www.ietdl.org

Published in IET Computer Vision
Received on 10th February 2014
Revised on 8th May 2014
Accepted on 10th June 2014
doi: 10.1049/iet-cvi.2014.0039

ISSN 1751-9632

Face recognition: handling data misalignments implicitly by fusion of sparse representations

Hugo Proença, João Neves, Juan Briceño

Department of Computer Science, Instituto de Telecomunicações, University of Beira Interior, 6200 Covilhã, Portugal
E-mail: hugomcp@di.ubi.pt

Abstract: Sparse representations for classification (SRC) are considered a relevant advance to the biometrics field, but are particularly sensitive to data misalignments. In previous studies, such misalignments were compensated for by finding appropriate geometric transforms between the elements in the dictionary and the query image, which is costly in terms of computational burden. This study describes an algorithm that compensates for data misalignments in SRC in an implicit way, that is, without finding/applying any geometric transform at every recognition attempt. The authors' study is based on three concepts: (i) sparse representations; (ii) projections on orthogonal subspaces; and (iii) discriminant locality preserving with maximum margin projections. When compared with the classical SRC algorithm, apart from providing slightly better performance, the proposed method is much more robust against global/local data misalignments. In addition, it attains performance close to the state-of-the-art algorithms at a much lower computational cost, offering a potential solution for real-time scenarios and large-scale applications.

1 Introduction

Sparse representations have been extensively reported in the computer vision literature. The idea is that a point $\boldsymbol{y}$ in a feature space can be appropriately represented by linear combinations of other points in a 'dictionary' $\boldsymbol{A}$. Such linear combinations are found by obtaining solutions to underdetermined systems $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$ (constrained by some norm). Finally, inference is done by reconstructing $\boldsymbol{y}$ with respect to some of the coefficients $\boldsymbol{x}$ and finding the minimal residuals.

According to the above concept, in the biometrics domain, the idea is that a sample (image) of a trait (e.g. iris, face) can be appropriately represented using only elements of the same class (identity). Hence, the classical sparse representation for classification (SRC) algorithm concatenates a set of images with known identity (called the 'gallery' elements) in the dictionary $\boldsymbol{A}$. Having an image of unknown identity (called the probe, $\boldsymbol{y}$), the recognition process is divided into two phases: (i) the probe is represented by a linear combination of the dictionary elements and (ii) the probe is reconstructed with respect to every identity in the gallery, that is, using exclusively the coefficients $\boldsymbol{x}$ that regard the corresponding identity. Then, the minimal residual between the probe and the reconstructed version is deemed to correspond to the identity of the probe.

Sparse representations are extremely effective in biometric recognition, provided that: (i) the gallery and probe images are aligned, to guarantee that the gallery elements constitute a 'vector basis' and (ii) a sufficient number of elements are included in the dictionary, guaranteeing that the corresponding system of linear equations is underdetermined and a sparse solution can be found.

Data misalignments are a major problem in sparse representations, as illustrated in Fig. 1: images in the upper row regard the same subject, but are misaligned. On the contrary, images in the bottom row regard notoriously different subjects but are accurately aligned. The residual of expressing one of the images as a linear combination of the other is much higher in the case of the upper row (same identity) than for the case in the bottom row (different identities). As this problem seriously affects sparse representations, it has been considered in several research works (e.g. [1, 2]).

However, we noted that in all of the methods published, either the gallery or probe images are explicitly aligned before the sparse representation, which augments the computational burden of recognition and turns difficult the use of sparse representations in large-scale identification scenarios. This was the main motivation behind this paper: to provide a simple recognition algorithm based on sparse representations that is robust to data misalignments, with two constraints: (i) do not explicitly align either gallery or probe data and (ii) do not substantially increase the computational burden of the recognition process.

Our solution cast the problem by evolving four concepts: 'score fusion, sparse representations, projections' into 'orthogonal subspaces', according to the 'discriminating locality preserving' and 'maximum margin' criteria. The foundations are as follows:

- The Johnson–Lindenstrauss lemma [3] states that points in a high-dimensional space can be projected into a much lower-dimensional subspace that preserves a lot of its structure in terms of inter-point distances and angles.
- Elad and Yavneh [4] observed that 'a plurality of sparse representations is better than the sparsest one alone' and

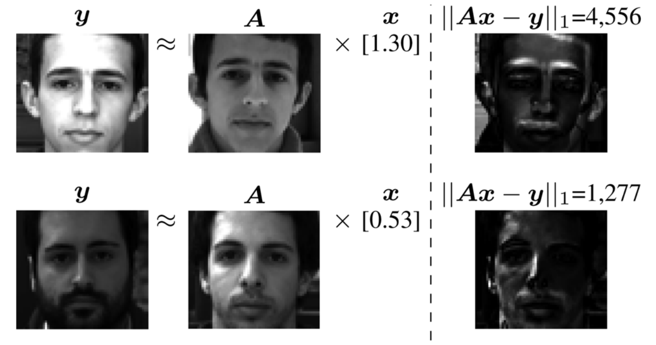

Fig. 1 *Illustration of the misalignment problem*

Even though images in the upper row are from the same subject and have neutral facial expression, they are misaligned (note the ear at the right band and the chin)
Residual of expressing $\boldsymbol{y}$ as a function of $\boldsymbol{A}$ (dictionary with a single element) is larger than in the case of the bottom row that regards notoriously different subjects

considered the set of slightly inferior representations more meaningful than the sparsest. Having merged several competing representations, they obtained a more accurate representation of the original signal in the mean-square error sense.

An empirical evaluation was conducted and the effectiveness of the proposed solution was compared with the original SRC algorithm and two alignments methods because of Pillai *et al.* [1] and Wagner *et al.* [2]. Regarding the data misalignments, two different kinds were considered: 'global' misalignments, as a result of failures in the detection of the region-of-interest (ROI), and 'local' misalignments, because of non-linear deformations in image patches (e.g. facial expressions).

The remainder of this paper is organised as follows: Section 2 summarises the most relevant works in terms of sparse representations in biometrics research. Section 3 provides a description of the proposed method. Section 4 contains the empirical validation procedure that was carried out. Finally, conclusions are given in Section 5.

2 Related work

This section is divided into two parts: (i) we start by summarising the most relevant works that use sparse representations in biometric research and (ii) next, we focus on the data misalignment problem and describe the major existing approaches to attenuate such problems. Finally, we highlight the key distinguishing point between such approaches and the work described in this paper.

2.1 Sparse representations in biometrics research

SRC-based algorithms have been argued as a relevant advance to the field of biometrics research, particularly since the pioneering approach of Wright *et al.* [5]. The idea is to concatenate a set of gallery images into a 'dictionary' $\boldsymbol{A} \doteq [\boldsymbol{v}_{11}, \ldots, \boldsymbol{v}_{\text{in}}]$ ($\boldsymbol{v}_{ij}$ is the *j*th image from the *i*th subject). Then, any probe $\boldsymbol{y}$ of unknown identity is represented as a linear combination in terms of $\boldsymbol{A}$, that is, $\boldsymbol{y} = \boldsymbol{Ax}$, $\boldsymbol{x}$ being the coefficients of that linear combination. The identity corresponding to $\boldsymbol{y}$ is found based on how well the $\boldsymbol{x}$ coefficients associated with each known identity reproduce $\boldsymbol{y}$. Let $\hat{\boldsymbol{y}}_i = \boldsymbol{A}\delta_i(\hat{x}_i)$ be the reconstruction of $\boldsymbol{y}$ using only coefficients from the *i*th identity (δ_i is a characteristic function). By obtaining the residuals of each identity $r_i = ||\boldsymbol{y} - A\delta_i(\hat{x}_1)||_2$, the minimal r_i corresponds to the identity of $\boldsymbol{y}$.

