Fusion of IR and Visible Light Modalities for Face Recognition

Pierre Buyssens, Marinette Revenu, Olivier Lepetit

Fusion of IR and Visible light Modalities for Face Recognition

Pierre Buyssens
Orange Labs
42 Rue des Coutures
Caen - France
pbuyssen@greyc.ensicaen.fr

Marinette Revenu
Laboratoire GREYC
CNRS UMR6072
Caen - France

Olivier Lepetit
Orange Labs
42 Rue des Coutures
Caen - France

Abstract— We present a low resolution face recognition technique based on a special type of convolutional neural network which is trained to extract facial features from face images and project them onto a low–dimensional space. The network is trained to reconstruct a reference image chosen beforehand, and it has been applied in visible and infrared light. Since the learning phase is achieved separately for the two modalities, the projections, and then the new spaces, are uncorrelated for the two networks. However, by normalizing the results of these two non–linear approaches, we can merge them according to a measure of saliency computed dynamically. We experimentally show that our approach obtain good results in terms of precision and robustness, especially on new and *unseen* subjects.

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, they can mainly be divided into two parts :

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

The local approaches first extract some features (like eyes, nose and mouth) by the use of special feature extractors. The recognition task is then performed using some measures (like the distance between the eyes) on these features.

The most popular local technique 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 [5] used geometric models like the distance between pairs of feature points to complete the face recognition. Wiskott *et al.*[17] 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 the local approaches is that the extractors have to be chosen by hand and can be sub–optimal. Moreover, it is difficult to deal with different scales and poses.

The global approaches perform a statistical projection of the images onto a face space. The most popular technique called *Eigenfaces* (first used by Turk and Pentland [16]) is based on a Principal Components Analysis (PCA) of the faces. It has also been applied to infrared faces by Chen *et al.* [7]. Jung *et al.* [10] 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 faces into classes according to the Fisher criterion. It has been applied early by Kriegman *et al.*[11].

A comparison of these methods is made by Socolinsky and Selinger in [15], or by Wu *et al.* in [18] which test also a Discrete Cosine Transform.

Other methods are specific to the infrared modality, like the work of Akhloufi *et al.* [4] where features are computed from extracted blood vessels.

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, the appearance of it undergoes a non–linear transformation, and due to the linear projection of the global approaches, the classification can fail.

Extensions of these linear approches have been proposed like kernel–PCA [13], or kernel–LDA [9] for face recognition. The drawback of these extensions is there is no invariance unless it is built into the kernel, and once again by hand. This is also the drawback of other machine learning technics like Support Vector Machine.

B. Our approach

We propose an approach that alleviates some of these problems by using a special type of Convolutional Neural Network (CNN). The network, called the *Face–Reconstruction Network* is based on the *diabolo* network model [14] where the output is the same vector as the input, with a less dimensional intermediate layer. The networtk then learns a compact code of the input. By applying some transformations to the input vector, while not changing the output, the network is then able to learn a compact code of the input

invariant to those transformations. Inspired by the work of Duffner and Garcia [8], the Face–Reconstruction Network acts like a diabolo network. It projects non–linearly the input onto a subspace and then tries to reconstruct a *reference* face that has been chosen beforehand. It can be seen as a kind of non–linear PCA, where a face is reconstructed using a set of reconstruction vectors. The Face–Reconstruction Network is used for the two modalities, visible an infrared.

This approach is based on convolutional neural networks architecture. CNN offers the advantage of learning how to extract the face features automatically, so no choice of a particular extractor or of a particular kernel has to be made. They are also designed to be more robust to illumination change, and pose variation.

The paper is organized as follow : The architecture is described in section II. The database, the preprocessing and the learning phase are detailed in section III. Sections IV, V and VI detail the three experiments we conducted. Section VII shows the importance of the gallery and the number of samples to enroll a subject. We present then our technique to merge the scores of the two modalities and its results in Section VIII. Finally, we present our conclusions and further work in section IX.

