

Fusion des modalités Visible et Infrarouge pour la Reconnaissance Faciale

Pierre Buysens *

Marinette Revenu

GREYC – CNRS UMR6072
ENSICAEN, Université de Caen Basse Normandie
14000 Caen

pierre.buysens@greyc.ensicaen.fr

Résumé

Nous présentons une technique de reconnaissance de visages fonctionnant pour de faibles résolutions, basée sur un type particulier de Réseau de Neurones Convolutionnels. Celui-ci a été entraîné pour extraire des caractéristiques faciales d'images de visages et les projeter sur un espace de faible dimension à des fins de comparaison, et a été appliqué aux modalités visible et infrarouge. Étant donné que les phases d'apprentissage ont été réalisées séparément pour les deux modalités, les projections, et donc les nouveaux espaces ne sont pas corrélés. Cependant, en normalisant les résultats de ces deux approches non-linéaires, nous pouvons les fusionner selon une mesure de pertinence calculée dynamiquement. Nous montrons expérimentalement que notre approche obtient de bons résultats en terme de précision et de robustesse, notamment sur des individus nouveaux et inconnus (i.e. n'ayant pas été utilisés lors de l'apprentissage).

Mots Clef

Reconnaissance de visages, Réseaux de Neurones Convolutionnels, Modalités Visible et Infrarouge, Fusion.

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 obtains good results in terms of precision and robustness, especially on new and unseen subjects.

Keywords

Face Recognition, Convolutional Neural Networks, Visible and Infrared Modalities, Fusion.

*Ce travail a été effectué dans le cadre d'une thèse Cifre chez Orange Labs Caen

1 Introduction

La reconnaissance faciale de personnes est un sujet dont l'intérêt n'est plus à démontrer : biométrie, vidéo-surveillance, IHM avancées ou encore indexation d'images/vidéos. Cependant, elle se heurte encore à de nombreux problèmes, dont le plus caractéristique est lié aux changements d'éclairage. Une possibilité pour pallier ce problème est l'utilisation d'autres modalités, telles que l'infrarouge qui n'est pas sujet aux changements d'éclairage. L'infrarouge permet en outre à un système biométrique de fonctionner, même lorsque la capture dans le domaine visible est impossible ou de mauvaise qualité, lors d'une capture nocturne par exemple. Le capteur infrarouge utilisé ici fonctionne dans le domaine des faibles longueurs d'ondes, générant ainsi une cartographie thermique des visages.

1.1 État de l'art

De nombreuses approches ont été proposées dans la littérature pour la reconnaissance faciale [15], elles peuvent principalement être classées en deux groupes :

- les approches locales qui extraient des caractéristiques faciales, puis les combinent au sein d'un modèle plus global pour ensuite effectuer une classification.
- les approches globales qui réalisent souvent une forme de projection linéaire de l'espace de grande dimension (i.e. les images de visage) dans un espace de dimension plus faible.

Les approches locales extraient dans un premier temps des caractéristiques (comme les positions des yeux, du nez ou de la bouche) en utilisant des extracteurs dédiés. La tâche de reconnaissance proprement dite est ensuite réalisée en effectuant certaines mesures (comme la distance entre les yeux) sur ces caractéristiques.

L'approche locale la plus populaire est l'*Elastic Graph Matching* (EGM) où un ensemble de point d'intérêts est extrait du visage, à partir duquel un graphe est créé. Brunelli et Poggio [9] utilisent des modèles géométriques comme la distance entre des paires de points caractéristiques pour réaliser la reconnaissance faciale. Wiskott *et al.* [7] utilisent des filtres de Gabor sur le voisinage de ces points pour calculer un ensemble de *jets* pour créer la méthode dite de l'*Elastic Bunch Graph Matching* (EBGM). Ici, la forme du visage est

modélisée grâce à ces jets pour améliorer la reconnaissance. Le principal problème des approches locales est que l'extracteur de caractéristiques doit être choisi par le concepteur du système, et est souvent particulier à un contrainte spécifique et donc sous-optimal pour d'autres contraintes.

Les approches globales réalisent une projection statistique des images (espace de grande dimension) dans un espace de visages (généralement de plus faible dimension). La méthode la plus populaire appelée *Eigenfaces* (introduite par Turk et Pentland [19]) est basée sur l'Analyse en Composantes Principales (ACP) des visages. Elle a également été appliquée à la modalité infrarouge par Chen *et al.* [8]. Jung *et al.* [11] l'utilisent conjointement à une analyse de la surface des visages. Une autre technique populaire appelée *FisherFaces* est basée sur une Analyse Discriminante Linéaire (LDA), qui divise les visages en classes selon le critère de Fisher. Elle a été appliquée très tôt par Kriegman *et al.* [12].

