IR and Visible Face Identification via Sparse Representation

Pierre Buyssens, Marinette Revenu

IR and Visible Face Identification via Sparse Representation

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

***Abstract*— We present a face recognition technique based on the sparsity principle. Parsimony is used both to compute the face feature vector and to process the classification of these vectors. Applied to visible and infrared modalities on the *Notre–Dame*, we showed 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 condition. 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. Another advantage of infrared light allows furthermore the system to run even in bad lighting condition, like night.

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 different approaches, the feature extraction phase in the litterature 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 relie then on the robustness of these extractors. The most popular local approach is the *Elastic 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 an analyse of the shape 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 applied early 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 distance 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 performed by the 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 of the coefficients are 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 discard 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 follow: 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 ϕ_i. In optimal sparse coding, the problem is formulated as :

$$\min \|X\|_0 \ s.t. \ Y = DX \tag{1}$$

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 :

$$\min_{X \in \mathbb{R}^m} \|Y - DX\|_2^2 + \lambda \|X\|_1 \tag{2}$$

where λ is a sparsity penalty term.

A lot of pre–defined dictionary exists in the litterature, 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 subspace*. 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 concatening 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 zero except the one associated with class k:

$$y = Ax \ with \ \|x\|_0 = 1 \tag{3}$$

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 then becomes :

$$\begin{aligned} \hat{x} &= \arg\min \|x\|_1 \ s.t. \ y = Ax \\ \text{or} & \\ \hat{x} &= \arg\min \|x\|_1 \ s.t. \ \|y - Ax\|_2^2 < \epsilon \end{aligned}$$

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:

$$\hat{y} = \sum_{i=1,n} x_i A_i \tag{4}$$

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 :

$$r_i(y) = \|y - A_i x_i\|_2 \tag{5}$$

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

$$identity(y) = \arg\min_i (r_i(y)) \tag{6}$$

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 *Training 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 expressions or head positions 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

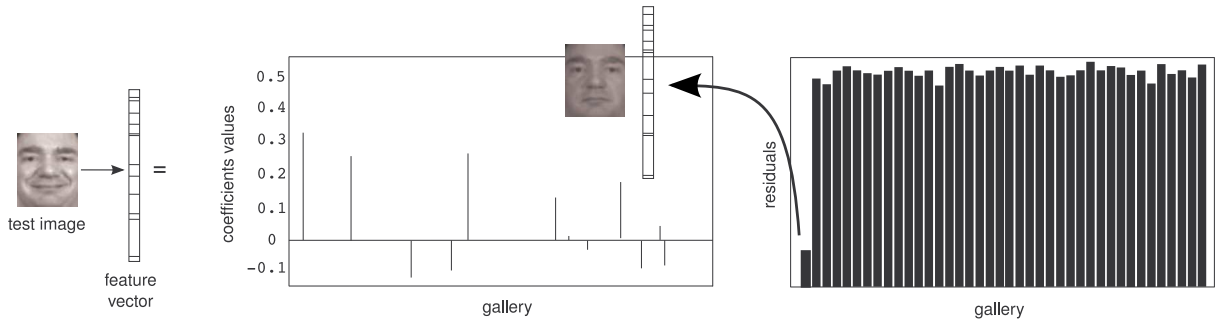


Fig. 1. Schematic view of the classification process: the feature vector of a probe face is decomposed onto the gallery feature vectors. From this decomposition, residuals are computed, and identity is deduced from the minimum residual.

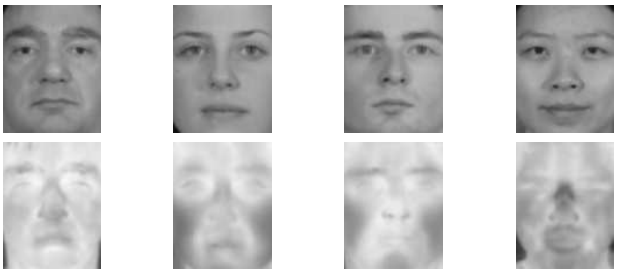


Fig. 2. Samples of the database for the Visible and IR modalities

two parts, called *Same–session* and *Time–lapse* sets in order to test the lighting problem, and the recognition through time respectively. For each of these subsets, there are files named *f*{*a*,*b*}*l*{*f*,*m*} which can be used for gallery or probe sets during the test. These subsets have been designed to test independantly the effect of a facial expression (*fa*: neutral expression, *fb*: smiling expression), under different lighting (*lf*: *Feret style lighting*, *lm*: *mugshot lighting*).

