IR and Visible Face Identification via Sparse Representation

Pierre Buysens and Marinette Revenu
GREYC Laboratory – CNRS UMR 6072
ENSICAEN, University of Caen, Caen, France
pierre.buysens@greyc.ensicaen.fr

Abstract—We present a face recognition technique based on the sparsity principle. Parsimony is used both to compute the face feature vectors and to process the classification of these vectors. Applied to visible and infrared modalities on the Notre-Dame database, we show that this approach has equal or better performances than those of the state-of-art on this database. This classification allows to use a simple method to merge the scores of these two modalities in order to enhance significantly the identification rates. We show also that this approach is quite robust to corrupted probe images.

I. INTRODUCTION

Face recognition is a topic which has been of increasing interest during the last two decades due to a vast number of possible applications: biometrics, video-surveillance, advanced HMI or image/video indexation. One of the main challenge in face recognition for the visible light modality is the illumination changes in uncontrolled conditions. A way to tackle this problem, and then to increase the global recognition rate, is to use other modalities, like infrared light, conjointly with visible light. Infrared light allows furthermore the system to run even in bad lighting condition, like night for example.

A. Classical approaches of the task

Several approaches have been proposed to the problem of automatic face recognition. Most of them are built with the same two–steps scheme:

- extract relevant features from faces
- classify these features

While it can be difficult to characterize the features classifiers, mainly due to the vast number of different approaches, the feature extraction phase in the literature can be divided into two parts :

- the local approaches, which extract features and then combine them into a global model,
- the global approaches which take the image as a whole to realize often a kind of linear projection of the high–dimensional space (i.e. the face images) onto a low–dimensional space.

The local approaches first extract some local features (like the location of the eyes, nose or mouth) by the use of special feature extractors. The saliency of the extracted features relies then on the robustness of these extractors. The most popular local approach is the Elastic Bunch Graph Matching (EGM) where a set of interest points is extracted from the face, and then a graph is created. Brunelli and Poggio [3] used geometric models like the distance between pairs of feature points to achieve the face recognition. Wiskott et al.[15] used some Gabor filters on the neighborhood of these points to compute a set of jets to create the Elastic Bunch Graph Matching method (EBGM). Here the shape of the face is modeled into the jets to enhance the recognition.

The main drawback of these local approaches is their sensitivity to the features extractors. Even the best feature classifier will fail if the extractor is not well chosen. Moreover, it is difficult to deal with different scales and poses.

The global approaches often take the face image as a whole and perform a statistical projection of the images onto a face space. The most popular technique called Eigenfaces (first used by Turk and Pentland [14]) is based on a Principal Components Analysis (PCA) of the faces. It has also been applied to infrared faces by Chen et al. [6], Jung et al. [9] use it conjointly with a shape analysis of the face. Another popular technique is the Fisherfaces method based on a Linear Discriminant Analysis (LDA), which divides the face images into classes according to the Fisher criterion. It has been early applied by Kriegman et al.[10].

A comparison of these methods is made by Socolinsky and Selinger in [12], or by Wu et al. in [17] where a Discrete Cosine Transform is also tested.

Many classifiers have been used conjointly with these global approaches: simple ones like distances between features, others more complex like Neural Networks, Support Vector Machine or some cascade of classifiers.

The main drawback of the global approaches is their sensitivity to the illumination changes for the visible light modality, and the thermal distribution of the face over time for the infrared modality. When the illumination (or the thermal distribution) of a face changes, its appearance undergoes a non–linear transformation, and due to the linear projection often performed by these global approaches, the classification can fail.

B. Contribution of Sparsity to Face Identification

A sparse representation of an input signal refers to a representation of this signal as a linear combination of base elements in which many coefficients are equals to zero. A parallel can be drawn between this principle and face identification. Wright et al. are the first (to our knowledge) in [16] to use sparse representations to process the classification for face identification. A sparse representation of a face supposes that many coefficients of the decomposition over the gallery are zero, which then discards on first sight all these identities. In this paper, we propose to use a projection...
of a face onto a sparse dictionary as the feature extractor, and the sparse principle for the classification.