Une comparaison de ces méthodes est effectuée par Socolinsky et Selinger dans [18], et par Wu *et al.* dans [20] qui testent également l'utilisation des Transformées en Cosinus Discrets (DCT).

D'autres méthodes sont spécifiques à la modalité infrarouge, comme les travaux de Akhloufi *et al.* [4] où des caractéristiques sont calculées à partir de l'extraction des vaisseaux sanguins.

Le principal inconvénient des approches globales est leur sensibilité aux changements de luminosité pour le visible, et aux changements dans le temps de la distribution de chaleur pour l'infrarouge. En effet, lorsque la luminosité (ou la distribution thermique) d'un visage change, son apparence subit une transformation non-linéaire, et étant donné l'aspect linéaire des approches globales, la classification peut échouer.

Des extensions de ces approches linéaires ont été proposées comme l'Analyse en Composantes Principales à Noyaux (Kernel-PCA) [16], ou l'Analyse Discriminante Linéaire à Noyaux (Kernel-LDA) [10] pour la reconnaissance faciale. L'inconvénient de ces extensions est qu'il n'y a pas d'invariance à certaines transformations à moins que celles-ci ne soient prises en compte lors de la création du noyau, et donc encore une fois manuellement. C'est également le défaut d'autres techniques d'apprentissage comme les Machines à Vecteurs de Support (SVM).

1.2 Notre Approche

Nous proposons une approche qui simplifie certains des problèmes énoncés ci-dessus en utilisant un type particulier de Réseau de Neurones Convolutionnels (CNN). Notre réseau, appelé *Réseau de Reconstruction* est basé sur le modèle dit *diabolo* [17] où la sortie souhaitée du réseau est identique à l'entrée, avec une couche intermédiaire de faible dimension. Le réseau apprend ainsi une représentation compacte de l'entrée. En appliquant certaines transformations au vecteur d'entrée sans changer la sortie désirée, le réseau est ainsi capable d'apprendre une représentation compacte in-

variante à ces transformations. Inspiré des travaux de Duffner et Garcia [6], le Réseau de Reconstruction fonctionne comme un réseau diabolo. Il projette de façon non-linéaire l'entrée sur un sous-espace puis reconstruit l'image d'un visage de *référence* choisi préalablement. Une instance différente du réseau de reconstruction est utilisée pour chacune des modalités.

Cette approche est basée sur les réseaux de neurones convolutionnels. Ces réseaux offrent l'avantage d'apprendre à extraire les caractéristiques faciales, ainsi aucun choix n'est effectué sur un extracteur ou un noyau particulier. Ils sont de plus conçus pour être invariants aux changements d'éclairage ainsi qu'aux variations de pose.

Le reste de l'article est organisé comme suit : l'architecture est décrite à la section 2. La base de données utilisée, les prétraitements et la phase d'apprentissage sont détaillés à la section 3. Les sections 4, 5 et 6 détaillent de façon chronologique les trois expériences que nous avons menées. La section 7 montre l'importance des images pour l'enrôlement ainsi que leur nombre. Nous présentons ensuite notre technique pour fusionner les scores des deux modalités et les résultats à la section 8. Finalement, nous concluons et présentons de futurs travaux à la section 9.

2 Architecture du réseau

Le réseau de reconstruction (voir Fig.1) prend en entrée une image de taille 56×46 (i.e. : la taille de la rétine du réseau) et la passe dans une succession de couches de convolution C_i , subsampling S_i et de neurones complètement connectés F_i de type *Multi-Layer Perceptron* (MLP). La sortie du réseau est une image, de même taille que l'entrée, qui est reconstruite par la dernière couche F_7 . Chaque pixel de la sortie est représenté par un neurone, il y a donc $56 \times 46 = 2576$ neurones sur la dernière couche.

Notre architecture, similaire au réseau *LeNet* de Y.Lecun [13], a été adaptée au problème. Plus précisément :

- *Input*. Nombre d'images : 1. Taille : 56×46 .
- C_1 . Nombre de cartes : 15; Taille des noyaux : 7×7 ; Taille des cartes : 50×40 . Toutes les cartes sont connectées à l'entrée.
- S_2 . Nombre de cartes : 15; Taille des noyaux : 2×2 ; Taille des cartes : 25×20 . Connexions 1 – 1.
- C_3 . Nombre de cartes : 45; Taille des noyaux : 6×6 ; Taille des cartes : 20×15 . Connexions partielles pour casser la symétrie.
- S_4 . Nombre de cartes : 45; Taille des noyaux : 4×3 ; Taille des cartes : 5×5 . Connexions 1 – 1.
- C_5 . Nombre de cartes : 250; Taille des noyaux : 5×5 ; Taille des cartes : 1×1 . Couche complètement connectée à S_4 .
- F_6 . Nombre de cartes : 50; Couche complètement connectée à C_5 .
- F_7 . Nombre de cartes : 2576; Couche complètement connectée à F_6 .

