On Robust Face Recognition via Sparse Encoding: the Good, the Bad, and the Ugly

Yongkang Wong, Mehrtash T. Harandi, Conrad Sanderson

SeSaMe Centre, National University of Singapore, Singapore
NICTA, GPO Box 2434, Brisbane, QLD 4001, Australia
University of Queensland, School of ITEE, QLD 4072, Australia
Queensland University of Technology, Brisbane, QLD 4000, Australia

***Abstract*—In the field of face recognition, Sparse Representation (SR) has received considerable attention during the past few years. Most of the relevant literature focuses on holistic descriptors in closed-set identification applications. The underlying assumption in SR-based methods is that each class in the gallery has sufficient samples and the query lies on the subspace spanned by the gallery of the same class. Unfortunately, such assumption is easily violated in the more challenging face verification scenario, where an algorithm is required to determine if two faces (where one or both have not been seen before) belong to the same person. In this paper, we first discuss why previous attempts with SR might not be applicable to verification problems. We then propose an alternative approach to face verification via SR. Specifically, we propose to use explicit SR encoding on local image patches rather than the entire face. The obtained sparse signals are pooled via averaging to form multiple region descriptors, which are then concatenated to form an overall face descriptor. Due to the deliberate loss spatial relations within each region (caused by averaging), the resulting descriptor is robust to misalignment and various image deformations. Within the proposed framework, we evaluate several SR encoding techniques: l_1-minimisation, Sparse Autoencoder Neural Network (SANN), and an implicit probabilistic technique based on Gaussian Mixture Models. Thorough experiments on AR, FERET, exYaleB, BANCA and ChokePoint datasets show that the proposed local SR approach obtains considerably better and more robust performance than several previous state-of-the-art holistic SR methods, in both verification and closed-set identification problems. The experiments also show that l_1-minimisation based encoding has a considerably higher computational cost when compared to SANN-based and probabilistic encoding, but leads to higher recognition rates.**

I. Introduction

Face based identity inference (normally known by the all-encompassing term “face recognition”), can be generalised into three distinct configurations: closed-set identification, open-set identification, and verification [10]. The task of closed-set 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) [5] separates the gallery such that small within-class scatter and large between-class scatter are achieved. However, this closed-set identification task assumes impostor attacks do not exist and that each probe face must match a person in the gallery. This is a necessarily limiting assumption (as the gallery cannot cover all people in existance), and hence algorithms specifically relying on the closed-set assumption do not readily translate to real-world applications [16]. In contrast, both open-set identification and verification explicitly take into account the possibilities of impostor attacks and previously unseen people. In open-set identification, a given face is assigned to one of $K+1$ classes, with the extra class representing an “unknown person”. The task of verification is to determine if two given faces (or two face sets) belong to the same person, where one or both identities may not have been observed beforehand.

Verification can be implemented as a pair-wise comparison, resulting in a distance or probability that is then thresholded to achieve the final decision (which is a binary yes/no). As such, open-set identification can be decomposed into a set of verification tasks (one for each person in the gallery), as long as the pair-wise verification distances or probabilities are used instead of the verification decisions. In addition to the task of biometric user authentication [10], [16], the ability to handle previously unseen people is useful in video surveillance [3], for applications such as person re-identification across multiple cameras [43].

Wright et al. [48] 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 contains 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. [48] 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. To handle the case of a preson not present in the gallery, the given query image is considered as an imposter if the minimum reconstruction error exceeds a predefined threshold.

This paper is a revised and extended version of our earlier work [47].

A more thorough discussion of class-based SRC can be found in [49].

A significant body of literature was proposed with the aim of improving the original SRC. For example, Yang and Zhang [50] extended the original approach to use a holistic representation derived from Gabor features. The Gabor-based SRC (GSRC) 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. [52], where sparse coding is modelled as a sparsity-constrained robust regression problem. RSC was shown to outperform the original SRC and GSRC, as well as being more effective in handling of face occlusions. However, RSC is computationally more expensive when compared to various SRC approaches. Yang et al. [51] explored the benefit of a structured dictionary, where each atom is associated to a class label. Using the Fisher discrimination criterion [19], a set of class-specified sub-dictionaries is learned, where each class has small within-class scatter and large between-class scatter.

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. [37] 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 (ie. 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 majority of SR-based systems represent faces in a rigid and holistic manner [48], [49], [50] (ie. holistic descriptors). That is, each face is represented by one feature vector that describes the entire face and implicitly embeds rigid spatial constraints between face components [10], [23]. Examples of such representation include classic techniques such as PCA-based feature extraction [44]. Such treatment implies ideal image acquisition (eg. perfect image alignment, perfect localisation/detection). In reality, especially for fully automated systems, attaining ideal images is very challenging (if not impossible) for low resolution moving objects [38]. The adverse impact of imperfect face acquisition on recognition systems that utilise holistic face descriptors has been demonstrated in [10], [33].

To tackle misalignment problems, Wagner et al. [45] recently proposed an SR-based face alignment algorithm. 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. Though this approach has shown promising results, it 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 (eg. face expressions variations).

In contrast to 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 [10], [24], [35], and in turn leads to a degree of inherent robustness to expression and pose changes [35], as well as robustness to misalignment (where the misalignment is a byproduct of automatic face locators/detectors [10]). Aharon et al. [1] showed that local features satisfy the sparsity assumption when an overcomplete dictionary (trained from a sufficient amount of samples) is presented. Therefore, in this paper we focus on the use of SR for encoding local features to handle the problem of imperfect image acquisition.

In the field of object recognition, *bag-of-words* (BoW) approaches [15], [26] have been shown to be robust and effective for general image categorisation problems. The underlying idea is to treat any given image as a set of local keypoints or patches, followed by assigning each patch to a predetermined word with a vector quantisation (VQ) algorithm. The given image can be represented as a vector of assignments, where each dimension of the vector indicates the count of patches assigned to a particular word. In the field of face recognition, an extension of BoW for face images, called Multi-Region Histograms (MRH), represents each image as a concatenated set of regional probabilistic histograms [36].

We first note that VQ and probabilistic approaches to BoW representations can be considered as a form of sparse coding [13]. With this in mind, we propose to employ more direct forms of SR within the MRH framework, namely l_1-minimisation and a Sparse Autoencoder Neural Network (SANN). We denote this approach as Locally Sparse Encoded Descriptor (LSED). As shown later, LSED in conjunction with l_1-minimisation outperforms MRH as well as previous holistic SR methods, obtaining state-of-the-art performance in various identity inference configurations (ie. both verification and identification).

A. Contributions

There are four main contributions in this paper:

- We briefly discuss why previous attempts with SR are not be applicable for verification tasks and show a possible rudimentary extension of SR (with holistic face representation) to such tasks.
- In contrast to following the traditional approach of using holistic face representation in conjunction with SR, we explicitly use a local feature-based face representation (based on the well-established bag-of-words literature [15], [26], [36]) and employ SR to encode local image patches. In addition to the probabilistic approach for SR implicitly used by MRH [36], we study the efficacy of

[1] However, it must be noted that not all local feature-based face representations automatically have relaxed spatial constraints. For example, in [18] 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.

two more direct SR techniques, namely l_1-minimisation and Sparse Autoencoder Neural Network (SANN).

- Via thorough evaluations on face images captured in controlled and uncontrolled environment conditions, as well as in various challenging situations such as pose mismatches, imperfect face alignment, blurring, etc., we show that the proposed local feature SR approach considerably outperforms state-of-the-art holistic SR approaches. The experiments are conducted in both verification and closed-set identification setups.
- We analyse the computation cost of the proposed local feature SR approach in conjunction with various SR encoding techniques. We show that l_1 encoding leads to the highest accuracy at the expense of considerably higher computation cost than the second best technique, which is implicit SR encoding via probabilistic histograms.

We continue the paper as follows. We first delineate 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. Section V is devoted to experiments on various identity inference experiments using still images. Image set matching experiments are given in Section VI. Section VII provides the main findings.

II. Background Theory

In this section, we delineate the background theory of three sparse encoding approaches, namely: **(a)** l_1-minimisation, **(b)** Sparse Autoencoder Neural Network (SANN), and **(c)** probabilistic approach. Consider a finite training set $\boldsymbol{Y} = [\,\boldsymbol{y}_1,\, \boldsymbol{y}_2,\, \cdots,\, \boldsymbol{y}_M\,] \in \mathbb{R}^{d\times M}$. Each sparse encoding approach requires a dictionary (or model), $\boldsymbol{D} \in \mathbb{R}^{d\times N}$, where each column $\boldsymbol{d}_i \in \mathbb{R}^d$ is called an atom. Given the learned dictionary $\boldsymbol{D}$, a probe vector $\boldsymbol{x}$ is then encoded as a sparse code $\widehat{\boldsymbol{\alpha}}$ by a chosen encoding scheme.