The paper is organized as follows: Section II describes the feature extraction process, classification method is explained in Section III, and experimental results are shown in Section IV. In Section V, we test the fusion of modalities, while in Section VI we confront our approach to corrupted images. Finally we present our conclusions and further work in Section VII.

II. FEATURES EXTRACTION

In order to extract relevant features, we decompose faces onto a dictionary, following a sparse scheme. The aim of the sparse coding algorithm is to find a representation $X \in \mathbb{R}^n$ for a given signal $Y \in \mathbb{R}^m$ by linear combination of an overcomplete set of basis vectors, which are the columns of a matrix $D \in \mathbb{R}^{m \times n}$ with $n > m$ [11]. These columns are often called atoms, and are noted $\phi_i$. In optimal sparse coding, the problem is formulated as:

$$\min_{X \in \mathbb{R}^n} \|Y - DX\|_2^2 \quad \text{s.t.} \quad X \in \mathbb{R}^n$$

(1)

where the $l^0$-norm is defined as the number of non-zero elements in a given vector. This problem is NP-hard; fortunately, under mild conditions, we can make a convex relaxation by turning the $l^0$-norm into a $l^1$-norm [7]. The problem can then be written as:

$$\min_{X \in \mathbb{R}^n} \left( \|Y - DX\|_2^2 + \lambda \|X\|_1 \right)$$

(2)

where $\lambda$ is a sparsity penalty term.

A lot of pre-defined dictionaries exist in the literature, such those based on wavelets, curvelets, ridgelets or DCT. Although these dictionaries are well suited for cartoon images, they are not very efficient to deal with textures. For our problem, it is more efficient to learn the dictionary directly from data. Starting from a random initialization of the atoms, learning the dictionary proceeds in an iterative way, alternating the two steps: 1) minimize Eq. 2 with respect to $X$ keeping $D$ constant, and 2) update the atoms of $D$ with $X$ found at previous step.

In this paper, we used for the two steps the OMP algorithm conjointly with the K–SVD algorithm respectively. The OMP algorithm (for Orthogonal Matching Pursuit) [13] is a greedy algorithm which selects atoms iteratively until the error reconstruction is low or the maximum number of atoms has been reached. The K–SVD algorithm [2] updates the atoms from the sparse representation provided by the first step. It is based on a Singular Value Decomposition, and is a generalization of the K–Means, hence its name.

III. CLASSIFICATION

A wide variety of approaches has been proposed to classify feature vectors. The popular subspace methods remain on the observation that the images of faces under varying lighting and expression lie on a special low–dimensional subspace [10], often called the face space. This is the assumption we have done in this work. We use a similar approach as the one presented in [16] to process the identification. Given a gallery with one image for each of the $n$ subjects, the matrix $A$ can be constructed by concatenating the $n$ feature vectors of gallery’s faces. In an optimal sparsity scheme, a test sample $y \in \mathbb{R}^m$ of class $k$ will then be decomposed into $x \in \mathbb{R}^n$, whose coefficients entries are equals to zero except the one associated with class $k$:

$$y = Ax \quad \text{with} \quad \|x\|_0 = 1$$

(3)

Unfortunately, this problem is hard to solve. It depends essentially on the matrix $A$ which represents the features of the gallery’s faces. Nevertheless, one can decompose the test image $y$ on $A$ into $x$ by relaxing the condition $\|x\|_0 = 1$, like as we have done at section II. The problem to solve becomes then:

$$\hat{x} = \arg\min \|x\|_1 \quad \text{s.t.} \quad y = Ax$$

or

$$\hat{x} = \arg\min \|x\|_1 \quad \text{s.t.} \quad \|y - Ax\|_2^2 < \epsilon$$

(4)