II. ARCHITECTURE OF THE NETWORK

The Face–Reconstruction Network (see Fig.1) takes in input an image of size 56×46 (i.e.: the size of the retina of the network) and passes it through a succession of convolution C_i, subsampling S_i and fully connected F_i layers. The output of the Network is an image, of the same size than the input, which is reconstructed by the last layer F_7. Each pixel of the output is one neuron, so there are $56 \times 46 = 2576$ neurons on the last layer.

We choosed a configuration of the first six layers, adapted to our problem, which is similar to the LeNet Network[12]:

- C_1. Feature maps : 15; Kernel size: 7×7; (Maps) Size: 50×40. Fully connected to the input.
- S_2. Feature maps: 15; Kernel size: 2×2; Size: 25×20.
- C_3. Feature maps: 45; Kernel size: 6×6; Size: 20×15. Partial connections to break symmetry.
- S_4. Feature maps: 45; Kernel size: 4×3; Size: 5×5.
- C_5. Feature maps: 250; Kernel size: 5×5; Size: 1×1. Fully connected to S_4.
- F_6. Neurons: 50; Fully connected to C_5.
- F_7. Neurons: 2576; Fully connected to F_6.

All the neurons use the sigmoid activation function, which is of the form : $\Phi(x) = 1.7159 \times tanh(\frac{2}{3}x)$.

Note that when testing the network, this is not the state of the last layer which is taken into account, but the compact code represented by the state of the penultimate layer (that to say 50 values). Some distances (L_1, L_2, Mahalanobis) have also been tested during the test, the *mahcosine* giving the best results has been retained for all the results presented here :

$$d(x,y) \;=\; -\frac{x \cdot y}{\|x\| \cdot \|y\|} \;=\; -\frac{\sum_{k=1}^{N_6} x_i y_i}{\sqrt{\sum_{k=1}^{N_6} (x_i)^2 \sum_{k=1}^{N_6} (y_i)^2}}$$

where N_6 is the number of neurons of layer F_6.

This architecture has already been tested on the visible ORL/AT&T database [2] which contains 10 images for each 40 subjects with variations of lighting and head positions. Tests on 50 images from *unseen* subjects (that has not been used for training) of this database are reported Table I. We can see that the face reconstruction network gets better results than the eigenfaces method (PCA), which validates the convolutional neural network approach for face recognition.

Rank	Face Reconstruction	PCA
0	38	29
1	45	33
2	45	38
3	47	40
4	47	42
5	49	44
6	50	44

TABLE I

Cumulative matches on *unseen* faces of the ORL/AT&T database. (*The last match for the Eigenfaces (PCA) is at rank* 23).

III. LEARNING PHASE

The Notre–Dame Database [1] (Collection $X1$) is used to train and test our networks. It has the advantage to present images of subjects with the two modalities, visible and infrared.

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 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*).

For a subject, a reference image has to be chosen beforehand for the training phase. It is in fact the face for the subject that the network has to recontruct.

The training is then performed using the descent gradient with the classical regression cost function:

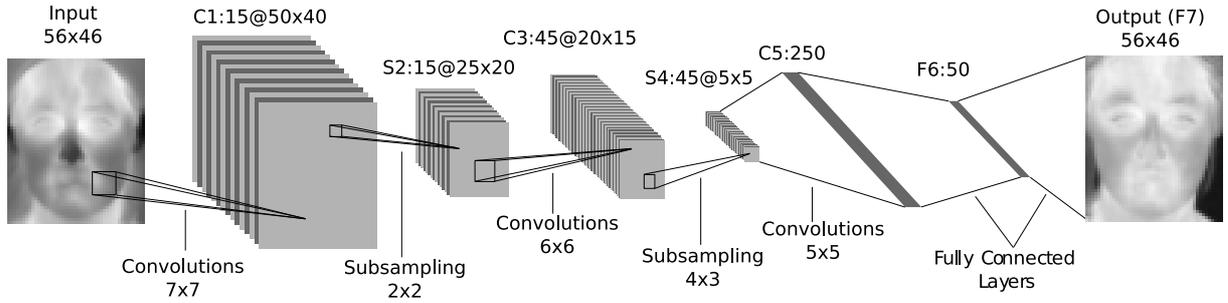$$E \;=\; \frac{1}{2}\|\boldsymbol{o_p} - \boldsymbol{t_p}\|^2$$

Fig. 1. Architecture of the Network

where o_p and t_p are the output values and the target values respectively for the pattern p.