A. Sparse Encoding via l_1-minimisation

Given the trained overcomplete dictionary $\boldsymbol{D}$ and a probe vector $\boldsymbol{x} \in \mathbb{R}^d$ that is compressible, a sparse solution $\widehat{\boldsymbol{\alpha}} \in \mathbb{R}^N$ exists such that $\boldsymbol{x}$ can be reconstructed with small residual. The sparse solution $\widehat{\boldsymbol{\alpha}}$ can be found by solving the following l_0-minimisation problem:

$$\min \|\boldsymbol{\alpha}\|_0 \text{ subject to } \|\boldsymbol{D\alpha} - \boldsymbol{x}\|_2^2 \le \epsilon \tag{1}$$

where the notation $\|\boldsymbol{\alpha}\|_0$ counts the nonzero entries of $\boldsymbol{\alpha}$ and ϵ is the threshold for the reconstruction error $\|\boldsymbol{D\alpha} - \boldsymbol{y}\|_2^2$.

Solving the l_0-minimisation problem is NP-hard and difficult to approximate. As shown in [42], the solution of Eqn. (1) can be approximated with the following l_1-minimisation (aka convex relaxation) problem:

$$\min \|\boldsymbol{\alpha}\|_1 \text{ subject to } \|\boldsymbol{D\alpha} - \boldsymbol{x}\|_2^2 \le \epsilon \tag{2}$$

which can be solved in polynomial time by linear programming methods [48], [12]. Another popular choice of sparse approximation technique is called the greedy pursuit approach, which approximates the sparse solution through iterative local approximation. However, the greedy pursuit approach can only produce the optimal solution under very strict conditions [40], whereas the convex relaxation has proven to be able to produce optimal or near optimal solutions for variety of problems [42].

As discussed in [14], the choice of the dictionary learning algorithm has minor influence 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 [1], which is effective for representing small image patches for sparse encoding problems [34]. 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:

$$\min_{\boldsymbol{D},\boldsymbol{\alpha}} \|\boldsymbol{Y} - \boldsymbol{D\alpha}^{\text{train}}\|_F^2 \text{ subject to } \forall i,\ \|\boldsymbol{\alpha}_i^{\text{train}}\|_0 \le T_0 \tag{3}$$

where the notation$\|A\|_F$ stands for the Frobenius norm, with $\|A\|_F^2$ is defined as $\sum_i \sum_j |a_{i,j}|^2$.

The first stage (sparse coding stage), with dictionary $\boldsymbol{D}$, the representation vectors $\boldsymbol{\alpha}_i^{\text{train}}$ in Eqn. (3) are obtained using any pursuit algorithm [41]. 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:

$$\boldsymbol{E}_i = \boldsymbol{Y} - \sum\nolimits_{j\neq i} \boldsymbol{d}_j \boldsymbol{\alpha}_j^{\text{train}} \tag{4}$$

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

An Artificial Neural Network (NN) is a non-linear statistical approach to modelling complex relationships between input and output data [7]. A generic configuration of a NN normally contains an input layer, a number or hidden layers, and an output layer. Each layer is comprised of a number of 'neurons' or 'nodes', which are basic computational units that take an input vector, an intercept term b (or a bias unit), and compute an output via:

$$h_{\boldsymbol{W},b}(\boldsymbol{x}) = f\left(\sum\nolimits_{i=1}^{N} \boldsymbol{w}_i \boldsymbol{x} + b\right) \tag{5}$$

where $\boldsymbol{w}_i$ is the weight associated to neuron i and $f(\cdot)$ is an activation function which maps the output to a fixed range.

The SANN [31], [22] is a NN for efficient feature encoding where the aim is to learn a sparse and compressed representation for a set of training data. More specifically, SANN can reconstruct the training data with small reconstruction error using a small set of nodes in the hidden layer. Under the framework of SANN, we employ unsupervised model training 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 [7] can be used for training by minimising the following cost function [13]:

$$J(\boldsymbol{W}, \boldsymbol{b}) = J_{\text{error}} + J_{\text{weight}} + \beta J_{\text{sparsity}} \tag{6}$$

where

$$\begin{aligned} J_{\text{error}} &= \frac{1}{M}\sum_{i=1}^{M}\left(\frac{1}{2}\|\widehat{\boldsymbol{x}}_i - \boldsymbol{x}_i\|^2\right) & (7)\\ J_{\text{weight}} &= \frac{\lambda}{2}\|\boldsymbol{W}\|^2 & (8)\\ J_{\text{sparsity}} &= \sum_{i=1}^{N}\text{KL}\left(\rho \parallel \widehat{\rho}_i\right) & (9)\\ &= \sum_{i=1}^{N}\left[\rho\log\left(\frac{\rho}{\widehat{\rho}_i}\right) + (1-\rho)\log\left(\frac{1-\rho}{1-\widehat{\rho}_i}\right)\right] & (10) \end{aligned}$$

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 $\boldsymbol{x}_i$ [22]. The regularisation term J_{weight} decreases the magnitude of the weights to prevent overfitting. J_{sparsity} constrains the network to achieve low "activation", where $\text{KL}\left(\rho \parallel \widehat{\rho}_i\right)$ is the Kullback-Leibler divergence between ρ and $\widehat{\rho}_i$. The parameter ρ controls the degree of sparsity and $\widehat{\rho}_i$ is the average activation of hidden node i. The parameter β in Eqn. (6) controls the contribution of J_{sparsity} (typically equal to 3).

Given the trained SANN and a probe vector $\boldsymbol{x}$, the elements of the sparse code $\widehat{\boldsymbol{\alpha}} = [\widehat{\alpha}_1, \widehat{\alpha}_2, \cdots, \widehat{\alpha}_N]$ are calculated using:

$$\widehat{\alpha}_i = \text{sig}(\boldsymbol{w}_i^T\boldsymbol{x} + b_i) \qquad (11)$$

where $\boldsymbol{w}_i$ 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 described previously, SANN has the advantage of avoiding the minimisation problem during the sparse encoding stage, resulting in a lower computational cost.

C. Implicit Sparse Encoding via Probabilistic Approach

In the context of probabilistic modelling, vectors are assumed to be independent and identically distributed (this assumption is often incorrect but necessary to make the problem tractable [32]). By assuming the vectors obey a Gaussian distribution, all data can be modeled as a mixture of Gaussians or Gaussian Mixture Model (GMM). GMM is a parametric probability density function represented as a weighted sum of Gaussian component densities [10], [32], [8]. Given a probe vector $\boldsymbol{x}$ and a trained model with relatively large number of Gaussians, the normalised likelihood of $\boldsymbol{x}$ belonging to each Gaussian can be represented as a sparse code $\widehat{\boldsymbol{\alpha}}$ with:

$$\widehat{\boldsymbol{\alpha}} = \left[\frac{w_1 p\left(\boldsymbol{x}|\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1\right)}{\sum_{n=1}^{N} w_n p\left(\boldsymbol{x}|\boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n\right)}, \cdots, \frac{w_N p\left(\boldsymbol{x}|\boldsymbol{\mu}_N, \boldsymbol{\Sigma}_N\right)}{\sum_{n=1}^{N} w_n p\left(\boldsymbol{x}|\boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n\right)}\right] \qquad (12)$$

where

$$p\left(\boldsymbol{x}|\boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n\right) = \frac{\exp\left[-\frac{1}{2}\left(\boldsymbol{x}-\boldsymbol{\mu}_n\right)^T\boldsymbol{\Sigma}_n^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_n)\right]}{(2\pi)^{\frac{d}{2}}\left|\boldsymbol{\Sigma}_n\right|^{\frac{1}{2}}} \qquad (13)$$

is a multi-variate Gaussian function [8], [17]. The variables w_n, $\boldsymbol{\mu}_n$, and $\boldsymbol{\Sigma}_n$ are, respectively, the weight, mean vector and diagonal covariance for Gaussian n. The dictionary is trained by first initialising the mean vectors with a k-means clustering algorithm followed by the Expectation-Maximisation algorithm [17]. We note that most of the entries in sparse code $\widehat{\boldsymbol{\alpha}}$ are typically not exactly zero but are small enough to be treated as zero.