Tous les neurones utilisent une fonction d'activation de type sigmoïde, qui est de la forme : $\Phi(x) = 1.7159 \times \tanh(\frac{2}{3}x)$.

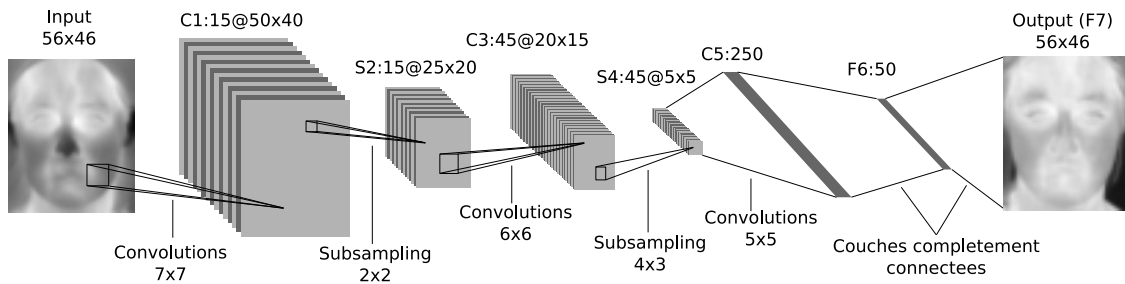


Fig. 1 – Architecture du réseau

Notons que lors des tests, ce n'est pas l'état de la dernière couche qui est prise en compte mais le code compact de l'avant-dernière (couche F_6 , soit un vecteur de dimension 50). Plusieurs distances (L_1 , L_2 , Mahalanobis) ont été testées durant la phase de test, la fonction de similarité cosinus dans l'espace de Mahalanobis, donnant les meilleurs résultats, a été retenue pour tous les résultats présentés ici :

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

où N_6 est le nombre de neurones de la couche F_6 .

La mise au point de cette architecture a été réalisée grâce à des tests sur la base visible ORL/AT&T [2] qui contient 10 images pour chacune des 40 personnes de la base. La base présente des variations de luminosité ainsi que de fortes variations de poses des visages. Les résultats de tests menés sur 50 images de personnes inconnues (n'ayant pas été utilisées lors de la phase d'apprentissage) sont présentés dans la table 1. Nous pouvons voir que le réseau de reconstruction donne de meilleurs résultats que la méthode des eigenfaces-ACP (testée dans les mêmes conditions), ce qui valide l'approche neuronale pour la reconnaissance faciale.

Rang	Réseau de Reconstruction	ACP
0	38	29
1	45	33
2	45	38
3	47	40
4	47	42
5	49	44
6	50	44

Tab. 1 – Correspondances cumulées pour des personnes inconnues de la base ORL/AT&T. (La dernière correspondance pour la méthode des eigenfaces (ACP) est au rang 23).

3 Méthodologie

Terminologie. La **phase d'apprentissage** pour un système neuronal consiste à trouver les poids du réseau par descente de gradient. Elle est réalisée grâce à une **base d'apprentissage**. Souvent, lors de l'apprentissage, une autre base (appelée **base de validation**) est utilisée pour réaliser

une validation croisée et permet ainsi d'éviter un surapprentissage de la base d'apprentissage.

Un système biométrique fonctionne en deux étapes principales : la phase dite d'**enrôlement** qui consiste en l'*enregistrement* des personnes via une base d'images (appelée **gallerie**), puis la phase d'identification qui consiste à faire correspondre un ensemble d'images de personnes (appelé **probe**) à la gallerie représentant les personnes enrôlées. Le **rang** pour une image testée en identification représente le nombre de faux positifs obtenus (i.e. plus il est bas, meilleure est la reconnaissance).

Données utilisées. La base Notre-Dame [1] (Collection X1) est utilisée pour l'apprentissage ainsi que pour les tests de l'approche. Celle-ci présente l'avantage d'avoir pour chaque image visible, l'image infrarouge correspondante.