For all the learnings, second order method is used to compute an approximation of the per-parameter optimal learning rate, in order to speed–up the learning process through the convergence of the network.

All the images have been resized to 56×46, their histogramm normalized and their pixel values scaled to ensure for each image $\mu \approx 0$ and $\sigma \approx 1$.

IV. FIRST EXPERIMENT

In a first experiment, we used the sets provided with the database which are explained in Sec. III. The first problem with the train set is that there is only one image per subject, so we created new images by applying some transformations to the original image, like a flip, a contrast enhancement or by adding some artificial lighting to parts of the image. We finally get $159 \times 12 = 1908$ images in the train set, which we divided into two parts: the first part composed of 159 images (one per subject chosen randomly) is used for the cross–validation, and the rest to train the network. For each subject, the reference image (i.e. the image to be reconstructed) is the original image. A cross–validation is performed during training after each iteration. It is useful to avoid the network overfitting the data, and then to improve its capacity of generalization.

The results for the two modalities of the two experiments are shown Fig.2, 3, 4, 5. Each curve has a name where the first part is the name of the gallery set, and the second part the name of the probe set. We can see that the results for both modalities are good for the *Same–session* experiment (Fig.2 and 4), but quite bad for the *Time–lapse* experiment (Fig.3 and 5) where the match rate at rank 0 is about 30%.

The main reason for the bad rates for the *Time–lapse* experiment is that we used only one image per subject during training. By applying some transformations (flip, contrast enhancement, blur ...) to the input, the network is able to learn them. But, some other variations (like facial expressions) are not taken into account (there are little or no facial expressions in the train set), so the network is not invariant to them, and since there are facial expressions both in gallery and probe sets, the recognition fails.
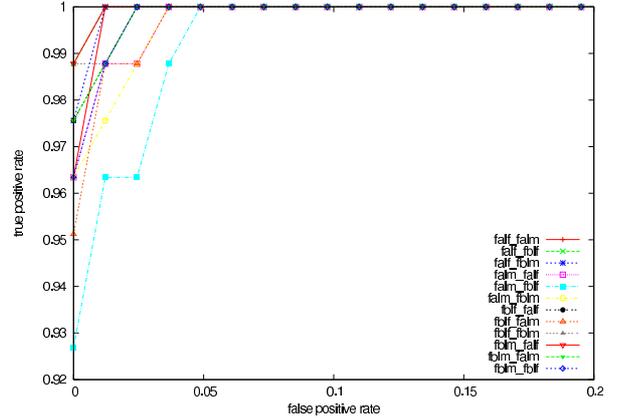

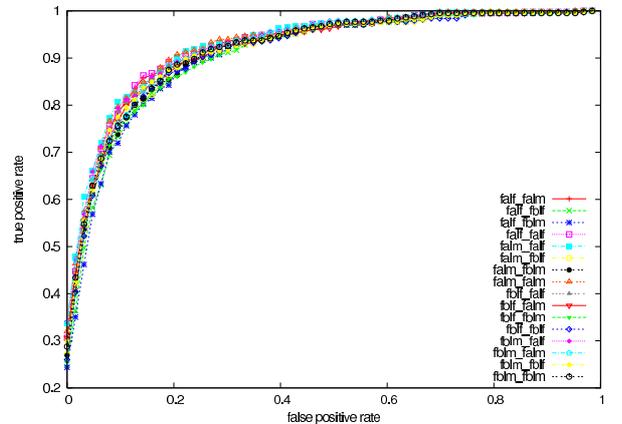Fig. 2. ROC curves for the *Same–session* experiment, Visible, first experiment