This is a typically lasso problem, for which many algorithms have been developed. We choose to process by an iterative soft–thresholding approach [8], which is efficient and fast. Once the solution has been computed, we have an estimate $\hat{y}$ of the test vector $y$ which is a linear combination of vectors of $A$:

$$\hat{y} = \sum_{i=1}^{n} x_i A_i$$

(5)

where $\hat{y}$ is the approximation of $y$, $A_i$ is the $i^{th}$ column of $A$, and most of $x_i$ are zeros. Finding the identity of $y$ is then processed by computing the residuals $r_i$ of $y$ for each feature vector $A_i$ of the gallery:

$$r_i(y) = \|y - A_i x_i\|_2$$

(6)

The smallest residual then corresponds to the vector $A_i$ that is the closest to $y$ in the meaning of $l^1$-norm:

$$\text{identity}(y) = \arg\min_i(r_i(y))$$

A schematic view of the classification process is shown on Fig. 1.

IV. EXPERIMENTS AND RESULTS

In order to test the approach, we used the Notre–Dame [1] (Collection X1) database (see Fig. 2 for samples of the database). It has the advantage to present images of subjects with two modalities, visible and infrared, taken at the same time. It can be divided into two parts: the first part, called Train set, is composed of 159 subjects who all have only one image in infrared light and its visible counterpart. The second part, called Test set, is composed of 82 subjects, for a total of 2292 infrared light images and 2292 visible light images.

While the train set contains no facial expression or head position variations, the test set is composed of several images containing variations in lighting, expressions, thermal changes and head positions. The test set is also divided into two experiments, called Same–session and Time–lapse which
mainly test the impact of illumination and facial expression changes in a short (minutes) and medium term (days or weeks) respectively. For each of these experiments, there are subsets named $F\{A,B\}L\{F,M\}$ which can be used for gallery or probe sets during the test. These subsets are:

- FA where faces have a neutral expression,
- FB where faces have a smiling expression,
- LF where faces are under the Feret Style Lighting,
- LM where faces are under a Mugshot Lighting.

In the rest of the paper, we assume that all the faces have been geometrically normalized according to the distance between the eyes, cropped and resized to $90 \times 110$, as we can see an example on Fig. 3.

### A. Learning of the Dictionary

In order to train the dictionary, we randomly extract 10000 patches of size $10 \times 10$ with sufficient standard deviation (to avoid too uniform patches) from the Train-set. The maximum number of atoms for the OMP algorithm has been fixed to 5, which means that each training pattern is decomposed into a sum of 5 atoms, the coefficients of the other atoms being 0. The redundancy of the dictionary has been set to 2 which means $2 \times 10 \times 10 = 200$ atoms to learn. The iterative process has been stopped after 100 iterations. A random selection of 100 atoms is presented Fig. 4. One can see that some atoms encode low frequency patterns, while others are more oriented edge selective.

### B. Creation of the Feature Vectors

Once the dictionary is learned, a face is then decomposed into non-recovering $10 \times 10$ patches. The size of the faces is $90 \times 110$, so there are 99 extracted patches. Each of these is then decomposed onto the dictionary, see Fig. 5. The decomposition consists on solving Eq. 2 without updating the atoms matrix $D$. In order to have a fast approximation of $X$, we used an iterative soft-thresholding approach [8] which minimizes $\|X\|_1$.

The $X$s of each patch are then stacked into one column vector to form the face feature vector. Since each patch is decomposed into a 200-dimensional vector, the size of the final face feature vector is $200 \times 99 = 19800$.

### C. Results of Identification

In order to test the approach, we used the imagelists provided with the database. The tests can be divided into two experiments: the Same-session and the Time-lapse experiments which mainly test the impact of illumination and facial expression changes in a short (minutes) and medium term (days or weeks) respectively. In both experiments, there is only one image per subject in the gallery, acting like a 1–image–to–enroll scenario. The Same-session experiment is composed of:
The Time–laplace experiment is composed of:
- 4 galleries, and 4 probe sets
- gallery sets: 1 image for each of the 63 subjects
- probe sets: 431 images of the 63 subjects.