La base est divisée en deux parties : la première partie, appelée *Ensemble d'apprentissage* (**TrS** dans la suite), est composée de 159 personnes, chacune ayant une seule image visible et son équivalent infrarouge. La deuxième partie, appelée *Ensemble de test* (**TeS** dans la suite), est composée de 82 personnes pour un total de 2292 images visibles et 2292 images infrarouges.

Alors que TrS ne contient ni expressions faciales, ni variations de la pose ou de luminosité, TeS contient de nombreuses images contenant des variations de luminosité, d'expressions faciales, de poses et de distribution thermique.

TeS est divisé en deux jeux de test, appelés *Same-session* et *Time-lapse* pour tester respectivement les problèmes de luminosité et la reconnaissance à travers le temps. Pour chacun de ces jeux de test, des fichiers livrés avec la base et nommés $f\{a,b\}l\{f,m\}$ correspondent à différentes gallerie et probes. Ces sous-ensembles ont été conçus pour pouvoir tester indépendamment les effets des expressions faciales (fa : neutral expression, fb : smiling expression), sous différentes conditions de luminosité (lf : *Feret style lighting*, lm : *mugshot lighting*).

Phase d'apprentissage. Pour chaque personne, une image de référence doit être choisie au préalable pour la phase d'apprentissage. Il s'agit de l'image cible que le réseau va essayer de reconstruire lors de l'apprentissage.

L'apprentissage est ensuite réalisé grâce à une descente de gradient en utilisant la fonction classique de coût :

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

où o_p and t_p sont respectivement les valeurs de sorties et les valeurs cibles pour une entrée p .

Pour tous les apprentissages, une méthode du second ordre a été mise en œuvre (voir [14]) pour calculer une approximation du taux d'apprentissage optimal de chaque paramètre, pour accélérer le processus d'apprentissage ainsi que pour améliorer la convergence du réseau.

Prétraitement. Toutes les images ont été redimensionnées à la taille 56×46 , leur histogramme normalisé, et leurs valeurs des pixels ramenés entre -1 et 1 , ceci pour assurer pour chaque image $\mu \approx 0$ et $\sigma \approx 1$.

4 Première approche

Dans un premier temps, nous avons utilisé les ensembles détaillés à la section 3. Le premier problème avec TrS est qu'il n'y a qu'une seule image par personne, nous avons donc créé de nouvelles images en appliquant des transformations aux images originales, telles des rotations, rehaussements de contraste, ou des ajouts de luminosité artificielle à certaines parties de l'image. Nous avons ainsi obtenu $159 \times 12 = 1908$ images que nous avons divisées en deux parties distinctes : la première partie composée de 159 images (une image par personne choisie aléatoirement) pour réaliser une validation croisée lors de l'apprentissage, et le reste pour l'apprentissage. Pour chaque personne, l'image de référence (i.e. l'image à reconstruire) est l'image originale. La validation croisée est réalisée après chaque itération lors de l'apprentissage pour éviter un surapprentissage, ce qui améliore ainsi la capacité de généralisation du réseau.

La moyenne et l'écart-type des différents jeux de tests pour les deux modalités des deux expériences (*Same-session* et *Time-lapse*) sont présentés aux figures 2, 3, 4, 5. Il s'agit de courbes ROC (Receiver Operating Characteristic) où l'ordonnée représente le taux de reconnaissance (*true positive rate*) et l'abscisse le taux de faux-acceptés (*false positive rate*). Chaque courbe a un nom où la première partie est le nom de la galerie, et la deuxième partie la probe. Nous pouvons voir que les résultats pour les deux modalités sont satisfaisants pour l'expérience *Same-session* (Fig.2 et 4), mais plutôt mauvais en ce qui concerne l'expérience *Time-lapse* (Fig.3 et 5) où le taux de reconnaissance au rang 0 est d'environ 30%.

La principale raison des mauvais scores pour l'expérience *Time-lapse* est qu'il n'y a qu'une seule image disponible par personne dans TrS. En appliquant certaines transformations (rotations, rehaussement de contraste ...) aux images présentées en entrée lors de l'apprentissage, le réseau est capable de les apprendre. Cependant, d'autres variations (comme les expressions faciales) ne sont pas prises en compte (il n'y a pas d'expressions faciales dans TrS), le réseau ne peut donc pas apprendre à y être invariant, et comme il y a des expressions faciales dans les galeries et les probes, la reconnaissance échoue.