Fig. 3. ROC curves for the *Time–lapse* experiment, Visible, first experiment

V. SECOND EXPERIMENT

In this second experiment, we tried to bypass the lack of facial expressions present in the train set. We applied the same transformations to the 159 images of the train set, and we added a subset of the FERET database [3] composed of 2708 face images from 994 subjects. This subset of the FERET database presents some head rotations, facial expressions and lighting variations. The train set is finally composed of about 4608 images from 1153 subjects. From this, we remove 355 images from different subjects to make the validation set (as in Sec.IV).
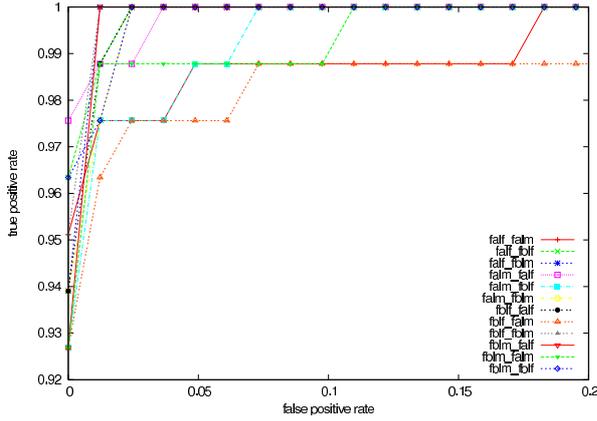
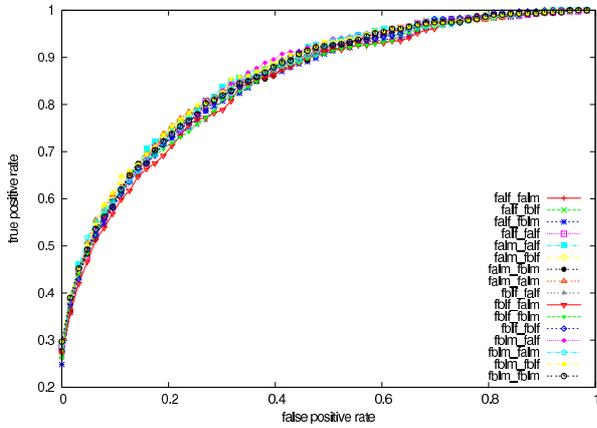Fig. 4. ROC curves for the *Same–session* experiment, IR, first experiment

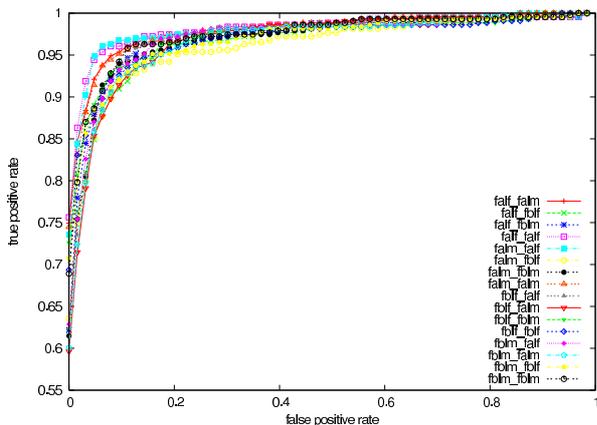Fig. 5. ROC curves for the *Time–lapse* experiment, IR, first experiment

Fig. 6. ROC curves for the *Time–lapse* experiment, Visible, second experiment

The results we obtained with this second experiment for the *Time–lapse* experiment in visible light are presented Fig. 6. The results for the *Same–session* experiment are quite the same than in the first experiment, so we do not display them here.

Compared to the first experiment, the results are better (the recognition rate at rank 0 is between 60% and 76%), which confirms the lack of expression variations of the previous train set.

The major problem with this second approach is that we can't do the same thing for the infrared modality due to the lack of images avalaible.

VI. THIRD EXPERIMENT

In order to increase the number of training images, and then the number of variations the network can learn, we decide to use some subjects of the test set (2292 images of 82 subjects) for the training phase. We split it randomly into two disjoint parts of 41 subjects each to form SET1 and SET2. SET1 and SET2 are then composed of 1256 and 1036 images respectively. Some variations have been applied to the original train set (composed of 159 subjects), and SET1 has been added to it. One image per subject has been retained to form the validation set, so we finally have a train set composed of $159 \times 11 + 1256 = 2964$ images of $159 + 41 = 200$ subjects, and a validation set composed of 200 images (of 200 different subjects).

