Fusion Levels of Visible and Infrared Modalities for Face Identification

Pierre Buyssens, Marinette Revenu

Fusion Levels of Visible and Infrared Modalities for Face Recognition

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 study on different levels of visible and infrared modalities fusion for face recognition. While visible modality is the most natural way to recognize someone, infrared presents thermal distribution that can be useful for face recognition. We compare the well–known eigenfaces method as a baseline to an approach based on sparsity for the feature extraction and the classification. Applied on the *Notre-Dame* database, we showed that the three levels of fusion considered are not equivalent in term of final identification rates. We also show that the sparse approach at the decision level outperforms the state-of-art on this database.**

I. Introduction

Although considerable progress has been made in the domain of face recognition over the last decade, especially with the development of powerful methods (such as the Eigenfaces or the Elastic Bunch Graph Matching methods), face recognition has shown to be not accurate enough in uncontrolled environments. Face recognition performances of a system can be degraded by many factors, including facial expression, head pose variation, occlusion and most importantly illumination changes.

In the particular case of uncontrolled illumination, previous studies have demonstrated that infrared imagery can be a promising alternative to visible imagery. An infrared capture of a face is nearly invariant to illumination changes, and allows a system to process in all the illumination conditions, including total darkness like night.

Despite these advantages, infrared imagery has other limitations. Since a face captured under this modality renders its thermal patterns, a temperature screen placed in front of the face will totally occlude it. This phenomena appears when a subject simply wears glasses. In this case, the captured face has two black holes, corresponding to the glasses, which is far more unconvenient than in the visible modality.

However, since these two modalities do not present the same advantages/limitations, using informations of both can decrease the disadvantages of each and globally enhance the recognition rates.

This paper adresses the question of how to fuse these modalities. Three different levels of fusion have been considered, an image-based level, a feature-based level and a score level. This paper adresses more particularly the question of robustness and feasability of the proposed algorithm across these levels, compared to the state-of-art *eigenfaces* algorithm.

A. Overview of Face Recognition Technics

Most of the approaches that have been proposed in the litterature for the problem of face recognition are built with the same three–steps scheme:

- preprocessing of the images
- extraction of features from faces
- classification of these features

The preprocessing step intends to locate a face, resize it if necessary and apply algorithms to enhance the quality of the images. Many systems use some illumination correction algorithms to simplify the features extraction step.

The most important step is the extraction of salient features from the faces. Two main strategies can be considered for this step:

- 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 (in this case, these technics are called *Subspace* methods).

A well–known local approach is the *Elastic Bunch Graph Matching* [10] method where interest points are extracted from the face. These points may then be treated as a weighted graph, local features extractors (like the widely used Gabor filters) may be applied in the neighborhood of these points to enhance the robustness of the final features. The main drawback of these local approaches is their sensitivity to the features extractors. Moreover, it is difficult to deal with different scales and poses. The eigenfaces [13] and fisherfaces [8] methods are probably the most popular global methods. Based on a Principal Component Analysis (PCA) and on a Linear Discriminant Analysis (LDA) respectively, these methods belong to the class of the *Subspace method*. This class of algorithms relies on the assumption that faces span a small area in the image space (called the *face space*). The aim of this class of algorithms is often to find a discriminative projection that maps faces onto this face space. The main drawback of the global approaches is the sensitivity of the projection to the illumination changes for the visible light modality, and the thermal distribution of the face over time for the infrared modality.

The last step intends to classify the extracted features. There are plenty of technics, simple ones based on distances between features, others based on learning methods such Support Vector Machine or Neural Networks. All these technics have their own advantages/disadvantages, their results depending in fact often on the robustness of the previously extracted features.

B. Multimodal Fusion technics

Some work has been realized in the face fusion technics domain. We can mainly divide it into three parts, according to the level of fusion applied :

At the sensor (or image) level, the fusion is realized in [7] as a weighted sum of the face patterns. The main challenge of this level of fusion is the high precision required (pixel–level typically). In [11], a multi–resolution fusion of images is achieved by merging wavelet coefficients obtained by the Haar transform. This kind of fusion mainly tries to bypass the illumination and the eyeglasses problem.

At the feature level, extracted features from different modalities are merged to create a single feature vector. In [11], the eigenfeatures (obtained with the eigenfaces method) are merged with the use of a Genetic Algorithm. A hierarchy of features is learned in [9] with a specific framework to produce local features combinations.