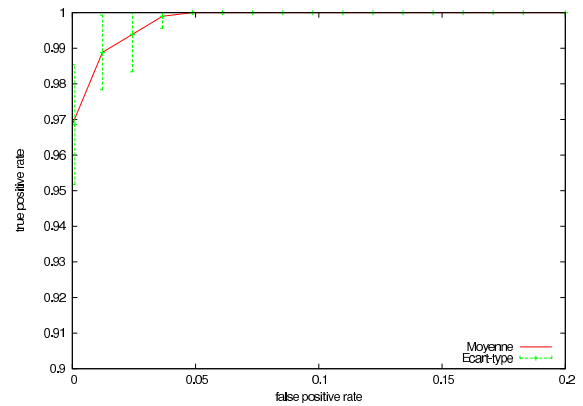


Fig. 2 – Courbe ROC pour l'expérience *Same-session*, Visible, première approche

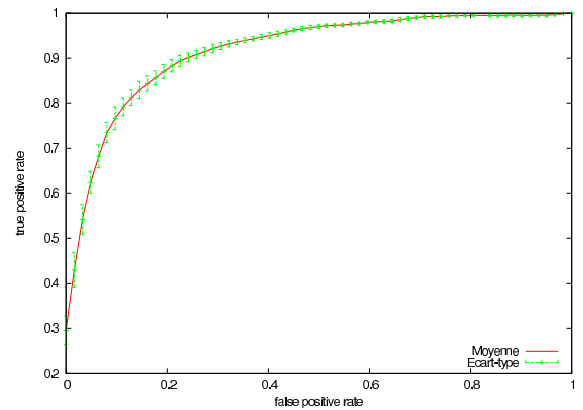


Fig. 3 – Courbe ROC pour l'expérience *Time-lapse*, Visible, première approche

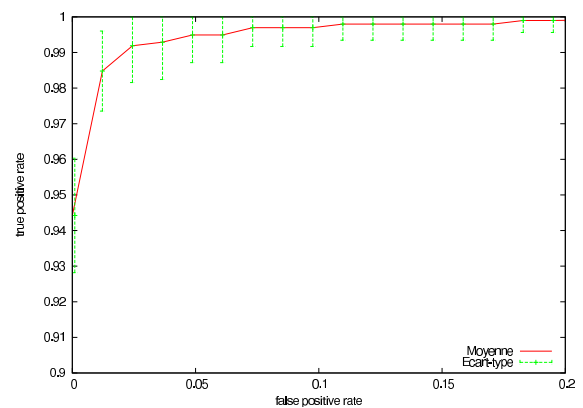


Fig. 4 – Courbe ROC pour l'expérience *Same-session*, IR, première approche

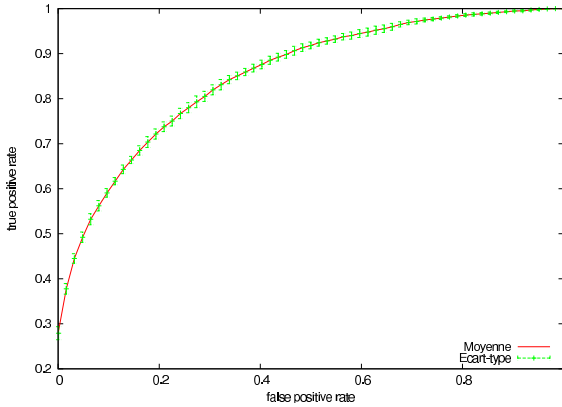


Fig. 5 – Courbe ROC pour l’expérience *Time-lapse*, IR, première approche

5 Deuxième approche

Dans cette deuxième approche, nous avons essayé de contourner le problème de la première approche, à savoir le manque d’images de visages présentant des expressions faciales dans l’ensemble d’apprentissage. Nous avons appliqué les mêmes transformations aux 159 images de TrS que lors la première expérimentation, et avons ajouté un sous ensemble de la base FERET [3] composé de 2708 images de visages de 994 personnes. Ce sous ensemble contient des variations de pose des visages, des variations d’éclairage ainsi que des expressions faciales. La base d’apprentissage est finalement composée de 4608 images de 1153 personnes. Nous en avons extrait 355 images de personnes différentes pour former l’ensemble nécessaire à la validation croisée (comme pour la première expérimentation, voir la section 4).