The test sets have been changed, because we do not want to test the network on subjects that have been seen during the training phase. So the 41 subjects of SET1 have been removed from the probe lists (but not from the gallery lists). The tests consist then to match the images of 41 subjects (from SET2) against 82 subjects (SET1+SET2).

Table II shows that the results for the *Same–session* experiment are good, with visible modality outperforming infrared modality. However results for *Time–lapse* experiment (see Tbl. III) is worse than those obtained by Chen *et al.* [6]. The main reason to explain this is that our approach runs in low dimension (the size of the images is 56×46), while Chen *et al.* use a PCA in a higher dimension, so they are able to extract more relevant and precise informations (the eigenvectors of the PCA), and the classes are finally more separable.

Probe / Gallery	FALF	FALM	FBLF	FBLM
FALF		1.00 0.90	0.97 0.87	1.00 0.87
FALM	1.00 0.95		0.97 0.87	0.97 0.87
FBLF	0.95 0.97	0.95 0.87		1.00 0.97
FBLM	1.00 0.95	1.00 0.85	1.00 0.92	

TABLE II

Rank–0 recognition rates for the *Same–session* experiment, third approach. Top: Visible, bottom: IR

Probe / Gallery	FALF	FALM	FBLF	FBLM
FALF	0.80	0.76	0.68	0.67
	0.41	0.44	0.37	0.38
FALM	0.73	0.75	0.68	0.65
	0.42	0.38	0.34	0.38
FBLF	0.72	0.71	0.77	0.78
	0.44	0.37	0.46	0.42
FBLM	0.73	0.71	0.73	0.73
	0.43	0.34	0.41	0.42

TABLE III

Rank–0 recognition rates for the *Time–lapse* experiment, third approach.
Top: Visible, bottom: IR

VII. IMPORTANCE OF THE GALLERY SET

The relative bad rates of the *Time–lapse* experiment are due to the gallery sets. In our approach, the intra–classe variance may be higher than the inter–classe variance. In the *one image to enroll* scenario (like in the experiments above), if the image which will define the class of a subject is not well chosen, the class may not be clearly separable, and false positives occur.

To show this, we have conducted experiments, where one image per subject is used to enroll and the rest to test.

The weight vector of the third experiment (see Sec.VI) has been reused to compute the projection of the images. Then one image per subject from SET2 has been chosen randomly to form the gallery set, the rest forming the probe set. Because of the randomness of the choice of the images in the gallery set, the process has been iterated 1000 times and the mean of the recognition rates has been calculated. The final result is shown on Fig.7.
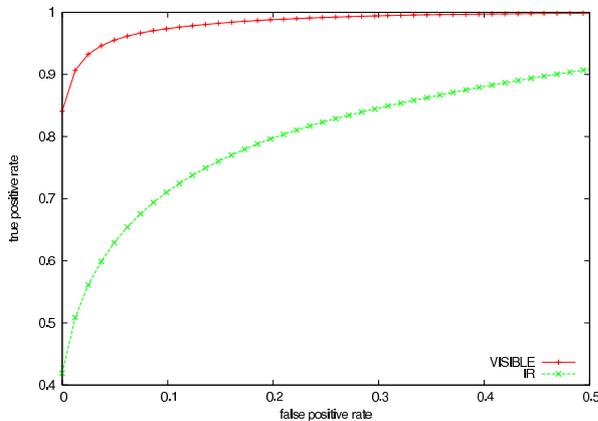

Fig. 7. Mean ROC curves of the random gallery experiment

Fig.7 shows us that recognition rate for visible light at rank 0 is about $84.\%$. It outperforms all the 16 *Time–lapse* tests made at experiment 3 (see III). For the infrared modality, the recognition rate for infrared light at rank 0 is about 41.9%. It is about the mean of the results of the 16 *Time–lapse* tests made at experiment 3 (see III). From this, we can conclude two things: first, the visible modality outperforms infrared modality in all cases, second, the gallery sets of the *Time–lapse* experiment for visible light offers less separability of

the classes than other gallery sets. The problem is then: what image will be the best to enroll one subject ? As we can have no *a priori* about the answer to this question, a possible way to tackle this problem is to enroll with more than one image.