At the decision (or score) level, the distances (or scores) of each of the two modalities probe images are merged. This is often realized in two steps: 1) the scores are first transformed to make them comparable, by applying linear, logarithmic or exponential rules [5]. In [4], these scores are weighted according to a measure of saliency, depending on the distribution of the distances. 2) the transformed scores are combined into one final score. This is classically realized with a sum rule, which has been demonstrated to be efficient, or by a weighted sum, with fixed or dynamic weights.

The rest of the paper is organized as follow : Section II recalls the two approaches we have tested in this paper (the eigenfaces method and a sparse approach). Section III details the *Notre–Dame* Database (UND) and the results we obtained for the three levels of fusion we have distinguished. Finally section IV summarizes and compares previous results on the UND database, and we present our conclusions in section V.

II. Tested Face Recognition Approaches

In this paper we have tested two different approaches to perform the recognition. The classical eigenfaces method and a sparse approach described in [3] and modified to process the identification faster, while improving slightly the results of identification. We now briefly recall the principle of these two approaches.

A. The eigenfaces method

The *Eigenfaces* method based on a Principal Component Analysis is one of the most popular method in face recognition. In this section, we briefly recall its principle. We also describe the preprocessing of the images and the important point of the choice of the distance measure.

1) Training and Projection: The *eigenfaces* algorithm aims to find a basis which maximizes the recovery of a sample according to the variance of its vectors. These eigenvectors are computed from the total scatter matrix of training samples. Then only relevant eigenvectors (those with the largest corresponding eigenvalues) are kept to form the projection basis. There are different ways to choose the number of eigenvectors retained. In this work, we choose to keep the eigenvectors whose eigenvalues represent at least 90% of the total energy (the sum of eigenvalues). Once the basis has been found, a face sample is first mean centered and then projected onto the basis. The projections are then the feature vectors of the face (and are sometimes called eigenfeatures).

2) Distance Measure and Classification: In order to compare the projected vectors of two images in the face space, we have to compute a distance between these vectors. Many distances have been tested such the Euclidean distance, the CityBlock distance or the Mahalanobis distance, which is computed as the distance between values of vectors taking into account the correlation between these values :

$$D_M(\mathbf{x},\mathbf{y}) = \sqrt{(\mathbf{x}-\mathbf{y})^T S^{-1}(\mathbf{x}-\mathbf{y})} \qquad (1)$$

We found that the Mahalanobis distance offered the best performance, greater than 10% in mean compared to other distances. This is then the Mahalanonis distance we chose for all results of PCA experiments presented in this paper.

B. Sparse Approach

The sparse approach (detailed in [3]) is based on the sparsity theory both for the features extraction and the classification. However, this method has been modified to have a faster computation while improving slightly the results of identification.

1) Learning of the Dictionary: In order to extract relevant features, we decompose faces onto a dictionary, following a sparse scheme. Although pre–defined dictionaries exist in the litterature (wavelets, curvelets, ridgelets or DCT), dealing with texture is more efficient with a dictionary that has been learned from data. In this work, we have learned three different dictionaries depending on the modality of the images (visible, infrared or the fusion of both). The size of the atoms has been fixed to 10×10. All these dictionaries have been learned with efficient algorithms such the OMP algorithm (for Orthogonal Matching Pursuit) [12] for the inversion of the input, and the K–SVD algorithm [2] to update the atoms. In each case, randomly extracted patches have served as the train set. A random selection of 50 atoms of the 200 learned from the image fusion of visible and infrared data is presented Fig. 1. One can see that some atoms encode low frequency patterns, while others are more oriented edge selective.

2) Feature Extraction: 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. 2. In order to have a fast approximation

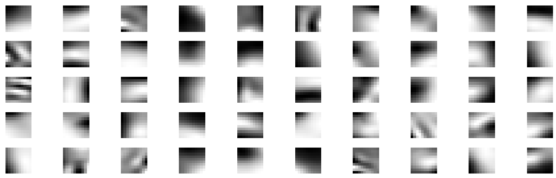

Fig. 1. Random selection of 50 atoms learned on the low–level fusion images.

of the sparse vector from a patch, we used an iterative soft–thresholding approach [6].