Patel [6] addressed the lighting and pose factors for unconstrained face recognition, having used a relighting approach based on the albedo estimation. The K-means singular value decomposition (K-SVD) algorithm was used to learn the dictionaries that best represent training data. The approximation vectors for each dictionary were found and the minimal reconstruction error among all dictionaries provided the identity for a probe. Buyssens and Revenu [7] fused at the decision level the rankings from sparse representations on multispectral data. A two-phase sparse representation was proposed by Guo [8] for the palmprint recognition problem. They selected a subset of the identities that more closely reconstruct the probe and obtained a second sparse representation for these elements. Classification resulted from the minimal reconstruction error on the latter phase. In the scope of ear recognition, Khorsandi *et al.* [9] used Gabor kernels for feature encoding before the sparse representation. Gong *et al.* [10] analysed spatiotemporal human gait signals by sparse representations, obtaining recognition performance comparable with state-of-the-art approaches and with higher robustness to clothing. Aiming at recognising irises in uncontrolled setups, Kumar *et al.* [11] used sparse representations of local patches of the iris described by the Radon transform, realigned in different ways to increase the robustness to segmentation errors. Wong *et al.* [12] argued that the sparse representation and classification algorithm is not applicable to verification problems, and proposed a model based on the 'bag-of-words'. This model regards images as sets of local patches and encodes each one by sparse representations. Concatenated coefficients feed a neural network that gives the binary response in case of verification problems.

2.2 Handling data misalignments in sparse representations

As above stated, SRC-based algorithms are particularly effective, provided that gallery and probe elements are minimally aligned, to guarantee that any query image can be appropriately represented by a linear combination of the gallery elements, that is, the dictionary is a vector basis. This problem was previously studied in the literature, and we divide the previously published techniques to compensate for data misalignments into two families: (i) the first one assumes that all gallery images are aligned and explicitly aligns probes, which augments the requirements about the quality of the training data and (ii) the other family does not have any alignment constraint and for every recognition attempt, aligns both the gallery and probe elements, which is practically impossible for large-scale applications, where dictionaries contain a large number of identities.

Pillai *et al.* [1] started by estimating the best alignments for the probe according to matched filters and obtained the final recognition score from a Bayesian fusion framework. For a probe $\boldsymbol{y}$, a set of filters with impulse response equal to shifted versions of y is built, resulting in the matrix $\boldsymbol{H} \doteq [\hat{\boldsymbol{y}}_1, \ldots, \hat{\boldsymbol{y}}_t]$. Next, the distances between $\hat{\boldsymbol{y}}$ and the dictionary elements are obtained, that is, $e = ||\hat{\boldsymbol{y}}_i - \boldsymbol{v}_{j,k}||_2$, $\forall j$, k, i. The alignment error associated with the *i*th shifted version is given by

$$e_i = \min_{k,j} \; e_{i,j,k}, \quad \forall j, \; k$$

$e_{i,j,k}$ being the residual between the ith alignment and the jth image of class k. The lowest values of e_i are considered the best possible alignments and the corresponding versions of $\boldsymbol{y}$ used in the matching phase. Next, the output class is deemed to correspond to the maximum accumulated score for all shifted versions of $\boldsymbol{y}$.

Wagner *et al.* [2] considered that probes are warped by some transformation $\tau \in T$, T being a finite-dimensional group of transformations in the image domain, that is, $\boldsymbol{y} = \boldsymbol{y}_0 \circ \tau$. Hence, $\boldsymbol{y}$ is not supposed to have a sparse representation of the form $\boldsymbol{Ax} + \boldsymbol{e}_0$. In this case, it is more appropriate to seek the best alignment of $\boldsymbol{y}$ with respect to each subject

$$\hat{\tau}_i = \arg \min_{x,e,\tau_i} ||\boldsymbol{e}||_1, \text{ subj. to } \boldsymbol{y} \circ \tau_i = \boldsymbol{A}_i \boldsymbol{x} + \boldsymbol{e} \qquad (1)$$

$\boldsymbol{A}_i$ being the gallery data of the ith subject. A sequential ℓ^1-minimisation schema is adopted, starting by an initial guess of the two-dimensional (2D) similarity transform τ_i provided by the face detector algorithm. Then, the estimate is refined by repeatedly linearising about τ_i, seeking representations of the form

$$\boldsymbol{y} \circ \tau + J\Delta r = \boldsymbol{A}_i \boldsymbol{x} + \boldsymbol{e} \qquad (2)$$

where $J = (\partial/\partial r)\boldsymbol{y} + \circ\tau$ is the Jacobian of $\boldsymbol{y} \circ \tau$ with respect to the transformation parameters and Δr is the step. They seek for a deformation step Δr that most sparsifies the residual e

$$\Delta_{\hat{r}_1} = \arg \min_{x,e,\Delta r} ||\boldsymbol{e}||_1 \text{subj. to } \boldsymbol{y} \circ \tau + J\Delta_r = \boldsymbol{A}_i \boldsymbol{x} + \boldsymbol{e} \qquad (3)$$

Authors regard (3) as a generalised Gauss–Newton method for minimising the composition of the ℓ^1-norm with a differentiable mapping from transformation parameters to the transformed images [13].

Huang *et al.* [14] assumed that a set of aligned training data is available. For a misaligned probe $\boldsymbol{y}$, they proposed to represent it in terms of the training images I and of their derivatives

$$\boldsymbol{y} = \sum_j \alpha_j \left(I_j + a_1 \frac{\partial I_j}{\partial x} + a_2 \frac{\partial I_j}{\partial y} \right) \qquad (4)$$

being (a_1, a_2) the translation parameters (other 2D similarity transforms apply similarly). This generates a linear model $\boldsymbol{B}$ with three times the initial number of images in the training set. A random projection is used to reduce dimensionality and the sparsest solution in this space derived, that is, $\boldsymbol{x}_0 = \arg\min ||\boldsymbol{x}||_1$, subj. to $||\boldsymbol{y}^* - \boldsymbol{B}^* \boldsymbol{x}||_2 < \epsilon$, where $\boldsymbol{y}^*$ and $\boldsymbol{B}^*$ are the representations in the space of reduced dimensionality. Next, the sparse solution x_0 is divided into $[z_0, z_1, z_2]^{\mathrm{T}}$ and the z_0 considered the aligned projection target. Based on it, an iterative process finds the parameters of the 2D transformation that better aligns $\boldsymbol{y}$ to the gallery samples.

As above described, the state-of-the-art algorithms to attenuate the data misalignment problem in SRC explicitly align the data for each recognition attempt, either the gallery (highest computational cost) or the probe images (with a smaller cost, but assuming that all gallery images are aligned). As described in the next section, the key distinguishing point of the method in this paper is that no explicit data alignment is carried out for every recognition attempt. Instead, by fiddling appropriate (and cheap) projections, the data misalignments can be compensated for in a way that is almost as effective as that attained by state-of-the-art techniques, at a much smaller computational cost.

3 Proposed method

We start by finding a set of orthogonal projections into subspaces. Then, each projection is optimised according to the locality preserving and maximum margin criteria, which is particularly attractive for biometric recognition purposes. Elements are projected into each subspace and sparse representations obtained independently in each one. At the end, results are fused at the score level, yielding the final response.

3.1 Random projections

As a result of the Johnson and Lindenstrauss lemma [3], it is known that a set of points in a high-dimensional Euclidean space can be embedded into a lower-dimensional space, so that all pairwise distances are maintained within an arbitrarily small factor. Over the years, the classical algorithms to perform such reduction in dimensionality project the input data onto a spherically random hyperplane through the origin, which amounts to multiply the input data with a dense matrix of real numbers. This might be a non-trivial task for many practical scenarios in terms of computational burden.