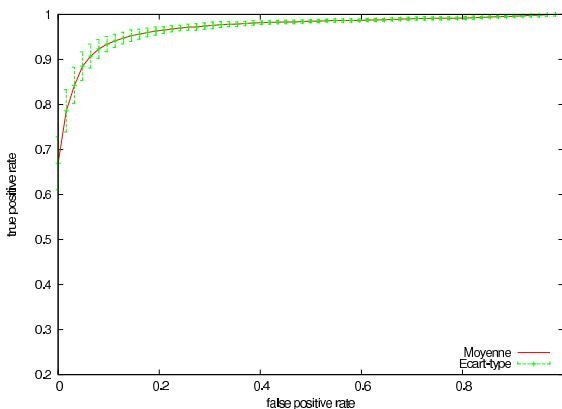


Fig. 6 – Courbe ROC pour l’expérience *Time-lapse*, Visible, deuxième approche

Les résultats obtenus pour l’expérience *Time-lapse* pour la modalité visible sont présentés à la figure 6. Les résultats pour l’expérience *Same-session* sont sensiblement identiques que lors de la première expérimentation, ils ne sont donc pas présentés ici.

Comparé à la première expérimentation, les résultats sont

meilleurs (le taux de reconnaissance au rang 0 est compris entre 60% et 76%, il était d’environ 30% pour la première approche), ce qui confirme le manque de variations (et notamment d’expressions) du précédent ensemble d’apprentissage. Le principal problème avec cette approche est qu’il nous est impossible de réaliser la même chose avec l’infrarouge étant donné le manque d’images disponibles pour cette modalité.

6 Troisième approche

En vue d’augmenter le nombre d’images de la base d’apprentissage, et donc le nombre de variations que le réseau peut apprendre, nous avons décidé d’utiliser des personnes de l’ensemble de test (TeS) pour la phase d’apprentissage. Nous avons divisé cet ensemble aléatoirement en deux parties disjointes de 41 personnes chacune pour former les ensembles SET1 et SET2 composés respectivement de $N1$ et $N2$ images chacun. Les mêmes variations que pour les deux approches précédentes ont été appliquées à TrS, et SET1 a été ajouté à cet ensemble d’apprentissage. Une image par personne a été retirée de cet ensemble pour former l’ensemble de validation, nous avons finalement une base d’apprentissage composée de $159 \times 11 + N1 = 2964$ images de $159 + 41 = 200$ personnes, et une base de validation composée de 200 images (de 200 personnes différentes).

Les probes ont été changés, en effet nous ne voulions pas tester le réseau avec des personnes qui avaient été utilisées lors de l’apprentissage. Ainsi, les 41 personnes de SET1 ont été enlevés des probes, mais pas des galleries. Ainsi, les tests consistent à tester les 41 personnes (de SET2) parmi les 82 de la base (SET1+SET2).

Le tableau 2 montre que les résultats obtenus pour l’expérience *Same-session* sont bons, la modalité visible ayant de meilleurs résultats que la modalité infrarouge. Cependant les résultats pour l’expérience *Time-lapse* sont moins bons (voir tab. 3) que ceux obtenus par Chen *et al.* [5]. La principale explication est que notre approche fonctionne dans de basses résolutions (les images sont de taille 56×46), tandis que Chen *et al.* utilisent une ACP avec des résolutions d’images plus grandes, ils sont donc capables d’extraire des informations plus pertinentes et précises (les vecteurs propres de l’ACP), et les classes sont finalement mieux séparables.

Galerie \ Probe		Probe			
		FALF	FALM	FBLF	FBLM
FALF	FALF		1.00	0.97	1.00
	FALM	0.90		0.87	0.87
FALM	FALM	1.00		0.97	0.97
	FBLF	0.95	0.95		1.00
FBLF	FBLF	0.97	0.87		0.97
	FBLM	1.00	1.00	1.00	
FBLM	FBLM	0.95	0.85	0.92	

Tab. 2 – Taux de reconnaissance au rang 0 pour l’expérience *Same-session*, troisième approche. Haut : Visible, bas : IR.

Gallerie \ Probe	FALF	FALM	FBLF	FBLM
	FALF	0.80 (0.86) 0.41 (0.62)	0.76 (0.85) 0.44 (0.61)	0.68 (0.66) 0.37 (0.55)
FALM	0.73 (0.88) 0.42 (0.56)	0.75 (0.59) 0.38 (0.58)	0.68 (0.66) 0.34 (0.51)	0.65 (0.68) 0.38 (0.51)
FBLF	0.72 (0.76) 0.44 (0.55)	0.71 (0.74) 0.37 (0.55)	0.77 (0.79) 0.46 (0.56)	0.78 (0.79) 0.42 (0.58)
FBLM	0.73 (0.76) 0.43 (0.53)	0.71 (0.76) 0.34 (0.53)	0.73 (0.82) 0.41 (0.57)	0.73 (0.82) 0.42 (0.58)

Tab. 3 – Taux de reconnaissance au rang 0 pour l’expérience *Time-lapse*, troisième approche. Haut : Visible, bas : IR. Entre parenthèses : les résultats obtenus par Chen *et al.* [5].

