *controlled* (columns 1 & 3), *degraded* (column 2), and *adverse* (column 4). (c) ChokePoint dataset contains 29 subjects captured in 4 distinct portals.

Fig. 3. Examples of cropped images.

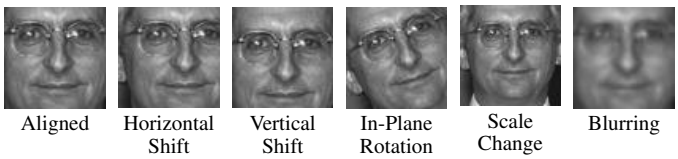

Fig. 4. Examples of simulated image variations on FERET.

In all experiments, we compared the proposed algorithm with the holistic SR descriptor described in Section III-B. We evaluated the holistic SR descriptor with two holistic features: **(1)** PCA based [2] (denoted as *PCA + SR*), and **(2)** Gabor based [33] (denoted as *Gabor + SR*). The similarity scores between these holistic SR descriptors were calculated via l_2-norm distance measurement.

To obtain PCA based face descriptors, we applied PCA on the vectorised image, which was attained by concatenation of the image pixels into a large vector. The Gabor feature based face descriptors were extracted with the same configuration as in [8], with PCA based dimensionality reduction. For both feature types, PCA preserved 98% of the total energy.

A. Face Verification with Alignment Errors and Blurring

In this section, we evaluate the robustness of the proposed method on blurring, as well as on four alignment errors using images taken from the 'fb' subset of FERET. Example images are shown in Figure 4. The generated alignment errors[2] are: horizontal shift and vertical shift (using displacements of ± 2, ± 4, ± 6, ± 8 pixels), in-plane rotation (using rotations of $\pm 10°$, $\pm 20°$, $\pm 30°$), and scale variations (using scaling factors of 0.7, 0.8, 0.9, 1.1, 1.2, 1.3). To simulate variations in sharpness, each original image was first downscaled to three sizes (48×48, 32×32 and 16×16 pixels), and then rescaled to the baseline size of 64×64 pixels. Using the frontal subset 'ba' and the expression subset 'bj', we randomly generated 800 matched and mismatched pairs for each alignment error. The experiments were conducted with 5-fold validations. We report the mean accuracy for each scenario.

The results, presented in Figure 5, show that PCA + SR performed poorly on all misalignment errors, with an overall accuracy of 58.4%. Moreover, Gabor + SR only improved the performance by a small margin. Overall, the proposed LSED methods consistently achieved robust performance in all simulated scenarios. LSED + SANN and LSED + l_1 achieved average accuracies of 85.8% and 89.2%, respectively.

[2]The generated alignment errors are representatives of real-life characteristics of automatic face localisation/detection algorithms [14].

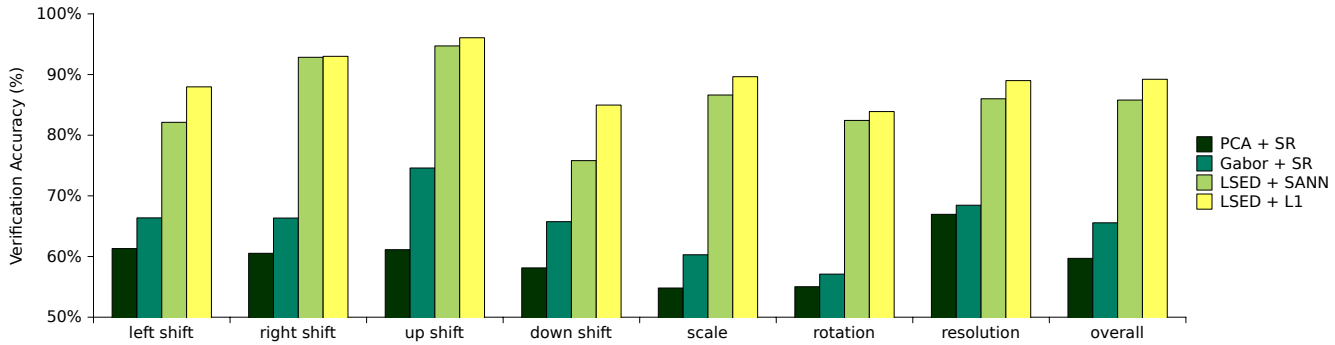

Fig. 5. The average verification accuracy on FERET images with stimulated alignment errors and sharpness variations (demonstrated in Fig. 4). Experiments were conducted with 5-fold validations.