One of the main contributions from Achlioptas [15] is that such projections into spherically random hyperplanes can be replaced by much simpler operations (multiplication by random vectors, built with step functions), without significant loss in the quality of embedding. Hence, we decided to use sparse random vectors based in these extremely simple step functions, which are computationally cheap to generate.

Let m be the dimension of the image feature space and d the dimension of subspaces. A set of d linearly independent vectors $\{\boldsymbol{v}_i\}$, $\boldsymbol{v}_i \in \mathbb{R}^m$ is generated, that is, $\sum_i c_i \boldsymbol{v}_i = 0 \Rightarrow c_i = 0$. Let $U \sim \mathbb{U}(0, 1)$ be a random variable that follows a uniform distribution. The jth coordinate of a random vector is given by

$$\boldsymbol{v}^{(j)} = \begin{cases} 1, & \text{if } u^{(j)} \le \dfrac{1}{3} \\ -1, & \text{if } u^{(j)} \ge \dfrac{2}{3} \\ 0, & \text{otherwise} \end{cases}$$

$u^{(j)}$ being a realisation of U. In our case, the problem of generating d linearly independent vectors was tackled iteratively: at the ith iteration $(1 \le i \le d)$, the random vector $\boldsymbol{v}_i$ was added to set $\boldsymbol{B}$ if it is linearly independent of all its elements, that is, $\boldsymbol{B}_i = [\boldsymbol{B}_{i-1}, \boldsymbol{v}_i]$ if $\mathrm{rank}([\boldsymbol{B}_{i-1}, \boldsymbol{v}_i]^{\mathrm{T}}) = i$, being $\boldsymbol{B}_0 = \emptyset$.

3.2 Orthogonal subspaces projections

The idea of an ‘orthogonal projection’ of $\boldsymbol{y} \in \mathbb{R}^m$ into an element in $\mathbb{R}^d$, $d \le n$, is to find a projected vector $\boldsymbol{y}^* \in \mathbb{R}^d$ orthogonal to all basis elements $\boldsymbol{v}_i$, that is, $\boldsymbol{y} - \boldsymbol{y}^* \perp \boldsymbol{v}_i$. Let $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_d\}$ be a set of linearly independent column vectors,

$\boldsymbol{y}_i \in \mathbb{R}^n$ and $\boldsymbol{B} = [\boldsymbol{v}_1, \ldots, \boldsymbol{v}_d]$ be an $n \times d$ matrix that results from the column-wise concatenation of vectors $\boldsymbol{v}_i$. $\boldsymbol{y}^*$ can be expressed as a linear combination of basis elements $\boldsymbol{v}_i$

$$\boldsymbol{y}^* = \sum_i c_i \boldsymbol{v}_i = \boldsymbol{B}[c_1, \ \ldots, \ c_d]^{\mathrm{T}} \tag{5}$$

being $\boldsymbol{c} = [c_1, \ \ldots, \ c_d]^{\mathrm{T}}$. As $\boldsymbol{B}^{\mathrm{T}}(\boldsymbol{y} - \boldsymbol{y}^*) = 0$, $\boldsymbol{B}^{\mathrm{T}}\boldsymbol{y} = \boldsymbol{B}^{\mathrm{T}}\boldsymbol{y}^*$. From (5), $\boldsymbol{B}^{\mathrm{T}}\boldsymbol{y} = \boldsymbol{B}^{\mathrm{T}}\boldsymbol{Bc}$ and under algebraic manipulation yields

$$\boldsymbol{B}(\boldsymbol{B}^{\mathrm{T}}\boldsymbol{B})^{-1}\boldsymbol{B}^{\mathrm{T}}\boldsymbol{y} = \boldsymbol{Bc} \tag{6}$$

As $\boldsymbol{y}^* = \boldsymbol{Bc}$, $\boldsymbol{P} = \boldsymbol{B}(\boldsymbol{B}^{\mathrm{T}}\boldsymbol{B})^{-1}\boldsymbol{B}^{\mathrm{T}}\boldsymbol{y}$ is the projection matrix and maps any vector $\boldsymbol{y} \in \mathbb{R}^m$ to its orthogonal projection $\boldsymbol{y}^* \in \mathbb{R}^d$.

3.3 Subspaces optimisation

Having a set of t projections $\{\boldsymbol{P}_1, \ \ldots, \boldsymbol{P}_t\}$, $\boldsymbol{P}_i = \{\boldsymbol{p}_i^{(1)}, \ldots, \boldsymbol{p}_i^{(m)}\}$, $\boldsymbol{p}_i \in \mathbb{R}^d$, an optimisation step was carried out to maximise the ratio between the 'between-class' scatter and the 'within-class' scatter, obtaining as many separate representations of each class (subject) as possible in each subspace. The idea of 'interesting' projections came from Friedman and Tukey [16] who attempted to find projections that preserve clusters, linear structures or outliers. In our case, this phase was tackled according to the idea of Lu *et al.* [17], using two criteria: discriminant locality preserving projections (DLPP) and maximum margin criterion (MMC). For a projection $\boldsymbol{P}_i$, DLPP maximises the objective function

$$J(\boldsymbol{P}_i) = \frac{\boldsymbol{P}_i^{\mathrm{T}}\boldsymbol{FHF}^{\mathrm{T}}\boldsymbol{P}_i}{\boldsymbol{P}_i^{\mathrm{T}}\boldsymbol{XLX}^{\mathrm{T}}\boldsymbol{P}_i} \tag{7}$$

where $\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{W}$ and $\boldsymbol{H} = \boldsymbol{E} - \boldsymbol{B}$ are Laplacian matrices, being $\boldsymbol{W}$ the 'within-class' weight matrix, $W_{i,j}^{(c)} = \exp(-||\boldsymbol{v}_i^{(c)} - \boldsymbol{v}_j^{(c)}||_2/\sigma^2)$, $\boldsymbol{W} = \mathrm{diag}(\boldsymbol{W}^{(c)})$, B the 'between-class' weight matrix $\boldsymbol{B}_{i,j} = \exp(-||\bar{\boldsymbol{v}}^{(i)} - \bar{\boldsymbol{v}}^{(j)}||_2/\sigma^2)$, $\boldsymbol{E}$ is a diagonal matrix which elements are column sum of $\boldsymbol{B}$ and $\bar{\boldsymbol{v}}^{(i)}$ is the mean of the ith class. The transformation that maximises (7) is given by the generalised eigenvalues problem

$$(\boldsymbol{FHF}^{\mathrm{T}})\boldsymbol{a}_i = \lambda_i(\boldsymbol{XLX}^{\mathrm{T}})\boldsymbol{a}_i, \ \lambda_i \geq \lambda_{i+1} \tag{8}$$

Furthermore, in order to obtain the MMC discriminant, the 'between-class scatter' $\boldsymbol{S}_{\mathrm{b}}$ and the 'within-class scatter' $\boldsymbol{S}_{\mathrm{w}}$ matrices are obtained

$$\begin{aligned} \boldsymbol{S}_{\mathrm{b}} &= \frac{1}{n}\sum_{i=1}^{c} n_i\left(\bar{\boldsymbol{v}}^{(i)} - \bar{\boldsymbol{v}}\right)\left(\bar{\boldsymbol{v}}^{(i)} - \bar{\boldsymbol{v}}\right)^{\mathrm{T}} \\ \boldsymbol{S}_{\mathrm{w}} &= \frac{1}{n}\sum_{i=1}^{c}\sum_{j=1}^{n_i}\left(\boldsymbol{v}_j^{(i)} - \bar{\boldsymbol{v}}^{(i)}\right)\left(\boldsymbol{v}_j^{(i)} - \bar{\boldsymbol{v}}^{(i)}\right)^{\mathrm{T}} \end{aligned} \tag{9}$$

where $\bar{\boldsymbol{v}}$ denotes the mean vector of elements in the dictionary and n_i is the number of elements in the ith class. According to the Fisher criterion, the objective function is used

$$J(\boldsymbol{P}_i) = \mathrm{tr}(\boldsymbol{P}_i^{\mathrm{T}}(\boldsymbol{S}_b - \alpha\boldsymbol{S}_w)\boldsymbol{P}_i) \tag{10}$$