We have then conducted similar experiment to the one explained above with more images to enroll (always chosen randomly), by simply averaging their projections. The process is iterated 1000 times. For some subjects who have few images, the maximum number of images avalaible has been used. More formaly:

$$n_{iep} = \min(\lambda, n_{ip}) - 1 \text{ for a subject who is tested}$$
$$n_{iep} = \min(\lambda, n_{ip}) \text{ for others}$$

where n_{iep} is the *n*umber of *i*mages used to *e*nroll subject *p*, λ is the desired number of images to enroll and n_{ip} the number of *i*mages avalaible for the subject *p*. The term -1 in the first case appears because we do not use the test image to enroll.

Modality	2–images enroll	3–images enroll	4–images enroll	10–images enroll
Visible	91.9	94.5	95.7	97.6
IR	55.4	61.6	65.4	72.9

TABLE IV

Rank–0 recognition rates according to the number of images used to enroll

As we can see on Tbl. IV, the rank–0 recognition increases with the number of images used to enroll a subject. The extremal case where all the images (except the one to test) are used gives a rank–0 recognition rate of 98.4% for visible light and 76.4% for infrared light. However, this extremal case is for example purpose and does not take into account the date of the images.

The reason of these results is by averaging the projections of multiple views, the signature of a subject is more stable to variations (facial expressions, lighting changes or head poses) and the classes become more separable. Moreover, in an operational scenario, the use of more than one image to compute a signature is not unrealistic, and an update of this signature through time can also be done easily.

VIII. FUSION

In this section we present the technique we proposed to merge the results obtained from the two modalities.

In order to enhance the recognition rate, we use the results on the two modalities and merge them according to a measure of saliency.

For a given test image I_v of a subject in visible modality, the distances of its projection to all the models m_k of the visible gallery are computed. We then have a distribution of distances. After a linear normalization of this distribution between 0 and 1, we compute its mean μ and its standard deviation σ. By fitting a gaussian curve on the distribution, we find a measure of saliency s_k for each of the distance d_k of the distribution :

$$s_k = \sigma\sqrt{2\pi}\frac{1}{e^{-\frac{1}{2}\left(\frac{d-\mu}{\sigma}\right)^2}}$$

The idea behind this is to give a big weight to a distance which is very different from the others (even if it is a big/bad distance), and inversely, to give a little weight to a distance which is quite the same as others, so we take the inverse of a gaussian.

This procedure gives us for the visible modality a distribution of distances d_{k_v}, each of it having a certain saliency s_{k_v}.

The same procedure is applied to the infrared modality with the infrared counterpart of I_v. It gives a distribution of distances d_{k_i}, each of it having a certain saliency s_{k_i}.

The final distances are obtained by computing a weighted sum of each couple of distances (d_{k_v}, d_{k_i}) according to their respective saliency (s_{k_v}, s_{k_i}):

$$d_k = \frac{d_{k_v} \times s_{k_v} + d_{k_i} \times s_{k_i}}{s_{k_v} + s_{k_i}} \ \forall k$$

Probe / Gallery	FALF	FALM	FBLF	FBLM
FALF		1.00 / 0.90 / 1.00	0.97 / 0.87 / 1.00	1.00 / 0.87 / 1.00
FALM	1.00 / 0.95 / 1.00		0.97 / 0.87 / 1.00	0.97 / 0.87 / 1.00
FBLF	0.95 / 0.97 / 1.00	0.95 / 0.87 / 1.00		1.00 / 0.97 / 1.00
FBLM	1.00 / 0.95 / 1.00	1.00 / 0.85 / 1.00	1.00 / 0.92 / 1.00	

TABLE V

Rank–0 recognition rates for the *Same–session* experiment, third experiment. Top: visible, middle: ir, bottom: fusion