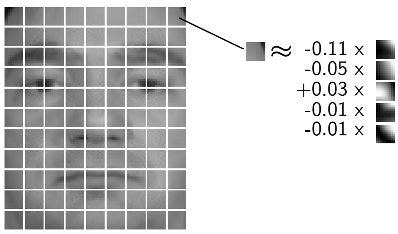

Fig. 2. Sparse decomposition of a face image.

Fig. 4. Geometric preprocessing of an image.

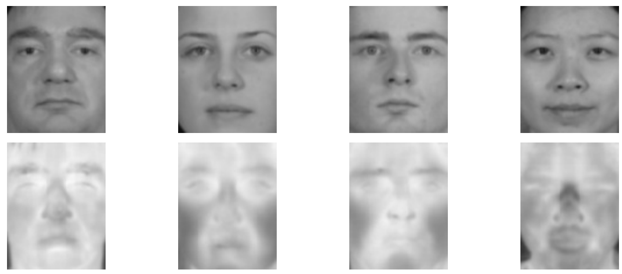

Fig. 5. (Preprocessed) Samples of the database for the Visible and IR modalities.

3) Classification: In order to perform the classification, we use a similar approach as the one presented in [14] or in [3]. A schematic view of the process is shown in Fig. 3. Each patch of a probe face image is processed independantly. Its sparse vector is assumed to span the corresponding sparse vectors of the gallery. A vector of residuals is then computed regarding the gallery. Each vector of residuals (one per patch) is then normalized between 0 and 1, and all these residual vectors are summed into one final residual vector. The identity returned by the system is then the one which corresponds to the minimum of this final residual vector.

Preprocessing of the images. The necessary steps for the preprocessing of the images are :

- the images have been cropped and rescaled to the size 90×110. This geometric normalization has been realized according to the distance between the eyes,
- an elliptical mask centered just below the eyes has been applied (PCA approach only) to obscur irrelevant parts of the faces (essentially the corners of the images),
- the pixel values have been normalized to ensure a mean pixel value of 0.0 and a standard deviation of 1.0. Note that, for the PCA approach, this transformation has been applied only on the 'visible' pixels, not on those that have been masked. For the sparse approach, all the pixels have been taken into account.

An example of the preprocessing of a visible face image for the subspace approach is shown on Fig. 4. Note that the preprocessing applied for the sparse approach is simpler and then faster than the one applied for the PCA approach.

III. Results of Fusion

In this section, we propose to study the impact of the level of fusion on the identification rates on the *Notre–Dame* database. Since identification proceeds in a classical two steps way (1. Extract features from an image, 2. Compute a distance between features in order to classify the image), we have distinguished three different levels of fusion :

- a low–level fusion, also called a sensor or image fusion, where images from the modalities are merged before any feature extraction,
- a mid–level fusion, also called features fusion, where features are extracted separately for the two modalities, and are then concatenated,
- a high–level fusion, also called decision level fusion, the features extraction and the distances to the gallery for probes are computed separately depending on the modalities. The distances are then merged.

Although the last is often preferred due to its simplicity and flexibility, 'earlier' fusion scheme have been less studied and may offer interesting alternatives. We now briefly recall the architecture of the *Notre–Dame* database, and in the rest of the paper, only the results of the fusion of visible and infrared modalities are shown. They present the identification rates for the *Same–session* and *Time–lapse* experiments for all the differents gallery–probe combinations of the *Notre–Dame* database.

A. Details of the Database

In order to test the approach, we used the *Notre–Dame* [1] (Collection $X1$) database (see Fig. 5 for samples of the database). The main advantage of this database is to present images of subjects both in visible and infrared modalities. Although corresponding images have not the same resolution, they are taken at the same time, which is usefull in order to test the gain of infrared modality accross illumination changes.

The database can be divided into two distinct parts : the first one, called *Training set*, is composed of 159 pairs of visible/infrared images for a total of 159 subjects. The second