α being a balancing weight. The optimal projection corresponds to the eigenvectors $\boldsymbol{a}_i$ associated to the largest eigenvalues

$$(\boldsymbol{S}_{\mathrm{b}} - \alpha\boldsymbol{S}_{\mathrm{w}})\boldsymbol{a}_j = \lambda_j\boldsymbol{a}_j, \ \lambda_j \geq \lambda_{j+1} \tag{11}$$

```
Algorithm 1
Require: Training set $A \in \mathbb{R}^{m\times n}$, k classes, test image $y \in \mathbb{R}^m$, $s \in \mathbb{N}$ number of subspaces, $d \in \mathbb{N}$
    subspace dimension.
for $i = 1$ to s do
    $B_i \leftarrow \emptyset$
    for $j = 1$ to d do
        Generate random vector $v_j, v_j \perp B_i$
        $B_i \leftarrow [B_i, v_j]$;
    end for
    $P_i^* \leftarrow$ DLPP/MMC projection pursuit($B_i$).
    $A_i^* \leftarrow P_i^* A$
    $y_i^* \leftarrow P_i^* y$
end for
for all $1 \leq i \leq k$ do
    for $j = 1$ to s do
        Obtain the sparse representation:
        $\hat{x}_j = \arg\min ||x||_1$, subj. to $||A_j^* x - y_j^*||_2 \leq \epsilon$.
        Obtain residuals: $r_j^{(i)}(y) = ||y_j^* - A_j^* \delta_i(\hat{x}_j)||_2$
    end for
    $f_i(y_j^*) \leftarrow$ fusion$\{r_1^{(i)}, \ldots, r_s^{(i)}\}$
end for
return identity($y_j^*$) $= \arg\max_i f_i(y_j^*)$
```

Fig. 2 *SRC for misaligned data*

doi: 10.1049/iet-cvi.2014.0039

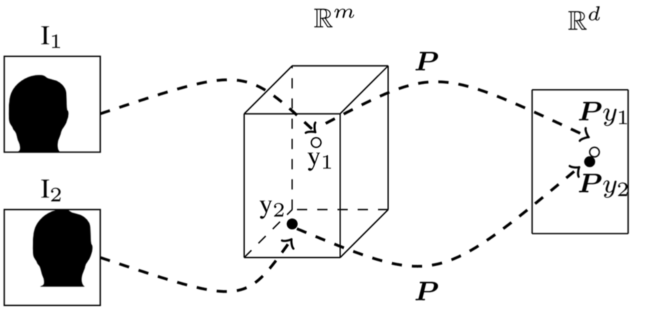

Fig. 3 *Key insight into the method proposed in this paper: having two misaligned samples represented in a space of high dimensionality m, for some interesting projections in subspaces of dimension d(d ≪ m), misalignments are not so evident*

Hence, samples are represented in these subspaces and sparse representations are obtained
Finally, results are fused at the score level

3.4 Fusion of representations

Let $\boldsymbol{P}_s^*$ denote the sth projection optimised by the DLPP/MMC criteria. In our algorithm, gallery and probe data are projected into each subspace, that is, $\boldsymbol{A}_s^* = \boldsymbol{P}_s^* \boldsymbol{A}$ and $\boldsymbol{y}_s^* = \boldsymbol{P}_s^* \boldsymbol{y}$. Let $r_s^{(i)}$ be the residual of the sth sparse representation for the ith class ω_i, given a probe y. Using the theoretical framework developed by Kittler *et al.* [18], all combinations of the responses given by sparse representations were tested according to the fusion rules: product, sum, min, max and median. Without any assumption of the prior probabilities, the posterior probability that a residual error $r^{(i)}$ belongs to class w_j was obtained by

$$p(\omega_j | r_s^{(i)}) = \frac{p(r_s^{(i)} | \omega_j)}{\sum_k p(r_s^{(i)} | \omega_k)} \tag{12}$$

k being the number of classes. The density $p(r_s^{(i)} | \omega_j)$ was estimated by kernel-based density methods [19]. Class w_c was assumed if $w_c = \arg_j \max \phi \ p(w_j / r^{(i)})$, where ϕ denotes the combination rule. Algorithm 1 (see Fig. 2) gives a cohesive perspective of the complete recognition process.

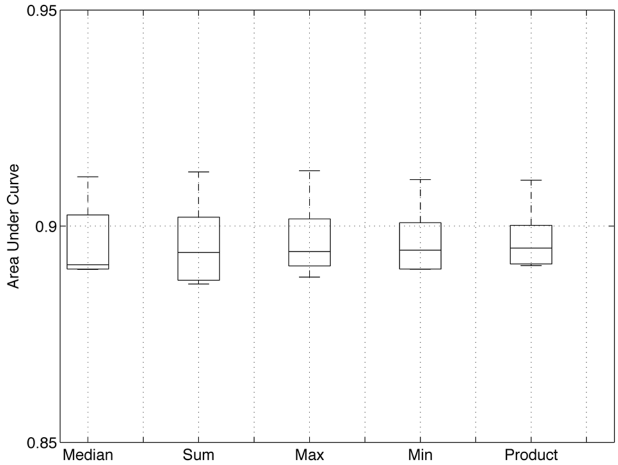

Fig. 4 *Area under curve values obtained with respect to different rules used to fuse scores*

Results regard the projection into ten orthogonal subspaces of dimension $0.98 \times m$, m being the dimension of the image feature space ($m = 90$ for 10×9 images)

Fig. 3 gives the rationale behind the proposed algorithm and Fig. 4 illustrates the residual effect of the different fusion rules in performance, as given by the area under curve (AUC) values. These results regard the fusion at the score level, using ten subspaces of dimension $0.98 \times m$, being m the dimension of the image feature space (images have 10×9 pixels, $m = 90$). This stability in performance was regarded as a positive indicator of the consistency of the proposed method, and allowed to conclude about a minimal dependence of the chosen fusion rule.

4 Results and discussion

Our experiments were conducted in two phases: at first, to validate our implementation and provide a comprehensible comparison term in terms of performance, effectiveness was tested in the AR dataset [20] that is widely used in face recognition experiments. Next, the FaceUBI dataset was used, whose annotation meta-data particularly fits our purposes and makes the experiments about data misalignments easier.

4.1 AR dataset

The AR dataset has over 4000 frontal images from 126 subjects, collected in two sessions. The images have white backgrounds and vary in terms of illumination, facial expressions and disguises. Similarly to the experiments described in [5], a set of 100 subjects was selected and 15 frames from each were considered (without occlusions, sunglasses and disguises). From these, ten frames were randomly selected for training (dictionary elements) and the remaining for probe data. Images were converted to greyscale and ROIs were marked manually, deliberately without particular alignment concerns, to simulate slight misalignments in the face detection step. Next, images were resized to 11×8 dimensions, ($n = 88$, dimension of the feature space). The left column of Fig. 5 illustrates some examples of the images in this dataset and of its major variation factors.