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

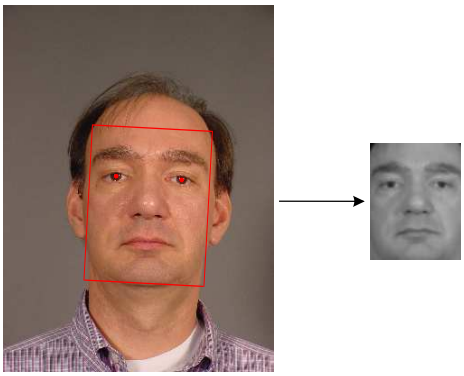


Fig. 3. Geometric preprocessing of the images

A. Learning of the Dictionary

In order to train the dictionary, we randomly extract 10000 patches of size 10×10 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 10×10 patches. The faces are of size 90×110, 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 Xs 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 final face feature vector is of size $200 \times 99 = 19800$.

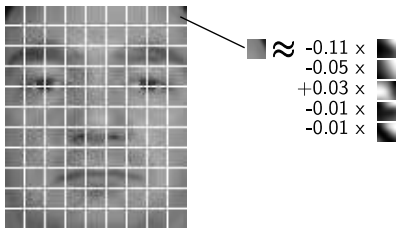


Fig. 5. Decomposition of a face image.

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:

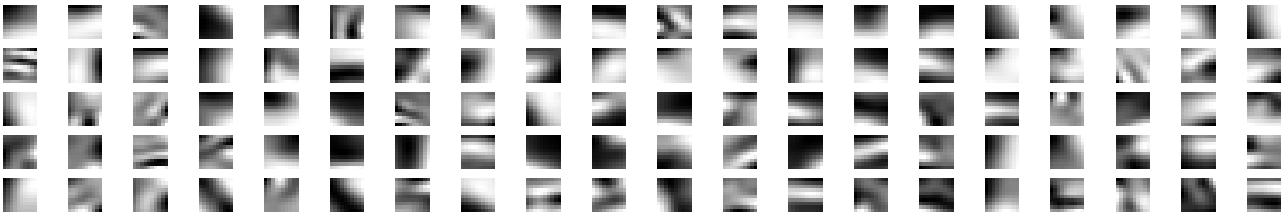


Fig. 4. Random selection of 100 atoms learned.

TABLE I
RANK–0 RECOGNITION RATES FOR THE *Same–Session* EXPERIMENT.
TOP: VISIBLE, BOTTOM: IR.

Gallery \ Probe	FA|LF	FA|LM	FB|LF	FB|LM
FA|LF		1.00 0.98	1.00 0.97	0.98 1.00
FA|LM	0.98 0.96		1.00 0.95	0.98 0.96
FB|LF	0.97 1.00	0.97 0.92		1.00 0.97
FB|LM	0.98 0.98	0.98 0.97	1.00 0.98	

TABLE II
RANK–0 RECOGNITION RATES FOR THE *Time–Lapse* EXPERIMENT.
TOP: VISIBLE, BOTTOM: IR.

Gallery \ Probe	FA|LF	FA|LM	FB|LF	FB|LM
FA|LF	0.95 0.83	0.92 0.79	0.87 0.76	0.87 0.77
FA|LM	0.95 0.83	0.93 0.81	0.87 0.77	0.85 0.77
FB|LF	0.86 0.77	0.83 0.74	0.93 0.79	0.91 0.80
FB|LM	0.92 0.79	0.87 0.80	0.88 0.78	0.88 0.82

- 4 sets used as galleries and probes
- sets: 1 image for each of the 82 subjects.

The *Time–lapse* 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.

Results for the two experiments are presented Tables I and II.

The results for the *Same–session* experiment, 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–lapse* experiment.