III. SR: Identification vs. Verification

In this section, we first briefly review the SR-based classification methodology for face identification problems. We then discuss why such methodology is not suitable for face verification problems and delineate a rudimentary extension to allow the use of SR with holistic descriptors in such problems. This rudimentary holistic approach is separate and distinct from using SR at the level of local patches.

A. Holistic SR for Face Identification

Consider a closed-set face identification problem with a gallery comprised of N samples. 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. (2). Using only the coefficients associated with the i-th class, Wright et al. [48] computed the residual, $r_i(\boldsymbol{x})$, using:

$$r_i(\boldsymbol{x}) = \|\boldsymbol{x} - \boldsymbol{D}\delta_i(\widehat{\boldsymbol{\alpha}})\|_2^2 \qquad (14)$$

where δ_i is a binary vector with the non-zero entries being associated 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 [50] and RSC [52].

B. Rudimentary Extension of Holistic 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 (ie. $\widehat{\boldsymbol{\alpha}}$) 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. (2). The similarity score between these descriptors can be calculated using:

$$s_{\text{SR}}(\boldsymbol{x}_a, \boldsymbol{x}_b|\boldsymbol{D}) = \text{dist}\left(\widehat{\boldsymbol{\alpha}}_a - \widehat{\boldsymbol{\alpha}}_b\right) \qquad (15)$$

where $\text{dist}(\cdot)$ is the distance function of choice, such as Euclidean or Hamming distance, with a smaller value indicating a higher similarity between $\boldsymbol{x}_a$ and $\boldsymbol{x}_b$. The classification decision (ie. whether $\boldsymbol{x}_a$ and $\boldsymbol{x}_b$ represent the same person) can be obtained by comparing s_{SR} to a decision threshold.

In the above approach, the sparse solutions can be obtained from holistic face representations, such as PCA-based feature extraction [44]. We therefore denote this approach as *holistic SR descriptor*.

IV. Locally Sparse Encoded Descriptor

In the previous section, we have shown an extension of holistic SR to verification problems. However, as shown later, the holistic SR descriptor delivers poor performance. In this section, we present an alternative way to utilise sparse coding in verification problems. Motivated by the benefits of local feature-based face representation and BoW approaches, we introduce a face descriptor termed Locally Sparse Encoded Descriptor (LSED), which can be seen as an extension of MRH [36]. In addition to the implicit probabilistic encoding used in the original MRH formulation, we propose to use two more direct sparse encoding techniques: l_1-minimisation and SANN, described in Sections II-A and II-B. We continue this section by first describing the face encoding framework, followed by brief discussions on the characteristics of each sparse encoding technique. We then elaborate on how the descriptor can be used for discriminating faces.

A. Framework

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.

From each normalised patch $\widehat{\boldsymbol{p}}_{r,i}$, a low dimensional texture descriptor, $\boldsymbol{x}_{r,i}$, is obtained via 2D DCT decomposition [21]. Preliminary experiments suggest that patches of size 8×8 pixels with 75% overlap (ie. adjacent patches are overlapped by either 6×8 or 8×6 pixels) lead to good performance [36]. Moreover, we selected the 15 lowest frequency components of the DCT coefficients, with the zeroth coefficient discarded (as it has no information due to the aforementioned normalisation step). We note that it is also possible to use other texture descriptors, such as raw pixels, Gabor wavelets [27] and Local Binary Patterns [2]. Preliminary experiments suggest that the DCT-based texture descriptors lead to better performance.

Each i-th texture descriptor from region r, $\boldsymbol{x}_{r,i}$, is then described by a sparse code $\widehat{\boldsymbol{\alpha}}_{r,i}$. In the original formulation of MRH [36], the sparse code is implicitly generated using the probabilistic encoding approach elaborated in Eqn. (12). Having each patch represented by a sparse code, each region r is then described via the following pooling strategy:

$$\boldsymbol{h}_r = \frac{1}{N_p} \sum\nolimits_{i=1}^{N_p} \widehat{\boldsymbol{\alpha}}_{r,i} \tag{16}$$

where $\widehat{\boldsymbol{\alpha}}_{r,i}$ is the i-th sparse vector in region r and N_p is the number of patches in region r. Due to the averaging operation, in each region there is a loss of spatial relations between face parts. As such, each region is in effect described by an orderless collection of local descriptors. A conceptual diagram of the framework is shown in Figure 1.

We propose to use two other sparse encoding techniques to generate the sparse code $\widehat{\boldsymbol{\alpha}}_{r,i}$, namely, l_1-minimisation (using Eqn. (2)) and SANN (using Eqn. (11)). For the l_1-minimisation based encoding, the generated sparse codes may consist of negative coefficients, which causes a problem with the averaging pooling strategy in Eqn. (16). To address this, the patch level sparse codes can be obtained with nonnegative encoding [9] or by splitting the positive and negative coefficients into two sparse codes followed by vector concatenation [14]. In preliminary experiments we found that the most robust performance can be obtained by simply applying an absolute function to each patch level sparse code.

The dictionary used by each sparse encoding approach is described in Section II. Examples of LSED with the three sparse encoding techniques are shown in Figure 2, where LSED with probabilistic encoding is the sparsest at both the patch level and the region level, whereas the SANN-based encoding produces relatively noisier descriptors while maintaining a good degree of sparsity. We discuss the differences of the encoding techniques below.

B. Characteristics of Sparse Encoding

In Section II, we presented three sparse encoding approaches (ie. l_1-minimisation, SANN and probabilistic encoding). We note that there are some fundamental differences between the approaches.

The probabilistic approach computes the normalised likelihood using each Gaussian in the GMM, which indirectly models each patch as a sparse vector. The sparsity in this case stems from a very small subset of the Gaussians (typically 2 or 3) being close to a given sample. The close Gaussians provide high normalised likelihoods, while the remaining Gaussians have likelihoods that are close to zero.

In contrast, the l_1-minimisation approach solves an optimisation problem based on the reconstruction error (ie. reconstruct a given patch as a linear combination of dictionary atoms), with the optimal solution obtained for each patch. The SANN-based approach uses a similar objective (ie. patch reconstruction). However, it avoids minimisation of the reconstruction error for each patch [31]. The sparse solution for any given local patch is obtained by feeding the given patch into the SANN, which is a very fast process that consists of straightforward linear algebra. SANN assumes that the training samples provide the generic distribution of the data and the optimisation is performed only on the training samples. As such, this encoding approach may not deliver the optimal solution for any given patch.

C. Similarity-Based Classification

Comparison between two faces is accomplished by comparing their corresponding regional descriptors. Using the method from [36], the matching score between faces A and B can be calculated via:

$$s_{\texttt{raw}}(A, B) = \frac{1}{R} \sum\nolimits_{r=1}^{R} \left\| \boldsymbol{h}_r^{[A]} - \boldsymbol{h}_r^{[B]} \right\|_1 \tag{17}$$

where R is the number of regions. To account for uncontrolled image conditions not already handled by the patch-based analysis, a cohort normalisation [16], [36] based distance can be employed:

$$s_{\texttt{norm}}(A, B) = \frac{s_{\texttt{raw}}(A, B)}{\sum_{i=1}^{N_C} s_{\texttt{raw}}(A, C_i) + \sum_{i=1}^{N_C} s_{\texttt{raw}}(B, C_i)} \tag{18}$$

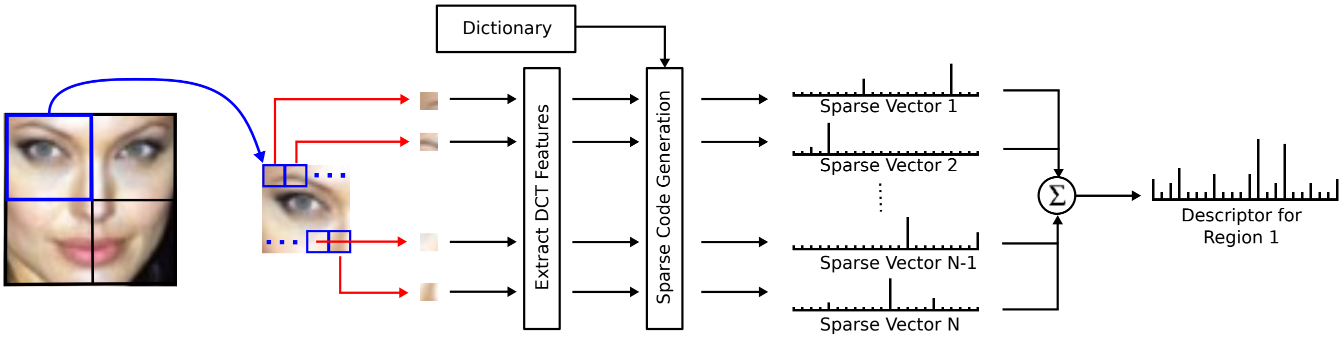

Fig. 1. Conceptual demonstration of the LSED framework. A given face image is divided into regions, followed by breaking each region into smaller patches. For each patch, a sparse vector is obtained by a sparse encoder using a learned dictionary. Each regional face descriptor is computed by pooling the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