The evaluation protocol was as follows: for a dictionary with k identities, for each probe y the reconstruction residuals r_i ($i \in \{1, \ldots, k\}$) were obtained, each one using a $\delta_i()$ function. The software package described in [21] was used to obtain the sparse representations. Next, all residuals were concatenated and a threshold acceptance level varied, obtaining a receiver operating characteristic (ROC) curve, which in our viewpoint carries much more information about the performance levels of the system than the recognition rate plots given in similar works. The results are summarised in the right plot of Fig. 5, and show consistent increases in performance of the proposed method when compared with the original sparse representation algorithm. Experiments were repeated 20 times, by randomly choosing images to be used in dictionary and probe sets. The median performance levels are represented by the lines series, whereas the best and worst performance at each point is represented by the horizontal bars around each data point.

4.2 FaceUBI dataset

About 4000 facial images were selected from the FaceUBI dataset, with the corresponding annotation files that

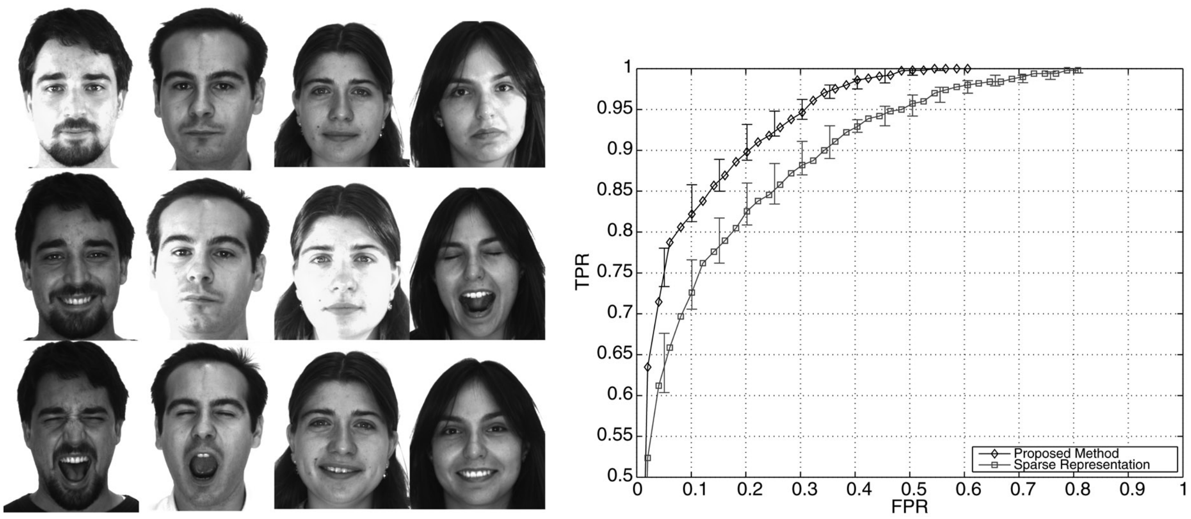

Fig. 5 *At left: examples of the AR images used in our experiments*

Variations are predominant in lighting and facial expression criteria

Right figure compares the ROC curves obtained by the classical SRC algorithm and the method proposed in this paper

delimitate the facial ROIs. For all these images, it was confirmed by visual inspection that the ROIs were coherently defined, that is, all images are aligned. Images are from 100 subjects, 40 frames per subject (20 frames from each session). They regard exclusively frontal subjects and were acquired under varying lighting conditions and in complex backgrounds. For the purpose of reproducibility of the results, both the images and annotation data are freely available [http://www.di.ubi.pt/~hugomcp/SparseAlign].

4.3 Effect of the amounts of training data

In the first level of analysis, we verified the levels of performance with respect to the amount of gallery data, that is, to the number of images per subject used in the dictionary. The experiments were repeated when using 1 to 25 images per subject in the dictionary. In addition, it should be stressed that gallery and probe data from each subject always regard different sessions, to minimise the dependence between training and test data. The obtained area under curve (AUC) values are given in Fig. 6, where the number of images per subject in the dictionaries appears in the horizontal axis and the corresponding AUC values in the vertical axis. For both the classical sparse representation algorithm and the proposed method, results tend to stabilise when more than five images were included in the dictionary. In this experiment, images had 10 × 9 pixels, thus it was guaranteed that even when using a single image per subject, the system of linear equations was underdetermined (100 subjects were used).

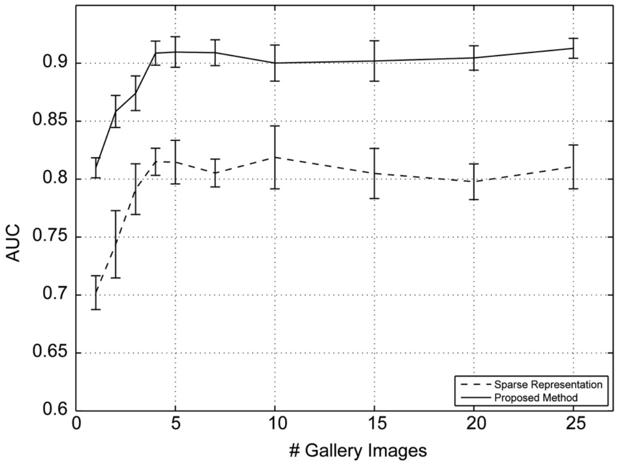

Fig. 6 *Comparison between the AUC values of the recognition system with respect to the number of images per subject used in the dictionary*

4.4 Effect of the number and dimension of subspaces

It is expected that the number of projections into subspaces and the dimension of these play important roles in the performance of the proposed method. Hence, the analysis of the AUC values with respect to both parameters was carried out. However, it should be considered that the number of subspaces should be kept as small as possible, as it determines the computational burden of the recognition process in a roughly linearly way, that is, a sparse representation must be found for each subspace and recognition attempt. Both parameters were varied in regular

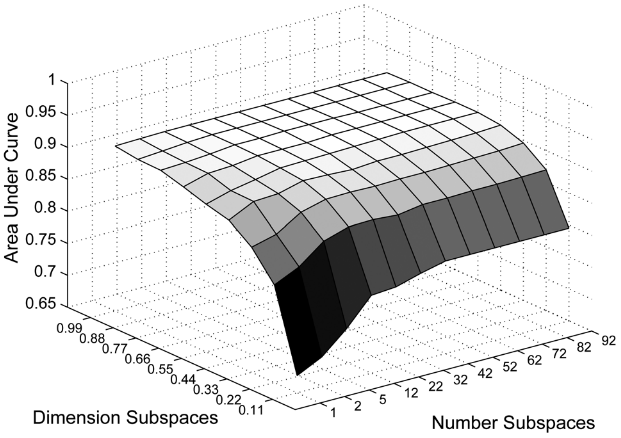

Fig. 7 *Effect of the number of subspaces used and of their dimensions in the recognition effectiveness of the proposed method*

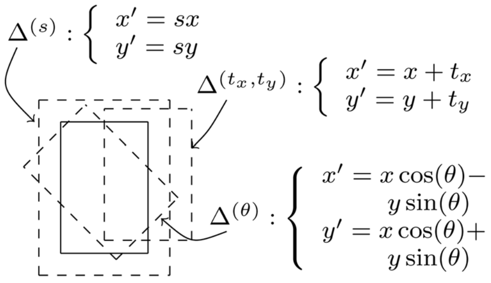
$\Delta^{(s)}: \begin{cases} x' = sx \\ y' = sy \end{cases}$
$\Delta^{(t_x,t_y)}: \begin{cases} x' = x + t_x \\ y' = y + t_y \end{cases}$
$\Delta^{(\theta)}: \begin{cases} x' = x\cos(\theta) - y\sin(\theta) \\ y' = x\cos(\theta) + y\sin(\theta) \end{cases}$

Fig. 8 *Illustration of the three kinds of global misalignments considered*

For a region-of-interest (continuous rectangle), translation $\Delta^{(t_x, t_y)}$, scale $\Delta^{(s)}$ and rotation $\Delta^{(\theta)}$ transformations of varying magnitude were used

intervals: from 1 to 92 subspaces projections, each one with dimensions that varied from 0.11 to $0.99 \times m$, m being the dimension of the image feature space ($m = 90$, for 10×9 images). Results are shown in the 3D plot of Fig. 7 and it is noteworthy to stress the stability of performance when more than five subspaces were used, each one with dimension higher than $0.33 \times m$. In addition, this analysis was restricted to subspaces of equal dimension, even though better performance might be possible to obtain, if the dimensions of each subspace are allowed to vary independently.