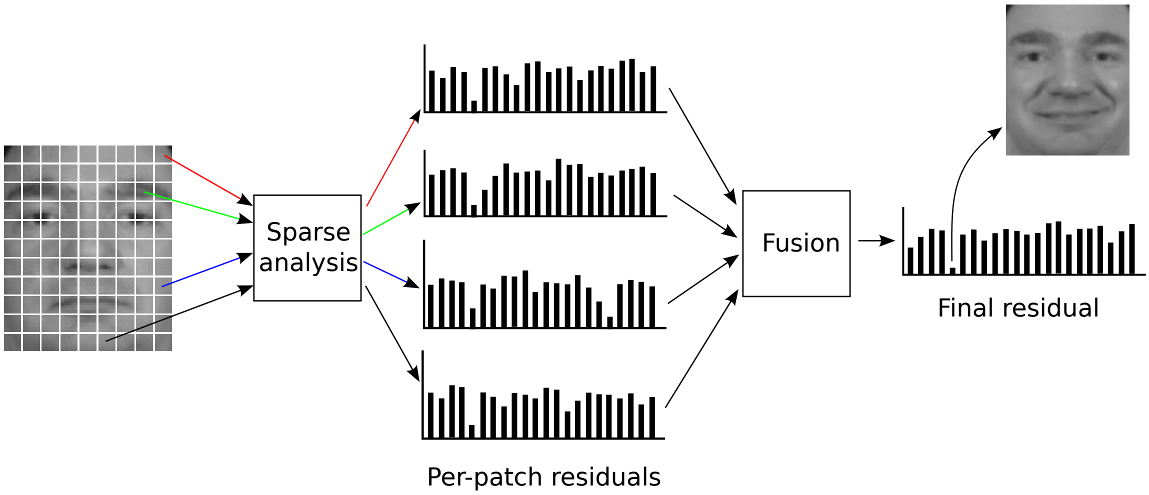

Fig. 3. Schematic view of the classification process: Each patch is processed separately through the sparse process. The final residual is the sum of normalized per-patch residuals.

one, called *Test set*, is composed of 82 subjects, for a total of 2292 image pairs.

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.

Two experiment protocols have been designed:

- In the *Same–session* experiment, gallery and probe images have been taken within seconds. There is no major changes in the thermal distribution, so this experiment is mainly usefull to test the effect of illumination on algorithms. 4 subsets are avalaible, serving as gallery or probe depending on the subexperiment that is conducted.
- In the *Time–lapse*, gallery and probe images have been taken within minutes, days or weeks. As images are taken at different time, the thermal distributions of infrared faces for a same subject have sometimes a great variation. 4 gallery and 4 probe subsets are avalaible. This experiment has mainly been designed to quantify the robustness of infrared through time.

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:

- 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.

The avalaible subsets are named $F\{A,B\}L\{F,M\}$ and are composed of :

- 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.

B. Low Level of Fusion

In the image fusion approach, we made the assumption that visible and infrared images have been taken at the same time. Visible images have the advantage to be more natural than IR ones, they present the texture of the face, but may be subject to illumination problems. Infrared images show a thermal distribution of a face, so they are not subject to illumination, but they are a far less natural way to identify a person. Nevertheless, we propose to merge them following a particular scheme :

- normalize the pixels values of the infrared face I_i between 0 and 1,
- multiply each pixel value of the visible face I_v by its infrared counterpart I_i to obtain the merged image I_f: $I_f(x,y) = I_v(x,y) \times I_i(x,y)$.

An example of such a fusion is shown in Fig. 6 for the two different preprocessing. This multiplicative way applied for the image fusion allows to take more into the thermal distribution of the infrared face than a simple sum of pixels of the two modalities.

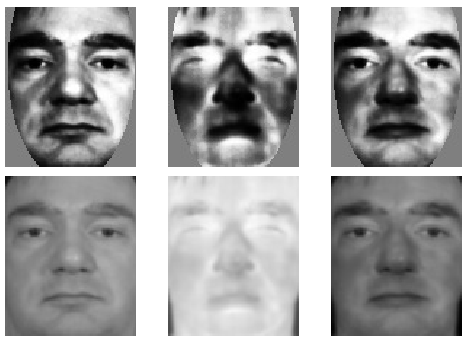

Fig. 6. Examples of image fusion. Rows differs from the preprocessing that has been applied. Left:Visible, Center:Infrared, Right: Fusion.

This fusion scheme has been applied to all the subsets of the database, especially to the *Train set* from which the

TABLE I

RANK–0 RECOGNITION RATES FOR THE LOW–LEVEL FUSION.
TOP TABLE: *Same-session*, BOTTOM TABLE: *Time-lapse* EXPERIMENT.
IN EACH CELL, TOP: PCA, BOTTOM: SPARSE APPROACH.