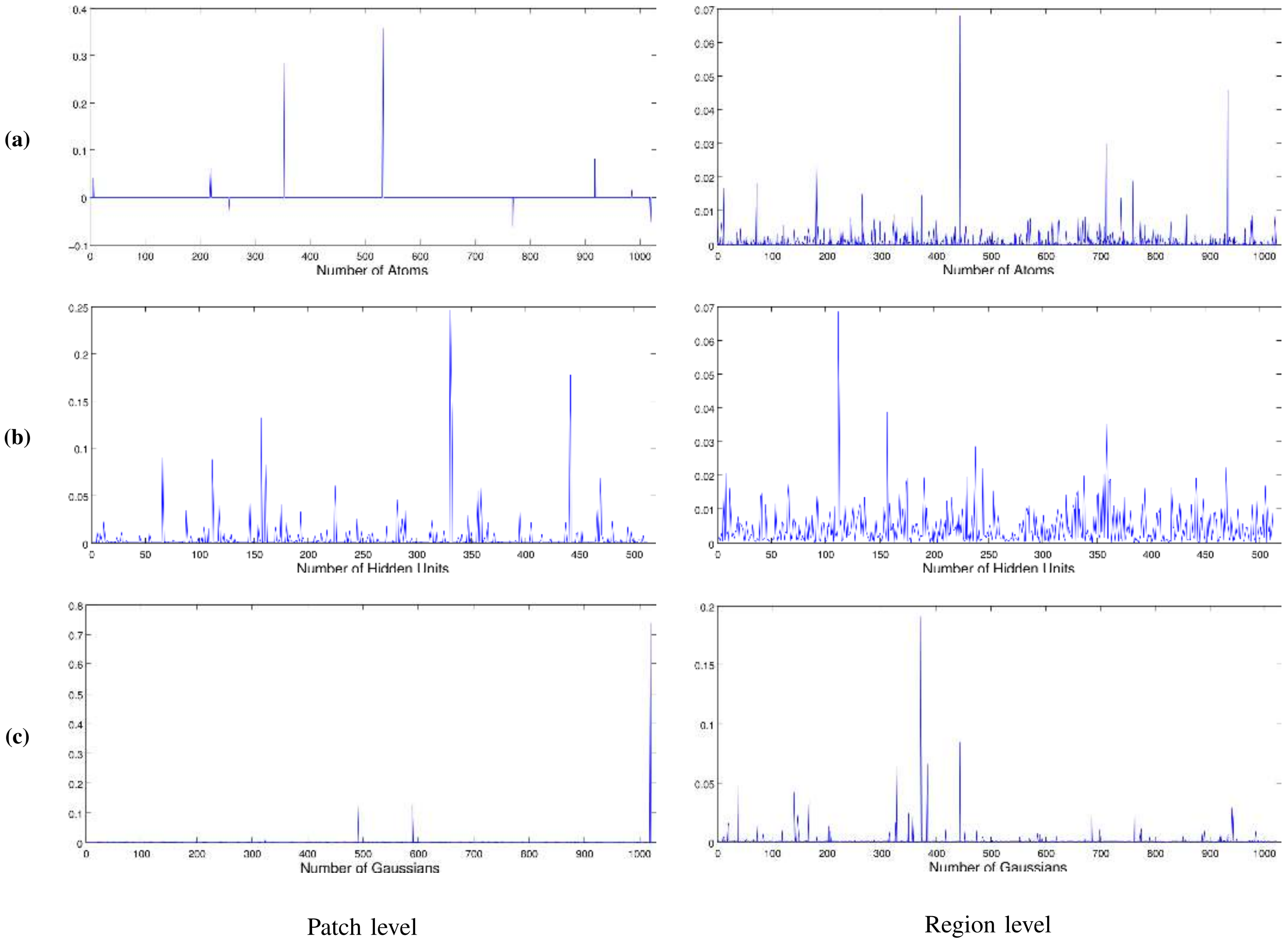

Fig. 2. Left column: examples of sparse codes for a single patch. Right column: examples of resultant region descriptors obtained via the pooling strategy in Eqn. (16). Three sparse encoding approaches are shown: **(a)** l_1-minimisation, **(b)** Sparse Autoencoder Neural Network, **(c)** probabilistic. For l_1-minimisation based encoding, an absolute function is applied to each patch level code prior to applying the pooling strategy. For probabilistic encoding, most of the coefficients are not exactly zero but are small enough to be treated as zero.

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 faces A and B belong to the same person, $s_{\texttt{norm}}(A, B)$ can be compared to a decision threshold.

V. Experiments with Still Images

In this section, we examine the performance of LSED on several identity inference configurations: **(a)** verification with various face alignment errors and sharpness variations, **(b)** verification with pose mismatches, **(c)** verification with controlled and uncontrolled images, and **(d)** closed-set identification. We also evaluate the computational cost for LSED generation as well as the query time in closed-set identification problems. In addition, we use synthetic data to demonstrate the weakness of the holistic SR descriptor (from Section III-B) on verification problems.

Experiments were conducted on five datasets: FERET [30], AR [29], BANCA [4], exYaleB [27], and ChokePoint [46]. Figure 3 shows example raw images. In all experiments, we used closely cropped face images with a size of 64×64 pixels. Each image was manually aligned so that the eyes were at fixed positions, except for experiments with simulated image variations. See Figure 4 for examples.

In the following subsections, we denote the original formulation of MRH with probabilistic encoding as LSED+prob. Forms of LSED with the Sparse Autoencoder Neural Network and l_1-minimisation based encoding approaches are denoted as LSED+SANN and LSED+l_1, respectively. The LSED framework has a number of parameters that affect performance. Based on preliminary experiments, we split each image into 3×3 regions and used 32 cohorts for the distance normalisation in Eqn. (18). LSED+SANN has 512 hidden units, where the parameters of the cost function (Eqns. (6) and (10)) were set as $\beta = 3$, $\rho = 0.1$, and $\lambda = 0.01$. LSED+prob and LSED+l_1 have a dictionary with 1024 Gaussians/atoms. The threshold for reconstruction error, ϵ, in Eqn. (2) was set to 0.1. These parameters were kept unchanged for all experiments.

Unless otherwise specified, all experiments were implemented in MATLAB using an in-house implementation. The l_1-minimisation problem was solved with SparseLab[2].

A. Face Verification 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 [20] as the training set. The CAS-PEAL dataset provides 1200 face images from 1200 unique individuals. Note that the face images for cohort normalisation are selected from the training set. The development and evaluation sets have a balanced number of matched and mismatched pairs.

Using the development set and the normalised matching scores from Eqn. (18), 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 (ie., the so-called Equal Error Rate point [16]). The threshold was then applied on the evaluation set, with the final accuracy defined as $1 - \frac{1}{2}(\mathrm{FAR} + \mathrm{FRR})$. The threshold was deliberately not found on the evaluation set as in real-life conditions it has to be selected *a priori* [10], [6].

In all experiments, we compared LSED with the holistic SR descriptor described in Section III-B. We used the holistic SR descriptor in conjunction with two feature extraction methods: **(1)** PCA based [5] (denoted as PCA+SR), and **(2)** Gabor based [28] (denoted as Gabor+SR). Based on preliminary experiments, the similarity scores between two PCA+SR descriptors were calculated via Hamming distance measurement, whereas Euclidean distance was preferred for Gabor+SR. Gabor based feature extraction followed the configuration in [50], with PCA based dimensionality reduction. For both feature extraction methods, PCA preserved 99% of the total energy.

We also evaluated verification performance of three baseline holistic face descriptors (ie., without sparse encoding): **(1)** PCA based (denoted as PCA), **(2)** Local Binary Patterns [2] (denoted as LBP), and **(3)** Gabor based (denoted as Gabor). The similarities between two face descriptors were calculated using Euclidean distance measurement.