4.5 Global misalignments in ROI location

As illustrated in Fig. 8, three kinds of misalignments were simulated, corresponding to translation, scale and rotation transforms of the manually defined ROIs. This way, the corresponding variations in effectiveness of the classical sparse representation algorithm and of the proposed method were observed. In addition, to contextualise the results, the proposals of Pillai *et al.* [1] and Wagner *et al.* [2] were used as comparison terms, selected because of their relevance in the sparse representation literature.

Fig. 9 provides the ROC curves for the proposed recognition method (continuous lines) and the classical sparse representation algorithm (dashed lines), as described by Wright *et al.* [5]. Results are shown with respect to misalignments $\Delta^{(\cdot)}$ of increasing magnitude (from right to left: translation, scale and rotation). The horizontal bars around each data point denote the performance range observed, when repeating each experiment 20 times, randomly choosing the gallery and probe elements.

Fig. 10 summarises the decreases in performance with respect to the magnitude of misalignments, in comparison with the techniques proposed by Pillai *et al.* [1] and Wagner *et al.* [2]. Results are given in terms of boxplots of the decidability of the pattern recognition system

$$d' = \frac{|\mu_G - \mu_I|}{\sqrt{\frac{1}{2}(\sigma_I^2 + \sigma_G^2)}} \tag{13}$$

being μ_G and μ_I are the means of the r_i values for genuine and impostor comparisons and σ_G and σ_I are the standard deviations. Four groups are shown in each figure, each one with three boxplots, where the luminance of each box directly corresponds to the magnitude of misalignments, that is, the black boxes regard minimal misalignments ($\Delta^{(tx,\ ty)}$

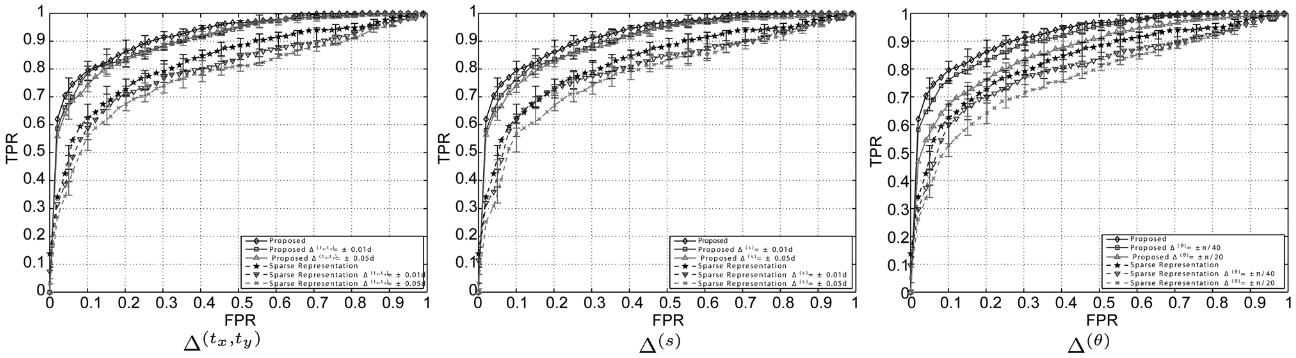

Fig. 9 *Comparison between the results observed for the classical algorithm of SRC, and for the method described in this paper*

Results are given with respect to the magnitude of misalignments $\Delta^{(t_x, t_y)}$, $\Delta^{(s)}$ and $\Delta^{(\theta)}$

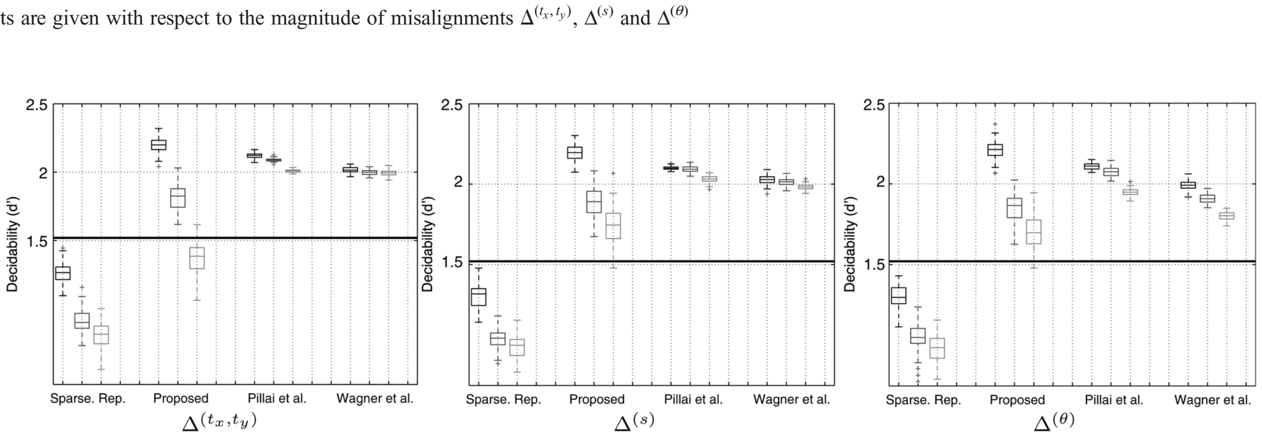

Fig. 10 *Summary of the decreases in the decidability index d' with respect of the magnitude of global misalignments in data*

$(\Delta^{(tx,\ ty)}) = \pm\{0.02, 0.050.1\}d$, $\Delta^{(s)} = \pm\{0.02, 0.050.1\}d$ and $\Delta^{(\theta)} = \pm\pi/\{40, 2515\}$

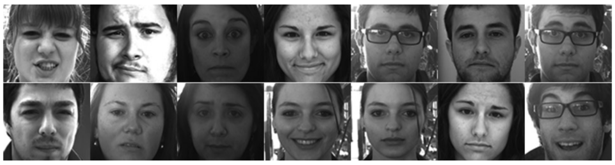

Fig. 11 *Examples of the seven types of facial expressions considered in the analysis of local misalignments, from left to right: anger, disgust, fear, happy, neutral, sad and surprise*

$= \pm 0.02d$, $\Delta^{(s)} = \pm 0.02d$ and $\Delta^{(\theta)} = \pm\pi/40$) ($d$ is the length of the diagonal of ROIs ($\simeq 13.45$ for 10×9 images) and the light grey boxes give the performance for misalignments of maximal magnitude ($\Delta^{(tx,ty)} = \pm 0.1d$, $\Delta^{(s)} = \pm 0.02d$ and $\Delta^{(\theta)} = \pm\pi/15$). The solid horizontal lines express the performance of the classical sparse representation algorithm for aligned data. Minimal decreases in performance were observed for the proposals of Pillai *et al.* and Wagner *et al.*, but in this case it should be stressed that both methods explicitly align either probes or gallery data, considerably augmenting the computational burden of the recognition process. The method of Pillai *et al.* outperformed for slight misalignments, but more critically degraded performance than Wagner *et al.*'s for more severe misalignments. Regarding the proposed method, the decreases in performance were larger than both Pillai and Wagner's methods but – with exception to translation misalignments of large magnitude – the levels of performance remained consistently above the horizontal black line, that is, even on misaligned data, results were better than in the classical sparse representation algorithm for aligned data. In addition, in every direct comparison with the original algorithm, the proposed method obtained best performance without interception of the upper/bottom performance range of both methods, pointing about the statistical significance of these results.