Gallery \ Probe	FA|LF	FA|LM	FB|LF	FB|LM
FA|LF		0.98 0.98	0.97 0.98	0.96 0.97
FA|LM	0.98 1.00		0.91 0.98	0.95 1.00
FB|LF	0.96 0.98	0.93 1.00		0.97 1.00
FB|LM	0.95 0.97	0.95 1.00	0.98 1.00	

Gallery \ Probe	FA|LF	FA|LM	FB|LF	FB|LM
FA|LF	0.70 0.91	0.68 0.88	0.55 0.85	0.58 0.85
FA|LM	0.70 0.90	0.67 0.88	0.58 0.84	0.60 0.84
FB|LF	0.61 0.83	0.58 0.82	0.63 0.97	0.65 0.89
FB|LM	0.64 0.88	0.61 0.82	0.67 0.80	0.64 0.90

TABLE II

RANK–0 RECOGNITION RATES FOR THE MID–LEVEL FUSION.
TOP TABLE: *Same-session*, BOTTOM TABLE: *Time-lapse* EXPERIMENT.
IN EACH CELL, TOP: PCA, BOTTOM: SPARSE APPROACH.

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

Gallery \ Probe	FA|LF	FA|LM	FB|LF	FB|LM
FA|LF	0.81 0.98	0.80 0.96	0.65 0.95	0.66 0.92
FA|LM	0.82 0.98	0.79 0.96	0.66 0.95	0.67 0.93
FB|LF	0.66 0.93	0.65 0.92	0.77 0.97	0.76 0.97
FB|LM	0.68 0.96	0.68 0.93	0.79 0.96	0.77 0.95

dictionary was learned. Results with this low–level fusion scheme for the *Same–session* and the *Time–lapse* experiments are shown successively in Tab. I. Results for the *Same–session*, which is an easy test, are quite similar for the two approaches. The *Time–lapse* results show that the PCA method performs poorly. We think this is due to the fact that illumination changes are amplified by our multiplicative merging technique based on the thermal distribution of the infrared face. The sparse approach gives decent results. This level of fusion can then become a realist alternative to other levels of fusion.

C. Middle Level of Fusion

Here we have processed the features extraction separately for the two modalities, and have realized a middle–level fusion by concatenating the features extracted from visible and infrared images. The feature vector for the PCA approach is a vector of size m, where m is the number of eigenvectors retained during the PCA. Given two vectors of size m_v and m_i for the visible and infrared images respectively, the final feature vector is then of size $m = m_v + m_i$. Distances are then computed between these vectors. Since we used the Mahalanobis distance for the PCA, distance between two final vectors is not just the sum of distances of visible and infrared (as it would be if we have used the L_1 distance for example).

A similar fusion is realized for the sparse approach. The two sparse vectors are concatenated, the size of the fused vector is then two times the size of the dictionary.

Results for the *Same–session* and the *Time–lapse* experiments are successively shown in Tab. II. The *Same–session* identification rates are slightly better than the already good results of the low–level fusion. Results of the *Time–lapse* experiment are improved over the low–level fusion results.

D. High Level of Fusion

In this part, we have applied the most popular fusion scheme, the decision level fusion. Features are extracted separately for the visible image and its infrared counterpart. Distances d_v and d_i (for the visible and infrared part respectively) are then computed between the features of the probe images and their corresponding galleries. The distances d_v and d_i are then merged and the identification decision is taken via the nearest neighbor classifier.

Before applying a merging technique to the distances d_v and d_i, it is necessary to normalize them. Distances to the gallery of a probe image can be seen as a vector of distances, so we simply normalize it between 0 and 1 : $d = d_v + d_i$.

Results for the *Same–session* and the *Time–lapse* experiments are successively shown in Tab. III. They are all improved over the mid–level fusion presented above. We can see that the PCA approach is less robust than the sparse approach to facial expression changes between the enrolment and the test.

IV. SUMMARY OF RESULTS

A summary of the results and a comparison to previous ones on the UND database with the same protocol are shown in Tab. IV. We can see that a decision level fusion leads to better identification rates in all cases. At this level, identification rates for the *Same–session* experiment are nearly perfect for all the approaches, it is mainly due to the quite ‘simplicity’ of this experiment, where gallery and probe images are taken within seconds. In [5] is published better identification rates than ours with the PCA approach. This is mainly due to the size of the images (bigger in [5]) and a more sophisticated transformation of the scores before their combination. The sparse approach presented here offers slightly better results than those in [3], and offer always the

TABLE III

RANK–0 RECOGNITION RATES FOR THE HIGH–LEVEL FUSION. TOP TABLE: *Same-session*, BOTTOM TABLE: *Time-lapse* EXPERIMENT. IN EACH CELL, TOP: PCA, BOTTOM: SPARSE APPROACH.