7 Importance de l’enrôlement

Les taux relativement mauvais obtenus pour l’expérience *Time-lapse* sont dus aux gallerie. Dans notre approche, la variance intra classe peut être supérieure à la variance inter classe. Dans un scénario où une seule image est utilisée pour l’enrôlement (comme dans les expériences ci dessus), si l’image utilisée pour l’enrôlement n’est pas bien choisie, les classes peuvent ne pas être clairement séparables, et des faux positifs peuvent apparaître.

Pour montrer cela, nous avons mené des expériences où une image par personne est utilisée pour l’enrôlement et le reste pour tester.

Le vecteur de poids du réseau de la troisième expérimentation (voir section 6) a été réutilisé pour calculer les vecteurs projetés des images. Puis une image par personne de SET2 a été choisie aléatoirement pour former la gallerie, le reste formant la probe. Etant donné le caractère aléatoire de la gallerie, le processus a été itéré 1000 fois et la moyenne du taux de reconnaissance a été calculée. Le résultat final est présenté à la figure 7.

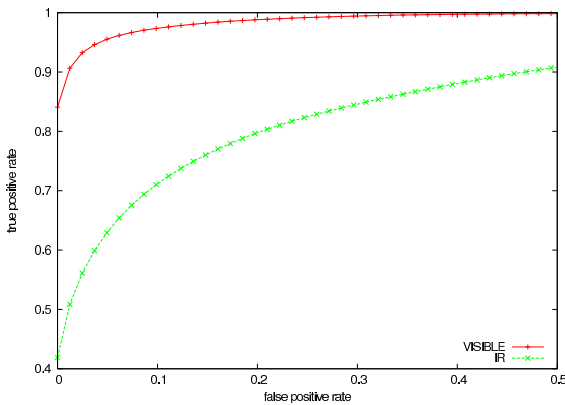


Fig. 7 – Courbe ROC moyenne avec les images d’enrôlement prises aléatoirement

Nous pouvons voir que le taux de reconnaissance moyen au rang 0 pour la modalité visible est d’environ 84.%. Cela surpasse les 16 tests de l’expérience *Time-lapse* réalisés lors de l’approche 3 (voir le tableau 3). Pour la modalité infrarouge, le taux de reconnaissance moyen au rang 0 est d’environ

41.9%. Cela correspond à environ la moyenne des taux de reconnaissance obtenus lors des 16 tests *Time-lapse* de l’approche 3 (voir le tableau 3). De ceci, nous pouvons tirer deux choses : premièrement, la modalité visible surpasse l’infrarouge dans tous les cas, et deuxièmement les gallerie des expériences *Time-lapse* séparent moins bien les classes que d’autres gallerie. Le problème est donc de trouver quelle image est la meilleure pour l’enrôlement d’une personne. Comme nous ne pouvons avoir d’*a priori* sur ce problème, une possibilité de contourner le problème est d’enrôler avec plus d’une image.

Nous avons ainsi mené une expérience similaire à celle décrite plus haut, mais cette fois ci en permettant l’enrôlement avec plusieurs images. Le vecteur caractéristique d’une personne étant alors simplement la moyenne des projetés de ses images d’enrôlement. Le processus a été itéré 1000 fois et la moyenne du taux de reconnaissance est calculée. Pour certaines personnes ne disposant pas d’assez d’images, le maximum d’images disponibles a été pris en compte. Plus formellement :

$$n_{iep} = \min(\lambda, n_{ip}) - 1 \text{ pour une personne qui est testée}$$

$$n_{iep} = \min(\lambda, n_{ip}) \text{ pour les autres}$$

où n_{iep} est le nombre d’images utilisées pour enrôler une personne p , λ est le nombre d’images désiré pour l’enrôlement et n_{ip} le nombre d’images disponibles pour la personne p . Le terme -1 dans le premier cas apparaît car nous ne voulons pas tester une image ayant été utilisée pour l’enrôlement.

Modalité	2 images	3 images	4 images	10 images
Visible	91.9	94.5	95.7	97.6
IR	55.4	61.6	65.4	72.9

Tab. 4 – Taux de reconnaissance au rang 0 selon le nombre d’images utilisées pour l’enrôlement

Comme nous pouvons le voir dans le tableau 4, le taux de reconnaissance au rang 0 augmente avec le nombre d’images utilisées pour l’enrôlement. Le cas extrême où toutes les images disponibles (sauf celle testée) d’une personne sont

utilisées donne un taux de reconnaissance de respectivement 98.4% et 76.4% pour les modalités visible et infrarouge. Cependant, ce cas extrême n'est pas réaliste puisqu'il ne prend pas en compte les dates des clichés.