4.6 Local misalignments because of facial expressions

Facial expressions induce misalignments in local image patches because of the action unit muscles evolved in the process. As in the case of global misalignments, this factor was regarded as a covariate and performance compared with respect to facial expressions in the gallery and probe data. Fig. 11 illustrates the seven categories considered:

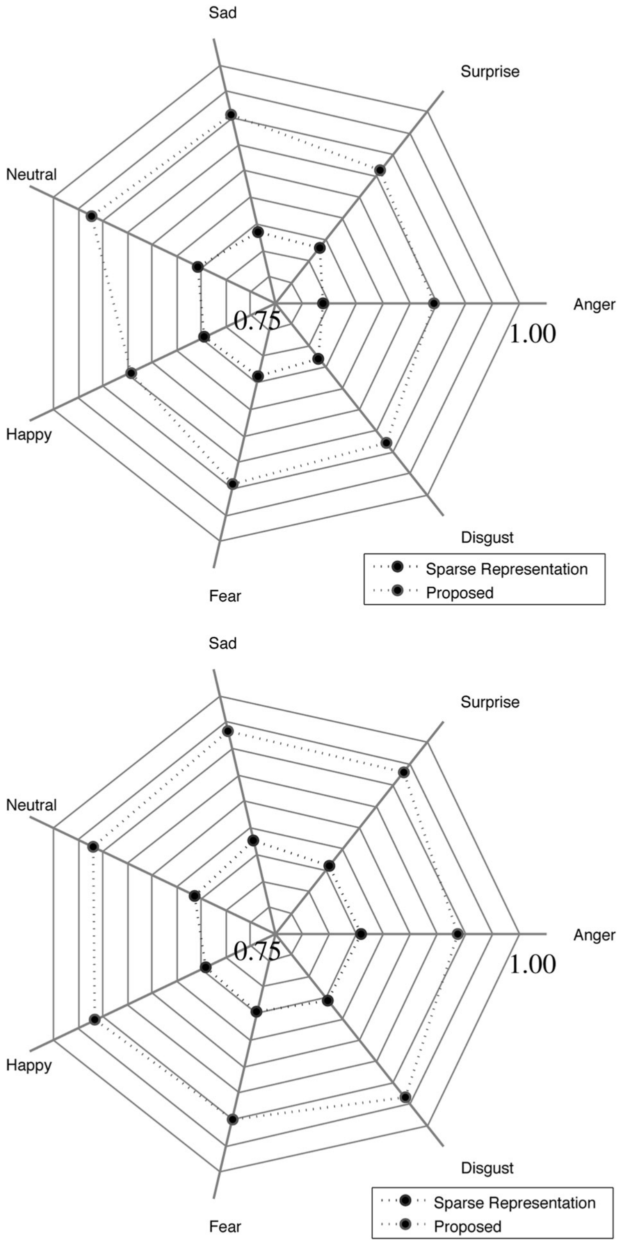

Fig. 13 *Variations in performance with respect to the facial expressions in probe data, when gallery samples are exclusively neutral (left plot), and when gallery data have different facial expressions (right plot)*

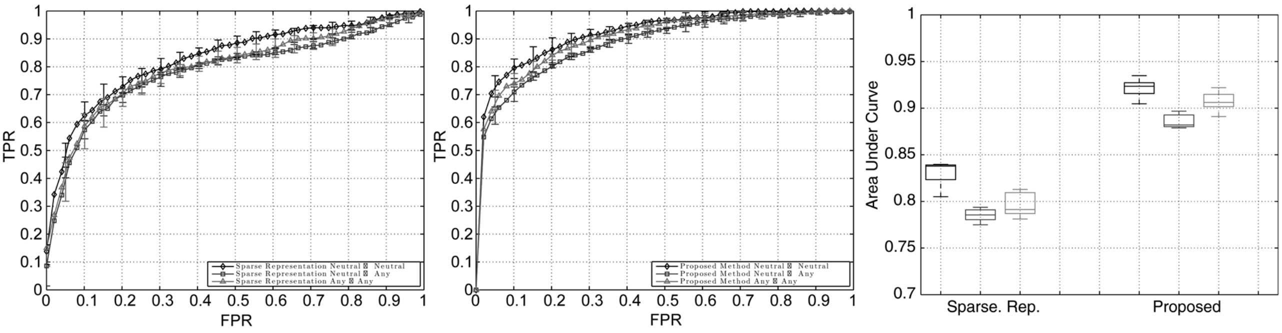

Fig. 12 *Effect of facial expressions in recognition performance*

Left plot expresses the results for the classical SRC algorithm
Plot at the centre gives the results for the proposed method
Boxplot at the right summarises the variations in performance

doi: 10.1049/iet-cvi.2014.0039

Table 1 Comparison between the average turnaround time (*S*) of the classical SRC algorithm (sparse representation Column) and of the method described in this paper

Methods	$s=2$, $m=0.11$	$s=2$, $m=0.99$	$s=42$, $m=0.11$	$s=42$, $m=0.99$	$s=82$, $m=0.11$	$s=82$, $m=0.99$
sparse representation (SRC)	0.05 ± 0.001	0.74 ± 0.003	0.05 ± 0.001	0.74 ± 0.003	0.05 ± 0.001	0.74 ± 0.003
proposed method	0.06 ± 0.002	1.05 ± 0.004	0.89 ± 0.030	15.75 ± 0.071	1.79 ± 0.03	31.05 ± 0.096

s Denotes to the number of subspaces used and *m* is the dimension of each subspace ($m \times 10 \times 9$ elements) (FaceUBI dataset).

'anger', 'disgust', 'fear', 'happy', 'neutral', 'sad' and 'surprise'. In this experiment, three variants were considered: (i) only neutral images were considered; (ii) gallery samples were exclusively neutral and probes had different expressions; and (iii) both gallery and probes samples might have any facial expression.

Results are summarised in Fig. 12 and do not show a relevant effect of expressions in performance, both for the classical sparse representation algorithm and ours method. This accords with previous reported results (e.g. [5]) and confirms the remarkable ability of sparse representations to handle non-linear deformations in images patches, because of occlusions, disguises or facial expressions. The AUC values shown at the rightmost plot confirm the minimal decreases in performance with respect to this factor, where the black boxes represent the 'neutral ↔ neutral' results, and the remaining boxes represent the 'neutral ↔ any' and 'any ↔ any' results.

Finally, to perceive the effect of each type of facial expression, experiments were repeated for each facial expression in an isolated way. Results are given in the radar charts of Fig. 13, where spokes represent facial expressions. The length of each spoke corresponds to the AUC value observed for that expression. The upper chart gives the results when matching probes of different expressions against exclusively neutral dictionary images, whereas the bottom chart expresses the results when both probes and gallery data have the same facial expression. In the case of neutral ↔ *X*, anger was observed to be the most problematic facial expression, whereas 'sad' and 'fear' almost did not affect the results. When matching data of the same expression, best results were observed for the 'disgust' and 'sad' expressions, both of them yielding better results than in 'neutral' data. These variations were consistent for both the classical sparse representation algorithm and our proposal.

As a summary, it can be concluded that the proposed method consistently outperformed the classical sparse representation and classification algorithm, at expenses of a slight increase in the computational burden of the recognition process. In addition, the proposed method is more tolerant to misalignments in the ROIs than the classical algorithm. Regarding this factor, it attains performance comparable with state-of-the-art alignment techniques for sparse representations, at a far lower computational burden for every recognition attempt.