Fig. 6. Examples of the FERET pose subset.

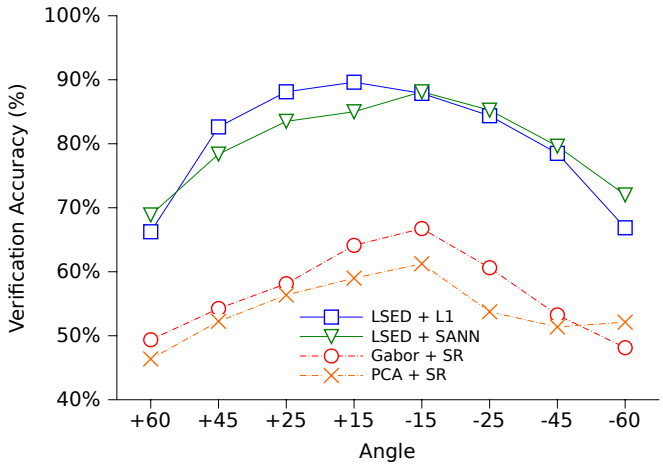

Fig. 7. Verification performance on pose mismatches for various angles. Faces from each pose angle are compared with the frontal subset 'ba' and the expression subset 'bj'. Experiments were conducted with 5-fold validations.

B. Face Verification with Pose Mismatches

In this section, we evaluate the robustness of the proposed method for handling pose mismatches. We selected the 'b' subset from the FERET dataset, which has 200 images per pose. The evaluation process on each pose angle was the same as the method described in the previous section. Example images are shown in Figure 6.

The results, shown in Figure 7, indicate that the proposed method considerably outperforms both PCA + SR and Gabor + SR. Both of the holistic SR descriptors obtained a maximum accuracy of 54% when the absolute value of the pose angle was $\geq 45°$. In contrast, LSED + SANN and LSED + l_1 achieved notably higher average accuracies of 74.7% and 73.6%, respectively. When the pose angle was between $-25°$ and $+25°$ (*i.e.* relatively frontal) the best performing holistic SR descriptor (Gabor + SR) achieved an average accuracy of about 63%, while the best performing LSED variant (LSED + l_1) achieved 87.5%.

C. Face Verification with Frontal Faces

In this experiment, we evaluated the performance on three datasets with images captured in various environment conditions. Example images are shown in Figure 3. The first dataset is AR [29], which contains 100 unique subjects with 14 images per subject. We randomly generated 9800 pairs of matched and mismatched pairs and evaluated the performance of each algorithm with 5-fold validations.

The second dataset is BANCA [30]. We report only the results on the 'P' protocol, where the algorithm was trained in controlled conditions and tested on a combination of controlled, degraded and adverse images. According to the protocol, the 52 subjects were divided into two groups, where each group played the role of the development set and evaluation set in turn. We randomly selected one image per person from each video.

The third dataset is ChokePoint [31], which was recorded under real-world surveillance conditions. It has 16 videos of 29 subjects recorded on four distinct portals[3]. We randomly generated 38,710 matched and mismatched image pairs where each pair consisted of images taken from different portals (*i.e.* cross environment matching). The experiments were evaluated with 5-fold validations.

The results, presented in Table I, show that the proposed LSED methods always obtained the best performance. Both PCA + SR and Gabor + SR performed at their best on the laboratory captured AR dataset and considerably worse on the more realistic ChokePoint dataset. This indicates that holistic SR descriptors are sensitive to image quality. The results also demonstrate the robustness of the proposed methods for images captured in various environment conditions.

Overall, LSED + l_1 achieved the best performance in all the verification experiments. However, it comes at the expense of considerably higher computational cost. As shown in Table II, LSED + l_1 requires 7739 milliseconds to generate a single face descriptor. In contrast, LSED + SANN is approximately 70 times faster.

[3]A portal is a location where a camera rig is placed to capture faces from multiple angles. Each portal has a unique background and lighting condition.

TABLE I
FACE VERIFICATION PERFORMANCE OVER SEVERAL DATASETS. THE FACE IMAGES WERE CLOSELY CROPPED TO EXCLUDE HAIR AND BACKGROUND, AND SCALED TO 64×64 PIXELS.