The results for the Same–session experiment (Table I), which is an easy test, are quite the same as those given in [4] based on a Convolutional Neural Network, or those in [5] using PCA. However, there is a significant improvement of recognition rates for the Time–laplace experiment (Table II), especially for the visible modalities. The method is then more robust to illumination changes than to thermal distribution changes.

V. Fusion

Results at section IV-C show that visible modality performs better than IR. This result has already been shown in [4] and [5]. However the mismatched probes of the two classifiers do not necessarily overlap. This suggests that the two modalities could offer complementary information about a probe face. A merging scheme then could enhance identification rates. Since the classifiers for the two modalities yield decision rankings as results, we chose to merge the results on the decision level. We have tested some algorithms like the one presented in [4], which realizes a weighted sum of the scores of the two modalities according to a measure of saliency computed dynamically. Nevertheless, we found that the simple sum rule on the residuals gives the best results. That is, for a probe image \( y \), each residuals \( r_{v_k} \) and \( r_{i_k} \) to sample \( k \) in the galleries for the visible and IR modalities are computed. Finals residuals for a sample \( k \) in the gallery are then:

\[
r_k = r_{v_k} + r_{i_k}
\]

The smallest residual then correspond to the identity of \( y \):

\[
identity(y) = \arg\min_k (r_k(y))
\]

Results of fusion scores for the Same–session and Time–laplace experiments are presented in Tables III and IV respectively. They show that the fusion scheme always improves the best result of one modality alone. They are also always better than those given in [4] and [5] (summary shown in Table V).

VI. Tests on Corrupted Images

In order to test the robustness of our approach, we apply two types of degradation on the probe images. Only the probe images are corrupted, not the images of the galleries. We apply the same protocol as above: 1) decomposition of the images onto the dictionary and 2) classification via
minimization of the $l^1$-norm. For the two types of corruption, we used the same test sets as above.

### A. Noisy probes

In this experiment, we corrupt the images by adding some gaussian noise. The standard deviation of the gaussian distribution is computed according to a ratios of the dynamic of the original image. The ratios we used are 10%, 20%, 30%, 40%, and 50%. An example of these noises is shown on Fig. 6. Results for the Same–session and Time–lapse experiments are shown on Fig. 7 and 8 respectively. These figures show the mean rank–0 identification rates of the 12 and 16 sub–experiments for the two experiments according to the amount of ‘missing pixels’ in the probe images.

We can see that the visible modality resists far better to missing pixels than infrared modality, which rank–0 identification rates quickly decrease.

### VII. CONCLUSION AND FUTURE WORK

We presented a face recognition method for visible and infrared light imagery. Based on the sparsity theory, it decomposes a face onto a dictionary that has been learned from data. Identification is then processed by considering this feature vector as a linear combination of the gallery’s feature vectors with the minimization of the $l^1$–norm as criterion. Results on the Notre–Dame database for the Time–lapse experiment are always better than the state-of-art (see Table V). Moreover, we show that a simple score fusion of the two modalities enhances always and significantly the identification rates. We also show that this approach is quite robust to restricted corruptions applied to the probe images. We are conducting experiments to adapt this method to bigger gallery sizes, to quantify the contribution of a multiscale sparse decomposition of faces, and to construct a cascade of sparse classifiers. Learning a rotation invariant dictionary could also help in case of non–frontal head pose.
Fig. 8. Results for the ‘noisy’ *Time-lapse* experiment.

Fig. 9. Percentage of ‘missing pixels’ in a probe image.

Fig. 10. Results for the ‘missing pixels’ *Same-session* experiment.

**REFERENCES**