1) Face Verification with Alignment Errors and Blurring: In this section, we evaluate the robustness of LSED on blurring, as well as on four alignment errors using images taken from the 'fb' subset of FERET. Example images are shown in Figure 5. The generated alignment errors[3] are: horizontal shift and vertical shift (using displacements of ±2, ±4, ±6, ±8 pixels), in-plane rotation (using rotations of $\pm10°$, $\pm20°$, $\pm30°$), 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 7, show that the three LSED approaches consistently achieved robust performance in all simulated scenarios. LSED+SANN and LSED+prob achieved average accuracies of 85.8% and 86.2%, respectively, whereas LSED+l_1 led the performance with an average accuracy of 89.2%. Overall, the accuracy of LSED+l_1 is about 12.2 percentage points better than the baseline Gabor approach and about 23.7 percentage points when compared to Gabor+SR. The results also show that PCA+SR and Gabor+SR performed poorly on all misalignment errors, with overall accuracies of 68.2% and 65.6%, respectively. The results suggest that scale changes and in-plane rotation variations are in general the hardest problems out of all alignment errors.

[2] SparseLab is available at http://sparselab.stanford.edu/

[3] The generated alignment errors are representatives of real-life characteristics of automatic face localisation/detection algorithms [33].

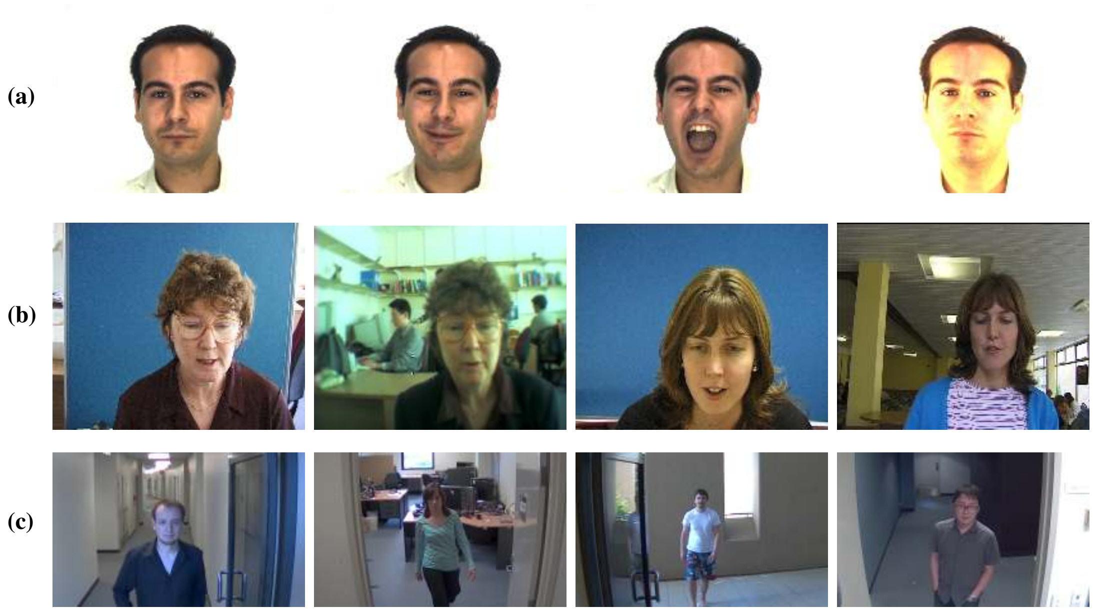

Fig. 3. Example raw images from several datasets. **(a)** The AR dataset contains 14 images per subject with various expressions and lighting conditions. **(b)** The BANCA dataset: each subject was recorded under 3 scenarios: *controlled* (columns 1 & 3), *degraded* (column 2), and *adverse* (column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

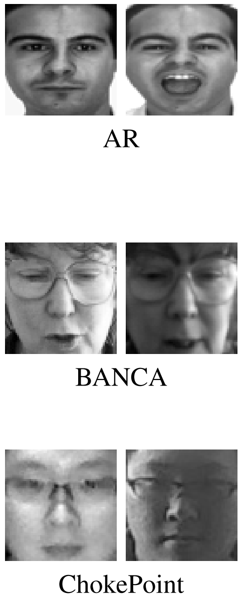

Fig. 4. Examples of cropped images.

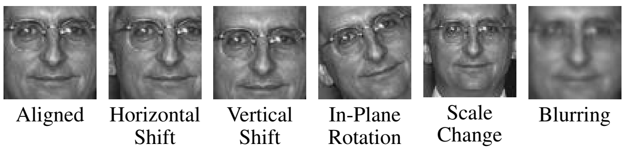

Fig. 5. Examples of simulated image variations on FERET.

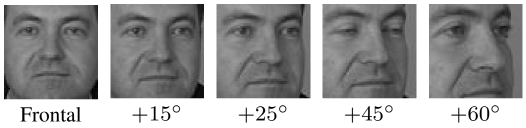

Fig. 6. Examples of the FERET pose subset.

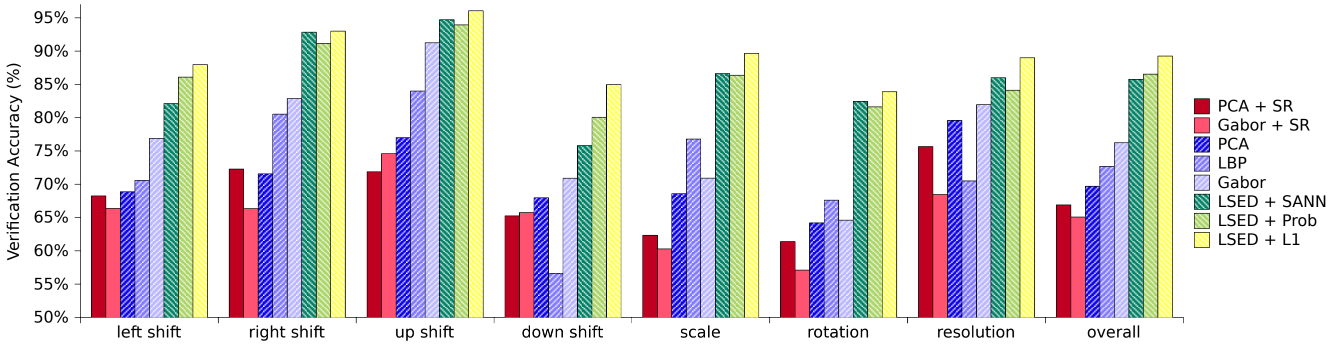

Fig. 7. The average verification accuracy on FERET images with stimulated alignment errors and sharpness variations (demonstrated in Fig. 5). Experiments were conducted with 5-fold validations.

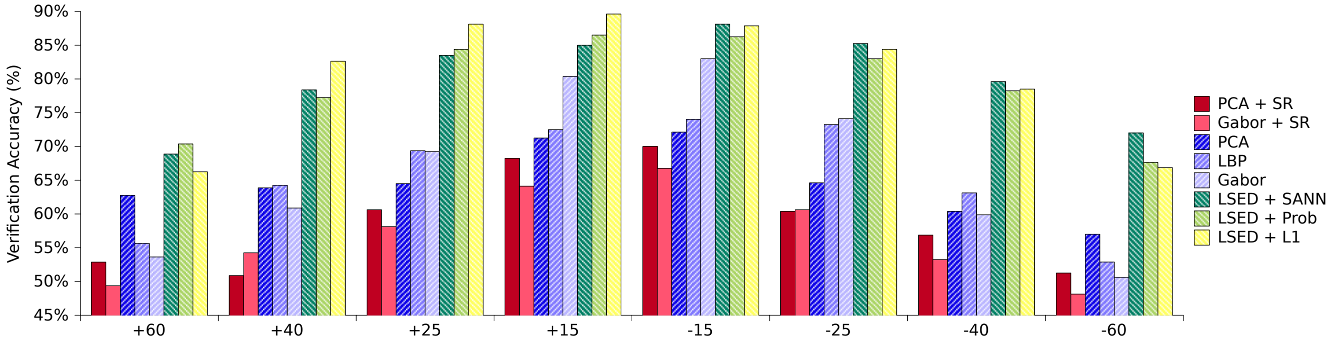

Fig. 8. Verification performance on pose mismatches for various angles. Faces from each pose angle are compared with the FERET frontal subset 'ba' and the expression subset 'bj'. Experiments were conducted with 5-fold validations.