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 sets of mismatched probes of the two classifiers do not necessarily overlap. This suggests that the two modalities could offer complementary informations 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 residual gives the best results. That is, for a probe image y, each residuals r_{v_k} and r_{i_k} of sample k in the galleries for the visible and ir modalities are computed. Finals residuals for a sample k in the gallery are :

$$r_k = r_{v_k} + r_{i_k} \tag{7}$$

The smallest residual then correspond to the identity of y :

$$identity(y) = \arg\min(r_k(y)) \tag{8}$$

TABLE III
RANK–0 RECOGNITION RATES FOR THE *Same–Session* EXPERIMENT.
TOP: VISIBLE, MIDDLE: IR, BOTTOM: FUSION.

Gallery \ Probe	FA|LF	FA|LM	FB|LF	FB|LM
FA|LF		1.00 0.98 **1.00**	1.00 0.97 **1.00**	0.98 1.00 **1.00**
FA|LM	0.98 0.96 **1.00**		1.00 0.95 **1.00**	0.98 0.96 **1.00**
FB|LF	0.97 1.00 **1.00**	0.97 0.92 **1.00**		1.00 0.97 **1.00**
FB|LM	0.98 0.98 **0.98**	0.98 0.97 **1.00**	1.00 0.98 **1.00**	

Results of fusion scores for the *Same–session* and *Time–lapse* experiments are shown 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].

VI. TESTS ON CORRUPTED IMAGES

In order to test the robustness of our approach, we apply two types of degradation to the probe images. Only the probe images are corrupted, not the images from 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.

TABLE IV

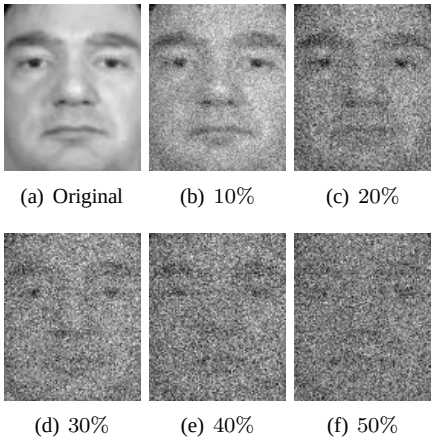
RANK–0 RECOGNITION RATES FOR THE *Time–Lapse* EXPERIMENT.

TOP: VISIBLE, MIDDLE: IR, BOTTOM: FUSION.

Gallery \ Probe	FA|LF	FA|LM	FB|LF	FB|LM
FA|LF	0.95	0.92	0.87	0.87
	0.83	0.79	0.76	0.77
	0.98	**0.96**	**0.94**	**0.93**
FA|LM	0.95	0.93	0.87	0.85
	0.83	0.81	0.77	0.77
	0.99	**0.97**	**0.95**	**0.92**
FB|LF	0.86	0.83	0.93	0.91
	0.77	0.74	0.79	0.80
	0.93	**0.91**	**0.97**	**0.95**
FB|LM	0.92	0.87	0.88	0.88
	0.79	0.80	0.78	0.82
	0.97	**0.93**	**0.95**	**0.95**

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 ratio of the dynamic of the original image. The ratio we used are 10%, 20%, 30%, 40%, and 50%. An example for 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 for the 12 and 16 sub–experiments of the two experiments according to the amount of noise in the probe images. The standard deviation to each measure is quite similar to those reported in section IV-C, so it has not been included in these figures to avoid burdening them.



(a) Original (b) 10% (c) 20%

(d) 30% (e) 40% (f) 50%

Fig. 6. Percentage of noise in a probe image.

As we might expect, identification rates decrease while noise strength increases. Moreover, this decrease is quite linear and is not significantly different for the two modalities.

B. *"Missing pixels" probes*

In this experiment, we corrupt the images by "removing" a ratio of pixels of the image. The value of these pixels is set to 0. We used ratios from 10% to 90% with a step of 10%. An example for this corruption is shown on Fig. 9. Results

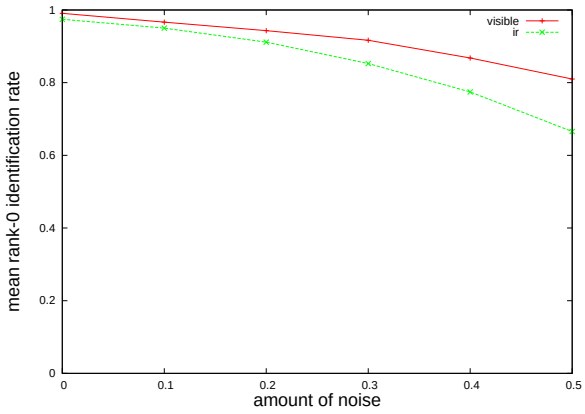


Fig. 7. Results for the "noisy" *Same–session* experiment.