Probe / Gallery	FALF	FALM	FBLF	FBLM
FALF	0.80 / 0.41 / 0.85	0.76 / 0.44 / 0.83	0.68 / 0.37 / 0.75	0.67 / 0.38 / 0.76
FALM	0.73 / 0.42 / 0.82	0.75 / 0.38 / 0.80	0.68 / 0.34 / 0.72	0.65 / 0.38 / 0.73
FBLF	0.72 / 0.44 / 0.82	0.71 / 0.37 / 0.80	0.77 / 0.46 / 0.80	0.78 / 0.42 / 0.88
FBLM	0.73 / 0.43 / 0.82	0.71 / 0.34 / 0.81	0.73 / 0.41 / 0.80	0.73 / 0.42 / 0.83

TABLE VI

Rank–0 recognition rates for the *Time–lapse* experiment, third experiment. Top: visible, middle: ir, bottom: fusion

Tables V and VI present the results we obtained for the *Same–session* and *Time–lapse* experiment respectively. The test sets are the same as in Sec.VI. Fusion of both modalities outperforms either modality alone in all cases, even when the scores of one modality are bad (like our scores on infrared modality).

IX. CONCLUSION AND FUTURE WORK

We presented a low resolution face recognition method for visible and infrared light imagery for telecom applications. Based on a special type of convolutional neural network, it receives a face image in input, and, for both modality, projects it onto a low–dimensional space where the recognition is performed. We successively show the importance of the training set for the training phase of the network, and the need of multiple samples to enroll in order to get a better classification. Results on the infrared modality give relative bad results, we think this is due to the high variablity of thermal distribution of the face over time. However, we present a fusion method of visible and infrared modalities which outperforms both. We are currently conducting experiments to correlate the projections of the two modalities to extend the possibilities of the multimodal face recognition.

References

[1] http://www.nd.edu/ cvrl/undbiometricsdatabase.html.
[2] www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html.
[3] www.itl.nist.gov/iad/humanid/feret/.
[4] M. Akhloufi and A. Bendada. Thermal faceprint: A new thermal face signature extraction for infrared face recognition. In *CRV*, pages 269–272, 2008.
[5] R. Brunelli and T. Poggio. Face recognition: Features versus templates. *IEEE PAMI*, 15(10):1042–1052, 1993.
[6] X. Chen, P. J. Flynn, and K. W. Bowyer. IR and visible light face recognition. *Computer Vision and Image Understanding*, 99(3):332–358, September 2005.
[7] Xin Chen, Patrick J. Flynn, and Kevin W. Bowyer. PCA-based face recognition in infrared imagery: Baseline and comparative studies. In *AMFG*, pages 127–134. IEEE Computer Society, 2003.
[8] S. Duffner and C. Garcia. Face recognition using non-linear image reconstruction. In *i-LIDS: Bag and vehicle detection challenge*, pages 459–464, 2007.
[9] Jian Huang, Pong Chi Yuen, Wensheng Chen, and Jian-Huang Lai. Choosing parameters of kernel subspace LDA for recognition of face images under pose and illumination variations. *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, 37(4):847–862, 2007.
[10] 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.
[11] D. J. Kriegman, J. P. Hespanha, and P. N. Belhumeur. Eigenfaces vs. fisherfaces: Recognition using class-specific linear projection. In *ECCV*, pages I:43–58, 1996.
[12] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. In *Intelligent Signal Processing*, pages 306–351. IEEE Press, 2001.
[13] Hichem Sahbi. Kernel PCA for similarity invariant shape recognition. *Neurocomputing*, 70(16-18):3034–3045, 2007.
[14] Holger Schwenk. The diabolo classifier. *Neural Computation*, 10(8):2175–2200, 1998.
[15] Diego A. Socolinsky and Andrea Selinger. Thermal face recognition in an operational scenario. In *CVPR (2)*, pages 1012–1019, 2004.
[16] M. A. Turk and A. P. Pentland. Face recognition using eigenfaces. In *Proceedings IEEE CVPR*, pages 586–590, Hawai, June 1992.
[17] L. Wiskott, J. M. Fellous, N. Kruger, and C. von der Malsburg. Face recognition by elastic bunch graph matching. *IEEE PAMI*, 19(7):775–779, July 1997.
[18] 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.