4.7 Comparison of turnaround times

To contextualise the computational cost of the algorithm proposed in this paper, we compared the average turnaround times for a query (recognition attempt), according to the SRC algorithm and our proposal. Our method starts by obtaining a set of *s* subspaces according to the DLPP/MMC projection pursuit algorithm. However in practical terms, this phase runs only once during the system initialisation, and it was neglected from the comparison. Then, for both the SRC and our method, the most significant cost of recognition depends on the algorithm that solves the underdetermined systems of linear equations. As stated above, the software package described in [21] was always used for that purpose.

Results are given in Table 1, for different number of subspaces *s* and two different levels of subspaces dimension $m = 0.11$ and 0.99 (feature spaces with dimension $m \times 10 \times 9$, as described in Section 4.4).

The number of elements in the dictionary was kept constant (ten elements per identity), as it is known that the time complexity of the algorithm that solves the system of linear equations is quadratic with respect to this parameter. To obtain an approximate confidence interval, the experiments were repeated 20 times, each iteration selecting a random sample of the data for the dictionary and using the remaining part as probes.

In summary, we observed that the turnaround time of our proposal is slightly higher than for the SRC algorithm, and varies in a roughly linear way with respect to the number of subspaces used. In addition, both the proposed method and the SRC slightly augment the turnaround times with respect to the parameter *m*. Even though the immediate comparison between the turnaround times might lead to conclusions about a significant higher computational cost of our proposal, it should be stressed that projecting the input data into independent subspaces and solving the corresponding systems of linear equations are easily parallelisable tasks, in case of system deployment. Hence, the differences of values with respect to the parameter *s* are easily reduced by parallel computing architectures. The important parameter here is *m*, where the differences in the average turnaround time of our method were similar to those observed for the SRC.

5 Conclusions

Sparse representations are a relevant advance on the biometrics field: they faithfully address data occlusions and different sources of noise, provided that a sufficient number of samples per class exist. However, a key requirement is that elements in the dictionary constitute a vector basis, which enforces that they should be aligned. Various algorithms were proposed to compensate for data misalignments, that attain remarkable effectiveness but either explicitly align the probes (assuming aligned gallery images) or even all gallery images (in case of unconstrained setups, maximum computational cost). This step considerably augments the computational burden of the recognition process, making its application in real-time or large-scale scenarios difficult.

The main goal of this paper is to provide a simple way to improve the robustness of sparse representations in case of misaligned data, without explicitly aligning either gallery or probe images. A method based in random projections into orthogonal subspaces is used to alleviate the effect of data misalignments. Each projection is optimised according to the discriminating locality preserving and maximum margin criteria. Sparse representations are obtained in these subspaces and the final response yields from fusion at the score level of the response from each subspace. When compared with the classical sparse representation algorithm, the empirical results point out consistent improvements in performance. In addition, with respect to the state-of-the-art techniques that compensate for data misalignments in sparse representations, similar performance was observed for slight to moderate misalignments, at a far lower computational cost in the recognition process.

6 References

1 Pillai, J., Patel, V., Chellappa, R., Ratha, N.: 'Secure and robust iris recognition using random projections and sparse representations', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2011, **33**, (9), pp. 1877–1893

2 Wagner, A., Wright, J., Ganesh, A., Zhou, Z., Mabahi, H., Ma, Y.: 'Toward a practical face recognition system: robust alignment and illumination by sparse representation', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2009, **34**, (2), pp. 371–386

3 Johnson, W., Lindenstrauss, J.: 'Extensions of Lipschitz mappings into a Hilbert space'. Proc. Conf. Modern Analysis and Probability, American Mathematical Society, 1984, vol. 26, pp. 189–206

4 Elad, M., Yavneh, I.: 'A plurality of sparse representations is better that the sparsest one alone', *IEEE Trans. Inf. Theory*, 2009, **55**, (10), pp. 4701–4714

5 Wright, J., Yang, A., Ganesh, A., Sastry, S., Ma, Y.: 'Robust face recognition via sparse representation', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2009, **31**, (2), pp. 210–227

6 Patel, V.: 'Dictionary-based face recognition under variable lighting and pose', *IEEE Trans. Inf. Forensics Sec.*, 2012, **7**, (3), pp. 954–965

7 Buyssens, P., Revenu, M.: 'IR and visible face identification via sparse representation'. Proc. Fourth IEEE Int. Conf. Biometrics: Theory Applications and Systems (BTAS), 2010, pp. 1–6

8 Guo, Z.: 'Palmprint recognition by a two-phase test sample sparse representation'. Proc. Int. Conf. Hand-Based Biometrics (ICHB), 2011, pp. 1–4

9 Khorsandi, R., Cadavid, S., Mottaleb, M.: 'Ear recognition via sparse representation and Gabor filters'. Proc. IEEE Fifth Int. Conf. Biometrics: Theory, Applications and Systems (BTAS), 2012, pp. 278–282

10 Gong, M., Xu, Y., Yang, X., Zhang, W.: 'Gait identification by sparse representation'. Proc. Eighth Int. Conf. Fuzzy Systems and Knowledge Discovery (FSKD), 2011, pp. 1719–1723

11 Kumar, A., Chan, T., Tan, C.: 'Human identification from at-a-distance face images using sparse representation of local iris features'. Proc. Fifth IAPR Int. Conf. Biometrics (ICB), 2012, pp. 303–309

12 Wong, Y., Harandi, M., Sanderson, C., Lovell, B.: 'Expression recognition using elastic graph matching'. Proc. Int. Joint Conf. Neural Networks (IJCNN), 2012, pp. 1–8

13 Jittorntrum, K., Osborne, M.: 'Strong uniqueness and second order convergence in nonlinear discrete approximation', *Numer. Math.*, 1980, **34**, pp. 439–455

14 Huang, J., Huang, X., Metaxas, D.: 'Simultaneous image transformation and sparse representation recovery'. Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2008, pp. 1–8

15 Achlioptas, D.: 'Database-friendly random projections'. Proc. Twentieth ACM SIGMOD-SIGACT-SIGART Symp. Principles of Database Systems, 2001, pp. 274–281

16 Friedman, J., Tukey, J.: 'A projection pursuit algorithm for exploratory data analysis', *IEEE Trans. Comput.*, 1974, **C-23**, (9), pp. 881–890

17 Lu, G., Lin, Z., Jin, Z.: 'Face recognition using discriminant locality preserving projections based on maximum margin criterion', *Pattern Recognit.*, 2010, **43**, pp. 3572–3579

18 Kittler, J., Hatef, M., Duin, R., Matas, J.: 'On combing classifiers', *IEEE Trans. Pattern Anal. Mach. Intell.*, 1998, **20**, (3), pp. 226–239

19 Matej, K., Ales, L., Danijel, S.: 'Multivariate online kernel density estimation with Gaussian kernels', *Pattern Recognit.*, 2011, **44**, pp. 2630–2642

20 Martinez, A., Benavente, R.: 'The AR face database'. CVC Technical Report, 24, 1998

21 vd-Berg, E., Friedlander, M.P.: 'Probing the Pareto frontier for basis pursuit solutions'. UBC Computer Science Technical Report, TR-2008-01, 2008

22 Chen, S., Donoho, D., Saunders, M.: 'Atomic decomposition by basis pursuit', *SIAM Review*, 2001, **43**, (1), pp. 129–159

23 Posse, C.: 'Projection pursuit exploratory data analysis', *Comput. Stat. Data Anal.*, 1995, **20**, pp. 669–687

24 Shejin, T., Sao, A.: 'Significance of dictionary for sparse coding based face recognition'. Proc. Int. Conf. Biometrics Special Interest Group (BIOSIG), 2012, pp. 1–6

doi: 10.1049/iet-cvi.2014.0039