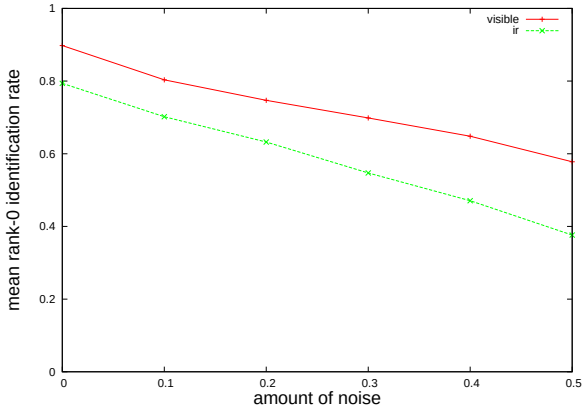


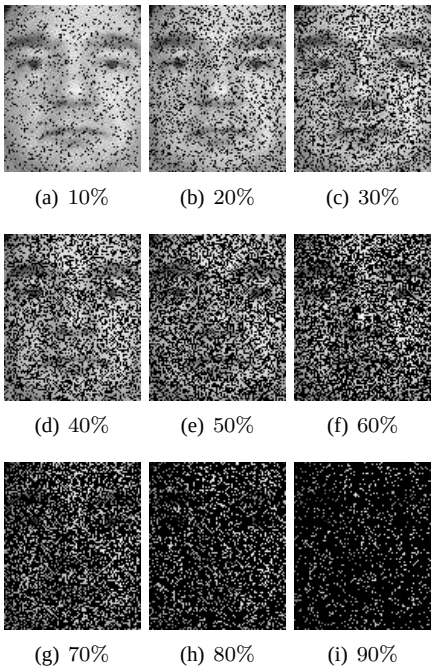
Fig. 8. Results for the "noisy" *Time–lapse* experiment.

for the *Same–session* and *Time–lapse* experiments are shown on Fig. 10 and 11 respectively. These figures show the mean rank–0 identification rates for the 12 and 16 sub–experiments of 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 as criterion the minimization of the l^1–norm. Results on the *Notre–Dame* database for the *Time–lapse* experiment are always better than the state-of-art (see Tab. V). Moreover, we show that a simple scores 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 galleries sizes, to quantify the contribution of a multiscale sparse decomposition of faces, and to construct



(a) 10% (b) 20% (c) 30%

(d) 40% (e) 50% (f) 60%

(g) 70% (h) 80% (i) 90%

Fig. 9. Percentage of "missing pixels" in a probe image.

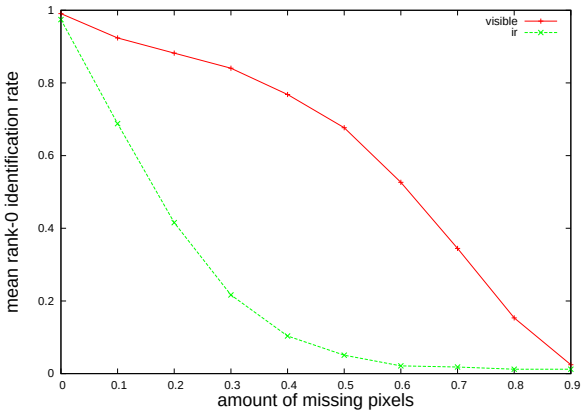


Fig. 10. Results for the "missing pixels" *Same–session* experiment.

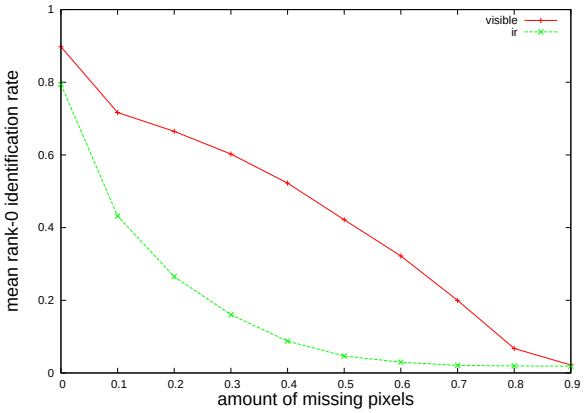


Fig. 11. Results for the "missing pixels" *Time–lapse* experiment.