Method	AR	BANCA	ChokePoint
PCA + SR	62.5%	60.1%	55.0%
Gabor + SR	66.1%	63.3%	59.5%
LSED + SANN	72.2%	73.4%	75.1%
LSED + l_1	**80.0%**	**82.0%**	**79.8%**

TABLE II
AVERAGE COMPUTATION TIME OF LSED GENERATION FOR ONE IMAGE.

Method	Time (milliseconds)
LSED + SANN	110
LSED + l_1	7739

VI. MAIN FINDINGS AND FUTURE DIRECTIONS

In this paper we have first shown that a natural extension of holistic SR-based face identification does not result in a robust and effective face verification system. Motivated by the obscurity of the ways sparse representation can be extended to the problem of face verification, we proposed a local and sparse face descriptor, namely Local Sparse Encoded Descriptor (LSED). In the proposed descriptor, sparse codes are obtained on local image patches using a learned dictionary. The local sparse codes are then pooled together to form the face descriptor. We selected two approaches to obtain the sparse codes, namely Sparse Autoencoder Neural Network (SANN) and l_1-minimisation.

The l_1-minimisation based LSED and SANN based LSED were evaluated on several face datasets, where they consistently achieved good performance for faces captured in various environment conditions, as well as being robust against pose mismatches, blurring, and face misalignment errors.

Overall, the l_1-minimisation based LSED always achieved better results when compared with SANN based LSED, but at the expense of considerably higher computational load.

Future avenues of research include a study of the performance of the proposed LSED on closed-set identification as well as image-set [34] and video-to-video matching, in particular on uncontrolled video sequences (*e.g.* YouTube Faces dataset [35]).

REFERENCES

[1] F. Cardinaux, C. Sanderson, and S. Bengio, "User authentication via adapted statistical models of face images," *IEEE Transactions on Signal Processing*, vol. 54, no. 1, pp. 361–373, 2006.

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

[3] G. R. Doddington, M. A. Przybocki, A. F. Martin, and D. A. Reynolds, "The NIST speaker recognition evaluation - overview, methodology, systems, results, perspective," *Speech Communication*, vol. 31, no. 2-3, pp. 225–254, 2000.

[4] T. Ali, R. Veldhuis, and L. Spreeuwers, "Forensic face recognition: A survey," Centre for Telematics and Information Technology, University of Twente, Tech. Rep. TR-CTIT-10-40, December 2010.

[5] P. H. Tu, G. Doretto, N. O. Krahnstoever, A. A. Perera, F. W. Wheeler, X. Liu, J. Rittscher, T. B. Sebastian, T. Yu, and K. G. Harding, "An intelligent video framework for homeland protection," in *Proceedings of SPIE Defence and Security Symposium - Unattended Ground, Sea, and Air Sensor Technologies and Applications IX*, vol. 6562, 2007.

[6] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 2, pp. 210–227, 2009.

[7] J. Yang, L. Zhang, Y. Xu, and J.-Y. Yang, "Beyond sparsity: The role of l_1-optimizer in pattern classification," *Pattern Recognition*, vol. 45, no. 3, pp. 1104–1118, 2012.

[8] M. Yang and L. Zhang, "Gabor feature based sparse representation for face recognition with Gabor occlusion dictionary," in *ECCV (6)*, ser. Lecture Notes in Computer Science, vol. 6316. Springer, 2010, pp. 448–461.

[9] M. Yang, L. Zhang, J. Yang, and D. Zhang, "Robust sparse coding for face recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2011, pp. 625–632.

[10] Q. Shi, A. Eriksson, A. van den Hengel, and C. Shen, "Is face recognition really a compressive sensing problem?" in *IEEE Conference on Computer Vision and Pattern Recognition*, 2011, pp. 553–560.

[11] M. T. Harandi, M. N. Ahmadabadi, and B. N. Araabi, "Optimal local basis: A reinforcement learning approach for face recognition," *International Journal of Computer Vision*, vol. 81, no. 2, pp. 191–204, 2009.

[12] M. Turk and A. Pentland, "Eigenfaces for recognition," *Journal of Cognitive Neuroscience*, vol. 3, no. 1, pp. 71–86, 1991.

[13] A. Torralba and P. Shina, "Detecting faces in improverished images," *Technical Report 028, MIT AI Lab*, 2001.

[14] Y. Rodriguez, F. Cardinaux, S. Bengio, and J. Mariéthoz, "Measuring the performance of face localization systems," *Image and Vision Computing*, vol. 24, no. 8, pp. 882–893, 2006.