2) Face Verification with Pose Mismatches: In this section we evaluate the robustness of LSED 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 8, indicate that the three LSED approaches considerably outperforms both PCA+SR and Gabor+SR. Both of the holistic SR descriptors obtained a maximum accuracy of 56.9% when the absolute value of the pose angle was $\geq 40^\circ$. In contrast, LSED+l_1 achieved an average accuracy of 73.6% under the same pose angles. Note that the LSED+l_1 was outperformed by LSED+SANN and LSED+prob for pose angles of $\pm 60^\circ$. When the pose angle was between -25° and $+25^\circ$ (ie. relatively frontal) the best performing holistic SR descriptor (PCA+SR) achieved an average accuracy of about 64.8%. All LSED approaches outperformed the holistic SR descriptors by a comfortable margin on the same range of pose angles, with LSED+l_1 obtaining an average accuracy of 87.5%.

3) 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 4. 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 [4]. 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 [46], which was recorded under real-world surveillance conditions. It has 16 videos of 29 subjects recorded on four distinct portals[4]. We randomly generated 38,710 matched and mismatched image pairs where each pair consisted of images taken from different portals (ie. cross environment matching). The experiments were evaluated with 5-fold validations.

The results, presented in Table I, show that the three LSED methods obtained the best overall performance. Both PCA+SR and Gabor+SR performed at their best on the laboratory captured AR dataset and considerably poorer on the more realistic ChokePoint dataset. The results also show that both the baseline LBP and Gabor methods outperformed the holistic SR descriptors. For example, the baseline Gabor approach obtained an overall accuracy of 73.2%, outperforming its sparse counterpart (Gabor+SR) which obtained an overall accuracy of 63.2%.

LSED+l_1 achieved the best overall accuracy of 80.7%. On the controlled AR dataset, The baseline LBP method outperformed both LSED+SANN and LSED+prob by 5.7 and 1.1 percentage points, respectively. However, the performance of LBP dropped considerably on both the BANCA and ChokePoint datasets, where LSED+prob outperformed LBP by 15.1 and 11.5 percentage points on the corresponding datasets. This indicates that while the LSED framework can be outperformed by baseline holistic methods in controlled conditions, LSED is more robust for face images obtained in uncontrolled conditions.

TABLE I

FRONTAL FACE VERIFICATION PERFORMANCE ON SEVERAL DATASETS. THE FACE IMAGES WERE CLOSELY CROPPED TO EXCLUDE HAIR AND BACKGROUND, AND SCALED TO 64×64 PIXELS. THE VALUES IN **bold** INDICATE THE BEST PERFORMING ALGORITHM FOR EACH DATASET.

Method	AR	BANCA	ChokePoint	Overall
PCA + SR	61.4%	58.8%	57.4%	59.4%
Gabor + SR	66.1%	63.3%	59.5%	63.2%
PCA	57.3%	63.5%	55.6%	59.0%
LBP	77.9%	60.3%	65.3%	68.1%
Gabor	74.5%	70.0%	75.6%	73.2%
LSED + SANN	72.2%	73.4%	75.1%	73.5%
LSED + prob	76.8%	75.4%	76.8%	76.3%
LSED + l_1	**80.0%**	**82.0%**	**79.8%**	**80.7%**

4) Experiments with Synthetic Data: The results obtained in the preceding sections indicate that holistic SR descriptors were consistently outperformed by baseline holistic face descriptors (ie., without sparse coding). In this section, we performed a set of verification experiments with synthetic data to study this phenomenon further.

We explicitly created a dictionary $\boldsymbol{D}$ which does not satisfy the underlying *sparsity* assumption. Each sample from the synthetic data is assumed to be a holistic representation of a face. The synthetic data comprised of 232 random classes, with the samples in each class obeying a normal distribution. The dimensionality of data was 16. For each class 128 samples were generated. We randomly selected 32 classes as the training set and the remaining 200 classes as the development set and evaluation set. The training set played the role of dictionary $\boldsymbol{D}$ in Eqn. (15). The experiments were conducted with 5-fold validations.

Several verification experiments with increasing difficulty were generated by fixing the mean of each class and increasing the class variance. The distribution of the class means was carefully controlled such that at the smallest class variance the mutual overlaps between classes are close to zero. We employed direct feature matching as the baseline. In other words, for two given samples, $\boldsymbol{x}_a$ and $\boldsymbol{x}_b$, the matching score is the Euclidean distance $\|\boldsymbol{x}_a - \boldsymbol{x}_b\|_2$. The holistic SR descriptor was evaluated with Hamming distance measurement, as this led to somewhat better performance than using the Elucidean distance. The Hamming distance compares two descriptors by measuring if the corresponding descriptors have the same set of nonzero entries. In other words, Hamming distance explicitly inspects if both descriptors are spanned by a set of common subspaces.

The results in Figure 9 show that the baseline performance is close to 100% when the class variance is small, and drops to 53.5% when variance is at its maximum value. In contrast, the holistic SR descriptor achieved poorer performance across

[4] 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 conditions.

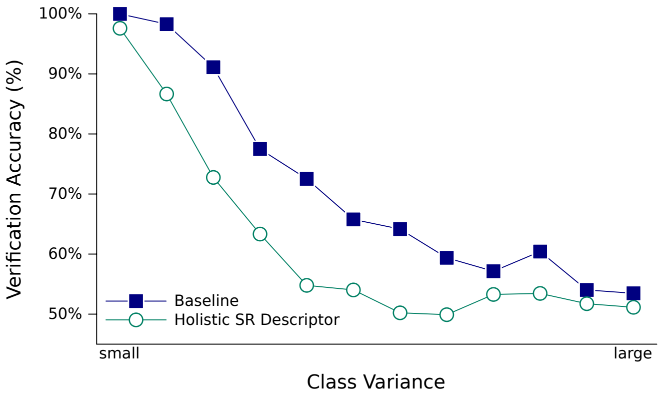

Fig. 9. Verification performance on synthetic data. Experiments were conducted with varying class variance, where large class variance indicates strong overlap between classes. The baseline was achieved by matching each feature pair using Euclidean distance. Experiments were conducted with 5-fold validations.

the variance range, with accuracies of 97.6% and 51.1% respectively for minimum and maximum class variance. This result agrees with our discussion in Section III and the findings from the preceding face verification experiments. Specifically, if the class information of the atoms is not given and the sparsity assumption does not hold for the dictionary $\boldsymbol{D}$, the resulting sparse solutions do not provide good discriminative ability when compared to the original holistic representation.

B. Face Identification Experiments

In the preceding set of experiments, we demonstrated that the proposed LSED framework outperforms holistic SR descriptors on various face verification problems. In this section, we evaluate the efficacy of LSED in closed-set face identification, which is the identity inference configuration typically used in SR related literature.

LSED was compared with five established holistic SR based classification algorithms: **(i)** SR with PCA feature extraction (denoted as PCA+SRC) [48], **(ii)** SR with PCA feature extraction and LDA (denoted as LDA+SRC) [48], **(iii)** SR with Gabor feature extraction (denoted as Gabor+SRC) [50], **(iv)** Robust Sparse Coding with PCA feature extraction (denoted as RSC) [52], and **(v)** orthonormal l_2-norm approach with vectorised raw image [37] (denoted as raw+l_2). Instead of solving an optimisation problem, raw+l_2 estimates the sparse code $\boldsymbol{\alpha}$ using $\boldsymbol{\alpha} = \boldsymbol{R}^{-1}\boldsymbol{Q}^T\boldsymbol{x}$, where $\boldsymbol{Q}$ and $\boldsymbol{R}$ are the result of QR factorisation [39] of dictionary $\boldsymbol{D}$.

The experiments were conducted on AR, exYaleB and ChokePoint datasets, with each gallery having 7, 16, and 16 images per class, respectively. To increase the difficulty, the gallery of the ChokePoint dataset was selected from a portal different than the portal used for the query images. Each portal has a unique background and illumination conditions. The identification performance of LSED was obtained with the Nearest Neighbour classifier. Note that the results shown for the established SR algorithms are slightly different from the literature, due to the image size and dataset splits being different.

TABLE III
AVERAGE COMPUTATION TIME FOR GENERATING A LOCALLY SPARSE ENCODED DESCRIPTOR FOR ONE IMAGE.