Gallery \ Probe	FA|LF	FA|LM	FB|LF	FB|LM
FA|LF		1.00 1.00	0.98 1.00	0.98 1.00
FA|LM	1.00 1.00		0.96 1.00	1.00 1.00
FB|LF	1.00 1.00	0.96 1.00		1.00 1.00
FB|LM	1.00 0.98	1.00 1.00	1.00 1.00	

Gallery \ Probe	FA|LF	FA|LM	FB|LF	FB|LM
FA|LF	0.92 0.98	0.92 0.97	0.75 0.94	0.76 0.94
FA|LM	0.91 0.99	0.91 0.98	0.77 0.96	0.79 0.94
FB|LF	0.81 0.96	0.78 0.94	0.87 0.98	0.87 0.97
FB|LM	0.86 0.98	0.86 0.95	0.86 0.96	0.86 0.97

TABLE IV

COMPARISON OF METHODS. TOP TABLE: *Same–session*, BOTTOM TABLE: *Time–lapse* EXPERIMENT. MEAN RECOGNITION RATE OVER THE 12 (OR 16 SUB–EXPERIMENTS) AND STANDARD DEVIATION IN PARENTHESIS. BEST SCORE PER LINE IN BOLD.

	This paper PCA	This paper Sparse	[5]	[3]
Low–level	0.95 (0.02)	**0.98** (0.01)	N/A	N/A
Mid–level	0.98 (0.01)	**0.99** (0.01)	N/A	N/A
High–level	**0.99** (0.01)	**0.99** (0.01)	N/A	**0.99** (0.01)
Low–level	0.63 (0.04)	**0.87** (0.04)	N/A	N/A
Mid–level	0.72 (0.06)	**0.95** (0.02)	N/A	N/A
High–level	0.84 (0.05)	**0.96** (0.01)	0.92 (0.02)	0.95 (0.02)

best score for the image and feature fusion levels. We think that the local normalization of each patch makes the system more robust to global changes (like illumination or thermal variations).

V. CONCLUSION

We presented a numerical study of different fusion levels of visible and infrared face images. The well-known eigenfaces method is compared to an approach based on the sparsity theory. As feature extractor, it decomposes a face onto a dictionary that has been learned from data, and processes the identification by considering this feature vector as a linear combination of corresponding gallery's feature vectors. Three levels of fusion have been considered : the low (image/sensor) level where images from the two modalities are merged, the middle (feature) level where features from both modalities are merged, and high (decision/score) level where scores from the two modalities are merged. Results on the *Notre-Dame* database show that a decision level fusion improves identification rates over image or feature level fusion, the sparse approach giving the best results in most cases. However, an image fusion scheme could be more interesting in case of eyeglasses for example. Moreover, it is not sure that the score fusion level would be the best choice if the illumination conditions are very bad. Nevertheless, our results for the first fusion levels are relevant and show the feasability of such approaches. The choice of the fusion scheme should then depend on the kind of extern conditions or the type of application. Further work has to be conducted on other merging technics at all levels.

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] P. Buyssens and M. Revenu. Ir and visible face identification via sparse representation. In *Biometrics: Theory, Applications and Systems*, 2010.

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

[7] Jingu Heo. Fusion of visual and thermal face recognition techniques : A comparative study. Master's thesis, 2003.

[8] 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.

[9] F. Scalzo, M. Nicolescu, L. Loss, and A. Tavakkoli. Feature fusion hierarchies for gender classification. In *International Conference on Pattern Recognition*, 2008.

[10] R. Senaratne, S. Halgamuge, and A. Hsu. Face recognition by extending elastic bunch graph matching with particle swarm optimization. In *Journal of Multimedia*, 2009.

[11] S. Singh, A. Gyaourova, G. Bebis, and I. Pavlidis. Infrared and visible image fusion for face recognition. In *SPIE Defense and Security Symposium*, 2004.

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

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

[14] 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.