[15] A. Wagner, J. Wright, A. Ganesh, Z. Zhou, H. Mobahi, and Y. Ma, "Towards a practical face recognition system: Robust alignment and illumination by sparse representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 2, pp. 372–386, 2012.

[16] H. K. Ekenel and R. Stiefelhagen, "Local appearance based face recognition using discrete cosine transform," in *European Signal Processing Conference*, 2005.

[17] B. Heisele, P. Ho, J. Wu, and T. Poggio, "Face recognition: component-based versus global approaches," *Computer Vision and Image Understanding*, vol. 91, no. 1-2, pp. 6–21, 2003.

[18] C. Sanderson, S. Bengio, and Y. Gao, "On transforming statistical models for non-frontal face verification," *Pattern Recognition*, vol. 39, no. 2, pp. 288–302, 2006.

[19] M. Aharon, M. Elad, and A. Bruckstein, "K-svd: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Transactions on Signal Processing*, vol. 54, no. 11, pp. 4311–4322, 2006.

[20] G. Csurka, C. R. Dance, L. Fan, J. Willamowski, and C. Bray, "Visual categorization with bags of keypoints," in *In Workshop on Statistical Learning in Computer Vision, ECCV*, 2004, pp. 1–22.

[21] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2006, pp. 2169–2178.

[22] C. Sanderson and B. C. Lovell, "Multi-region probabilistic histograms for robust and scalable identity inference," in *Lecture Notes in Computer Science (LNCS)*, vol. 5558, 2009, pp. 199–208.

[23] A. Coates and A. Y. Ng, "The importance of encoding versus training with sparse coding and vector quantization," in *Proceedings of International Conference on Machine Learning*, June 2011, pp. 921–928.

[24] J. A. Tropp, A. C. Gilbert, and M. J. Strauss, "Algorithms for simultaneous sparse approximation. Part I: Greedy pursuit," *Signal Processing*, vol. 86, no. 3, pp. 572–588, 2006.

[25] M. Ranzato, Y.-L. Boureau, and Y. LeCun, "Sparse feature learning for deep belief networks," in *NIPS*, 2007.

[26] I. J. Goodfellow, Q. V. Le, A. M. Saxe, H. Lee, and A. Y. Ng, "Measuring invariances in deep networks," in *Advances in Neural Information Processing Systems*, 2009, pp. 646–654.

[27] R. Gonzalez and R. Woods, *Digital Image Processing*, 3rd ed. Prentice Hall, 2007.

[28] P. J. Phillips, H. Wechsler, J. Huang, and P. J. Rauss, "The FERET database and evaluation procedure for face-recognition algorithms," *Image and Vision Computing*, vol. 16, no. 5, pp. 295–306, 1998.

[29] A. Martínez and R. Benavente, "The AR face database," Computer Vision Center, Universitat Autónoma de Barcelona, CVC Technical Report 24, June 1998.

[30] E. Bailly-Bailliére, S. Bengio, F. Bimbot, M. Hamouz, J. Kittler, J. Mariéthoz, J. Matas, K. Messer, V. Popovici, F. Porée, B. Ruíz, and J.-P. Thiran, "The BANCA database and evaluation protocol," in *Audio- and Video-based Biometric Person Authentication (AVBPA), Lecture Notes in Computer Science (LNCS)*, vol. 2688, 2003, pp. 625–638.

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

[32] W. Gao, B. Cao, S. Shan, X. Chen, D. Zhou, X. Zhang, and D. Zhao, "The CAS-PEAL large-scale Chinese face database and baseline evaluations," *IEEE Transactions on Systems, Man, and Cybernetics—Part A: Systems and Humans*, vol. 38, no. 1, pp. 149–161, 2008.

[33] C. Liu and H. Wechsler, "Gabor feature based classification using the enhanced Fisher linear discriminant model for face recognition," *IEEE Transactions on Image Processing*, vol. 11, no. 4, pp. 467–476, 2002.

[34] M. T. Harandi, C. Sanderson, S. Shirazi, and B. C. Lovell, "Graph embedding discriminant analysis on Grassmannian manifolds for improved image set matching," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2011, pp. 2705–2712.

[35] L. Wolf, T. Hassner, and I. Maoz, "Face recognition in unconstrained videos with matched background similarity," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2011, pp. 529–534.