Method	Time (milliseconds)
LSED + SANN	110
LSED + prob	2021
LSED + l_1	7739

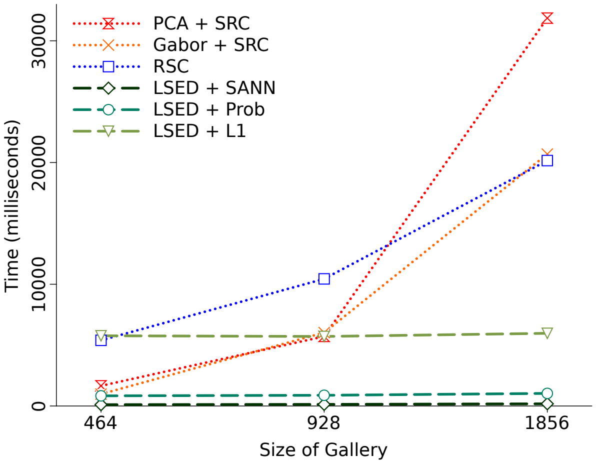

Fig. 10. Average computation time (including feature extraction, sparse encoding and identification) for matching a probe image against galleries of various sizes.

The results, shown in Table II, indicate that LSED+prob and LSED+l_1 consistently outperformed all SRC algorithms in closed-set identification. The improvement on the ChokePoint dataset is the most notable among the three datasets, where LSED+l_1 outperformed the closest SRC algorithm (ie. RSC) by 20.5 percentage points. It also outperformed the raw+l_2 approach by 9 percentage points.

C. Computation Time

The preceding verification and identification experiments indicate that the LSED+l_1 technique achieves the best overall performance. However, the superior performance of LSED+l_1 comes at the expense of considerably higher computational cost. As shown in Table III, LSED+l_1 requires 7739 milliseconds (ms) to generate a single face descriptor, mainly due to solving multiple expensive l_1-minimisation problems (one for each small patch). In contrast, LSED+SANN is approximately 70 times faster, as it requires only 2.9ms to generate the entire face descriptor. LSED+prob, which achieved the closest performance to LSED+l_1, requires approximately a quarter of time when compared with LSED+l_1. We note that the computation cost for all three LSED variants can be considerably reduced via parallelisation, as each patch can be processed independently prior to the pooling operation in Eqn. (16).

Other than the computational cost generating each face descriptor, the cost to match one probe against a large gallery is also important. Using galleries with various amount of face images, we evaluated the average time to recognise a single probe using a closed-set identification setup. For each method,

TABLE II

Closed-set identification performance of the proposed method and various SR based approaches. The values in brackets are the number of images per class in the gallery. The values in **bold** indicate the best performing algorithm for each dataset.

Method	AR (7)	exYaleB (16)	ChokePoint (16)	Overall
PCA + SRC [48]	81.0%	67.9%	17.5%	52.1%
LDA + SRC [48]	89.7%	52.8%	65.3%	64.9%
Gabor + SRC [50]	91.7%	61.4%	63.4%	68.3%
RSC [52]	95.7%	72.8%	64.5%	74.4%
raw + l_2 [37]	90.3%	75.1%	76.0%	78.5%
LSED + SANN	96.3%	66.0%	77.0%	76.2%
LSED + prob	97.9%	76.7%	80.5%	82.4%
LSED + l_1	**98.9%**	**90.9%**	**85.0%**	**90.4%**

we measured the time for feature extraction, sparse encoding (or approximate sparse solution) and identification. The raw+l_2 method is not included in this evaluation as it does not solve an optimisation problem.

The results, shown in Figure 10, indicate that the identification time for LSED framework is almost constant. In contrast, the computational cost of traditional SRC-based methods and RSC increased considerably as the gallery size increased.

VI. Experiments with Image Sets

In the previous section we presented experiments using a single face image per person at a time. In contrast, in this section we evaluate the verification performance of LSED using multiple images per person at a time. This recognition task is also known as image set matching, with the aim of determining if two face sets, $\mathbb{A}$ and $\mathbb{B}$, belong to the same person.

We first describe two image set matching approaches (Hausdorff distance and mean descriptors), followed by presenting results on BANCA and ChokePoint datasets. We also contrast the computational costs of the two matching approaches.

A. Image Set Matching via Hausdorff Distance

Given two finite image sets, $\mathbb{A} = \{a_1, a_2, \ldots, a_{N_{\mathbb{A}}}\}$ and $\mathbb{B} = \{b_1, b_2, \ldots, b_{N_{\mathbb{B}}}\}$, the Hausdorff distance is defined as:

$$H(\mathbb{A}, \mathbb{B}) = \max \{ \, h(\mathbb{A}, \mathbb{B}), h(\mathbb{B}, \mathbb{A}) \, \} \tag{19}$$

where

$$h(\mathbb{A}, \mathbb{B}) = \max_{i \in \mathbb{A}} \{ \, \min_{j \in \mathbb{B}} \{ s(a_i, b_j) \} \, \} \tag{20}$$

and $s(\cdot)$ measures the similarity between two images. The function $h(\mathbb{A}, \mathbb{B})$ is called the directed Hausdorff distance from $\mathbb{A}$ to $\mathbb{B}$. In general, if the Hausdorff distance between image set $\mathbb{A}$ and $\mathbb{B}$ is d, each image in $\mathbb{A}$ is within distance d to some of the points in $\mathbb{B}$, and vice-versa [25].

B. Image Set Matching with Mean Descriptors

The Hausdorff distance measurement is a computationally expensive approach for image set matching. This is in particularly a problem for video surveillance of public spaces, where the volume of surveillance video can be very high. To address this problem, each image set can be represented by an overall descriptor via straightforward averaging of the corresponding face descriptors [11]. Specifically, given descriptors from image set $\mathbb{A}$, the mean descriptor is represented as $\frac{1}{N_{\mathbb{A}}} \sum_{n=1}^{N_{\mathbb{A}}} \boldsymbol{h}_{\mathbb{A},n}$, where $\boldsymbol{h}_{\mathbb{A},n}$ is the n-th descriptor of $\mathbb{A}$. The similarity between two mean descriptors can be then computed using Eqn. (18).

In contrast to image set matching using the Hausdorff distance, the total number comparisons between $\mathbb{A}$ and $\mathbb{B}$ is reduced from $N_{\mathbb{A}} \times N_{\mathbb{B}}$ to one.

C. Results

We evaluate image set matching performance on two datasets, with images captured under uncontrolled environment conditions. The first dataset is BANCA dataset, where we randomly generate 900 pairs of matched and mismatched pairs, and each image-set contains 9 face images. The experiments were evaluated with 5-fold validations. The second dataset is the ChokePoint video dataset. We selected 16 images with the highest quality as per [46] and randomly generated 5000 matched and mismatched pairs. The experiments were evaluated using 10-fold validations. For comparison, we used the same face descriptor methods as in Section V-A. The results are shown in Figure 11.

On the BANCA dataset, the performance of the three LSED approaches is very similar for both the Hausdorff and mean descriptor matching approaches. Among the LSED variants, LSED+l_1 in conjunction with mean descriptor matching obtains the highest accuracy, with the computationally less expensive LSED+prob variant not far behind. The performance of baseline LBP and Gabor approaches are considerably lower than LSED+l_1. PCA+SR and Gabor+SR achieved poor performance for both the Hausdorff and mean descriptor matching approaches.

The verification performance on the ChokePoint dataset has two notable differences when compared to the performance on the BANCA dataset. All LSED variants achieved notably better performance using the mean descriptor matching approach rather than the Hausdorff distance based approach. Secondly, the traditional PCA approach obtained the worst verification accuracy among all face descriptors, with PCA+SR outperforming it by 4.5 percentage points when using the Hausdorff distance. The poor performance is mainly due to image quality variations (ie. stemming from surveillance environments), which was also shown in the face identification experiments in Section V-B.

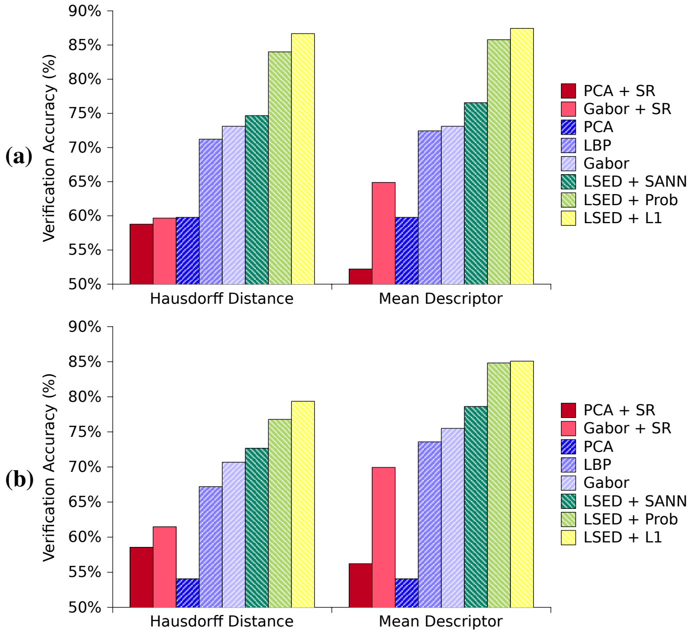

Fig. 11. Image set verification performance on **(a)** BANCA dataset with 5-fold validations, and **(b)** ChokePoint dataset with 10-fold validations.