TABLE V

COMPARISON OF METHODS FOR THE TWO EXPERIMENTS. MEAN RECOGNITION RATE OVER THE 12 (OR 16 SUB–EXPERIMENTS) AND STANDARD DEVIATION IN PARENTHESIS. BEST SCORE IN BOLD.

	same-session			time-lapse		
	[5]	[4]	this paper	[5]	[4]	this paper
Visible	97.08 (3.13)	98.41 (1.97)	**98.66** (1.17)	82.66 (7.75)	72.50 (4.01)	**89.31** (3.56)
IR	**97.41** (2.01)	90.5 (4.27)	97.00 (2.08)	77.81 (3.31)	40.06 (3.47)	**78.87** (2.46)
Fusion	N/A	**100.** (0.)	99.83 (0.54)	92.5 (2.71)	80.12 (4.13)	**95.00** (2.17)

a cascade of sparse classifiers. Learning a rotation invariant dictionary could also help in case of non–frontal head pose.

REFERENCES

[1] http://www.nd.edu/cvrl/undbiometricsdatabase.html.

[2] M. Aharon, M. Elad, and A. Bruckstein. K–svd: Design of dictionaries for sparse representation. *IEEE Transactions On Signal Processing*, 2006.

[3] R. Brunelli and T. Poggio. Face recognition: Features versus templates. *IEEE Pattern Analysis and Machine Intelligence*, 1993.

[4] P. Buyssens, M. Revenu, and O. Lepetit. Fusion of ir and visible light modalities for face recognition. In *Biometrics: Theory, Applications and Systems*, 2009.

[5] X. Chen, P. J. Flynn, and K. W. Bowyer. IR and visible light face recognition. *Computer Vision and Image Understanding*, 2005.

[6] Xin Chen, Patrick J. Flynn, and Kevin W. Bowyer. PCA-based face recognition in infrared imagery: Baseline and comparative studies. In *AMFG*. IEEE Computer Society, 2003.

[7] David L. Donoho. Sparse components of images and optimal atomic decomposition. Technical report, Department of Statistics, Stanford University, 1998.

[8] M.J. Fadili and J.L. Starck. Sparse representation-based image deconvolution by iterative thresholding. In *Astronomical Data Analysis ADA'06*, 2006.

[9] Soon-Won Jung, Youngsung Kim, Andrew Jin Tech, and Kar-Ann Toh. Robust identity verification based on infrared face images. In *International Conference on Convergence Information Technology*, 2007.

[10] D. J. Kriegman, J. P. Hespanha, and P. N. Belhumeur. Eigenfaces vs. fisherfaces: Recognition using class-specific linear projection. In *European Conference on Computer Vision*, 1996.

[11] B. A. Olshausen and D. J. Field. Sparse coding with an overcomplete basis set: A strategy employed in V1? *Vision Research*, 1997.

[12] Diego A. Socolinsky and Andrea Selinger. Thermal face recognition in an operational scenario. In *Computer Vision and Pattern Recognition*, 2004.

[13] Joel A. Tropp. Greed is good: algorithmic results for sparse approximation. *IEEE Transactions on Information Theory*, 2004.

[14] M. A. Turk and A. P. Pentland. Face recognition using eigenfaces. In *IEEE Computer Vision and Pattern Recognition*, 1992.

[15] L. Wiskott, J. M. Fellous, N. Kruger, and C. von der Malsburg. Face recognition by elastic bunch graph matching. *IEEE Pattern Analysis and Machine Intelligence*, 1997.

[16] J. Wright, A.Y. Yang, A. Ganesh, S.S. Sastry, and Y. Ma. Robust face recognition via sparse representation. *IEEE Pattern Analysis and Machine Intelligence*, 2008.

[17] Shi-Qian Wu, Li-Zhen Wei, Zhi-Jun Fang, Run-Wu Li, and Xiao-Qin Ye. Infrared face recognition based on blood perfusion and sub–block dct in wavelet domain. In *International Conference on Wavelet Analysis and Pattern Recognition*, 2007.