TABLE IV

AVERAGE TIME FOR MATCHING TWO IMAGE SETS. EACH IMAGE SET CONTAINS 32 IMAGES. THE EXPERIMENTS WERE CONDUCTED WITH LSED + l_1, WHERE THE DIMENSIONALITY OF EACH DESCRIPTOR IS 9216.

Method	**Time (milliseconds)**
Matching via Hausdorff distance	1018
Matching via mean descriptor	7

The approximate computational cost for the mean and Hausdorff matching approaches, using LSED+l_1 descriptors, is shown in Table IV. The straightforward mean descriptor approach is approximately 2 orders of magnitude faster than the computationally intensive Hausdorff approach, while obtaining similar or better results.

VII. MAIN FINDINGS

Most of the literature on Sparse Representation (SR) for face recognition has focused on holistic face descriptors in closed-set identification applications. The underlying assumption in SR-based methods is that each class in the gallery has sufficient samples and the query lies on the subspace spanned by the gallery of the same class. Unfortunately, such assumption is easily violated in the more challenging face verification scenario, where an algorithm is required to determine if two faces (where one or both have not been seen before) belong to the same person.

We first discussed why previous attempts with SR might not be applicable to verification problems. We then proposed an alternative approach to face verification via SR. Specifically, we proposed to use explicit SR encoding on local image patches rather than the entire face. The obtained sparse signals are pooled via averaging to form multiple region descriptors, which are then concatenated to form an overall face descriptor. Due to the deliberate loss spatial relations within each region (caused by averaging), the resulting descriptor is robust to misalignment and various image deformations. Within the proposed framework, we evaluated several SR encoding techniques: l_1-minimisation, Sparse Autoencoder Neural Network (SANN), and an implicit probabilistic technique based on Gaussian Mixture Models.

Thorough experiments on AR, FERET, exYaleB, BANCA and ChokePoint datasets show that the proposed local SR approach obtains considerably better and more robust performance than several previous state-of-the-art holistic SR methods, in both verification and closed-set identification problems. The proposed approach is particularly suited to dealing with face images obtained in difficult conditions, such as surveillance environments. The experiments also show that l_1-minimisation based encoding has a considerably higher computational cost when compared to SANN-based and probabilistic encoding, but leads to higher recognition rates.

VIII. ACKNOWLEDGEMENTS

NICTA is funded by the Australian Government as represented by the Department of Broadband, Communications and the Digital Economy, as well as the Australian Research Council through the ICT Centre of Excellence program.

REFERENCES

[1] M. Aharon, M. Elad, and A. Bruckstein. K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Trans. Signal Processing*, 54(11):4311–4322, 2006.

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

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

[4] 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)*, volume 2688, pages 625–638, 2003.

[5] 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*, 19(7):711–720, 1997.

[6] S. Bengio and J. Mariéthoz. The expected performance curve: A new assessment measure for person authentication. In *Proc. Odyssey 2004: The Speaker and Language Recognition Workshop*, pages 279–284, 2004.

[7] C. M. Bishop. *Neural Networks for Pattern Recognition*. Oxford University Press, 1st edition, 1995.

[8] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[9] A. M. Bruckstein, M. Elad, and M. Zibulevsky. On the uniqueness of nonnegative sparse solutions to underdetermined systems of equations. *IEEE Transactions on Information Theory*, 54(11):4813–4820, 2008.

[10] F. Cardinaux, C. Sanderson, and S. Bengio. User authentication via adapted statistical models of face images. *IEEE Trans. Signal Processing*, 54(1):361–373, 2006.

[11] S. Chen, S. Mau, M. T. Harandi, C. Sanderson, A. Bigdeli, and B. C. Lovell. Face recognition from still images to video sequences: A local-feature-based framework. *EURASIP Journal on Image and Video Processing*, 2011.

[12] S. S. Chen, D. L. Donoho, and M. A. Saunders. Atomic decomposition by basis pursuit. *SIAM Review*, 43(1):129–159, 2001.

[13] A. Coates, H. Lee, and A. Y. Ng. An analysis of single-layer networks in unsupervised feature learning. *Journal of Machine Learning Research - Proceedings Track*, 15:215–223, 2011.

[14] 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*, pages 921–928, June 2011.

[15] 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*, pages 1–22, 2004.

[16] 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*, 31(2-3):225–254, 2000.

[17] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. Wiley, 2 edition, 2001.

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

[19] R. A. Fisher. The use of multiple measurements in taxonomic problems. *Annals Eugen.*, 7:179–188, 1936.

[20] 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*, 38(1):149–161, 2008.

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

[22] 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*, pages 646–654, 2009.

[23] 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*, 81(2):191–204, 2009.

[24] B. Heisele, P. Ho, J. Wu, and T. Poggio. Face recognition: component-based versus global approaches. *Computer Vision and Image Understanding*, 91(1-2):6–21, 2003.

[25] D. P. Huttenlocher, G. A. Klanderman, and W. Rucklidge. Comparing images using the Hausdorff distance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(9):850–863, 1993.

[26] 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*, pages 2169–2178, 2006.

[27] K.-C. Lee, J. Ho, and D. J. Kriegman. Acquiring linear subspaces for face recognition under variable lighting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(5):684–698, 2005.

[28] C. Liu and H. Wechsler. Gabor feature based classification using the enhanced Fisher linear discriminant model for face recognition. *IEEE Transactions on Image Processing*, 11(4):467–476, 2002.

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

[30] 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*, 16(5):295–306, 1998.

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

[32] D. A. Reynolds. Gaussian mixture models. In *Encyclopedia of Biometrics*, pages 659–663. Springer, 2009.

[33] Y. Rodriguez, F. Cardinaux, S. Bengio, and J. Mariéthoz. Measuring the performance of face localization systems. *Image and Vision Computing*, 24(8):882–893, 2006.

[34] R. Rubinstein, A. M. Bruckstein, and M. Elad. Dictionaries for sparse representation modeling. *Proceedings of the IEEE*, 98(6):1045–1057, June 2010.

[35] C. Sanderson, S. Bengio, and Y. Gao. On transforming statistical models for non-frontal face verification. *Pattern Recognition*, 39(2):288–302, 2006.

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

[37] 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*, pages 553–560, 2011.

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

[39] L. N. Trefethen and D. Bau. *Numerical Linear Algebra*. SIAM: Society for Industrial and Applied Mathematics, 1997.

[40] J. A. Tropp. Greed is good: algorithmic results for sparse approximation. *IEEE Transactions on Information Theory*, 50(10):2231–2242, 2004.

[41] J. A. Tropp, A. C. Gilbert, and M. J. Strauss. Algorithms for simultaneous sparse approximation. Part I: Greedy pursuit. *Signal Processing*, 86(3):572–588, 2006.

[42] J. A. Tropp and S. J. Wright. Computational methods for sparse solution of linear inverse problems. *Proceedings of the IEEE*, 98(6):948–958, June 2010.

[43] 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*, volume 6562, 2007.

[44] M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3(1):71–86, 1991.

[45] 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*, 34(2):372–386, 2012.

[46] 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)*, pages 74–81, 2011.

[47] Y. Wong, M. T. Harandi, C. Sanderson, and B. C. Lovell. On robust biometric identity verification via sparse encoding of faces: Holistic vs local approaches. In *IEEE International Joint Conference on Neural Networks*, pages 1762–1769, 2012.

[48] 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*, 31(2):210–227, 2009.

[49] J. Yang, L. Zhang, Y. Xu, and J.-Y. Yang. Beyond sparsity: The role of l_1-optimizer in pattern classification. *Pattern Recognition*, 45(3):1104–1118, 2012.

[50] M. Yang and L. Zhang. Gabor feature based sparse representation for face recognition with Gabor occlusion dictionary. In *ECCV (6)*, volume 6316 of *Lecture Notes in Computer Science*, pages 448–461. Springer, 2010.

[51] M. Yang, L. Zhang, X. Feng, and D. Zhang. Fisher discrimination dictionary learning for sparse representation. In *IEEE International Conference on Computer Vision*, pages 543–550, 2011.

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