L'explication de ces résultats est qu'en moyennant les projections de différentes vues d'une même personne, le vecteur signature d'une personne est plus stable aux variations (expressions faciales, luminosité, poses) et les classes deviennent plus facilement séparables. De plus, pour une utilisation opérationnelle, l'utilisation de plusieurs images pour l'enrôlement n'est pas irréaliste, et une mise à jour des vecteurs signatures peut être réalisée facilement au cours du temps.

8 Notre méthode de Fusion

Nous présentons ici la technique que nous proposons pour fusionner les résultats des deux modalités.

Pour améliorer les taux de reconnaissance, nous utilisons les résultats des deux modalités et les fusionnons selon une mesure de pertinence.

Pour une image test I_v d'une personne pour la modalité visible, les distances entre sa projection et tous les modèles m_k de la base visible sont calculées. Nous obtenons donc une distribution de distances. Après avoir normalisé cette distribution linéairement entre 0 et 1, sa moyenne μ et son écart type σ sont calculés. Nous pouvons ainsi calculer pour chaque distance d_k une mesure de pertinence associée s_k selon la fonction :

$$s_k = \left(1 + \frac{1}{2} \tanh \left(\frac{1}{\sigma} (d_k - \mu) \right) \right)^{-1}$$

L'idée ici est de donner beaucoup de poids à une distance qui est vraiment différente des autres pour autant qu'elle soit faible, et de donner un poids faible aux distances dès que le réseau n'arrive plus à bien séparer les modèles de la base.

Cette procédure nous donne ainsi une distribution de distances d_{k_v} , chacune d'elles ayant une certaine pertinence s_{k_v} .

La même procédure est appliquée pour la modalité infrarouge avec l'image infrarouge correspondante à I_v . Nous obtenons ainsi une deuxième distribution de distances d_{k_i} , chacune d'elles ayant une certaine pertinence s_{k_i} .

La figure 8 montre un exemple de calcul des pertinences des distances à un modèle de la base pour la modalité visible (à gauche) et infrarouge (à droite). Sur cet exemple, bien que la distance au modèle considéré soit plus grande pour la modalité visible que pour la modalité infrarouge, la pertinence de la distance pour l'infrarouge a un poids plus faible et est donc moins prise en compte pour le calcul de la distance finale au modèle.

Les distances finales sont obtenues en calculant la moyenne pondérée de chaque couple de distances (d_{k_v}, d_{k_i}) selon leur pertinences respectives (s_{k_v}, s_{k_i}) :

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

Gallery \ Probe	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	

Tab. 5 – Taux de reconnaissance au rang 0 pour l'expérience *Same-session*, troisième approche. Haut : Visible, milieu : IR, bas : Fusion

Gallery \ Probe	FALF	FALM	FBLF	FBLM
FALF		0.76 0.44 0.83	0.68 0.37 0.77	0.67 0.38 0.77
FALM	0.73 0.42 0.82		0.68 0.34 0.71	0.65 0.38 0.74
FBLF	0.72 0.44 0.83	0.71 0.37 0.82		0.78 0.42 0.89
FBLM	0.73 0.43 0.85	0.71 0.34 0.81	0.73 0.41 0.81	

Tab. 6 – Taux de reconnaissance au rang 0 pour l'expérience *Time-lapse*, troisième approche. Haut : Visible, milieu : IR, bas : Fusion

Les tableaux 5 et 6 présentent les résultats obtenus respectivement pour les expériences *Same-session* et *Time-lapse*. Les jeux de tests sont les mêmes que lors de l'expérimentation 3 (voir section 6). La fusion des deux modalités surclasse chacune des deux modalités prises seules, même lorsque les scores pour une modalité ne sont pas bons (comme les nôtres pour la modalité infrarouge).

9 Conclusion et Perspectives

Nous présentons une méthode de reconnaissance de visages fonctionnant dans de basses résolutions pour les modalités visible et infrarouge. Le réseau de neurones convolutionnels reçoit une image de visage en entrée, et pour chaque modalité la projette dans un espace de faible dimension où la reconnaissance est réalisée. Nous montrons successivement l'importance de la base d'apprentissage pour rendre le réseau invariant aux transformations, et l'apport obtenu lorsque plusieurs images sont utilisées lors de l'enrôlement. Les résultats obtenus pour la modalité infrarouge ne sont pas bons, nous pensons qu'ils sont dus à la trop grande variation de la distribution de chaleur pour une même per-

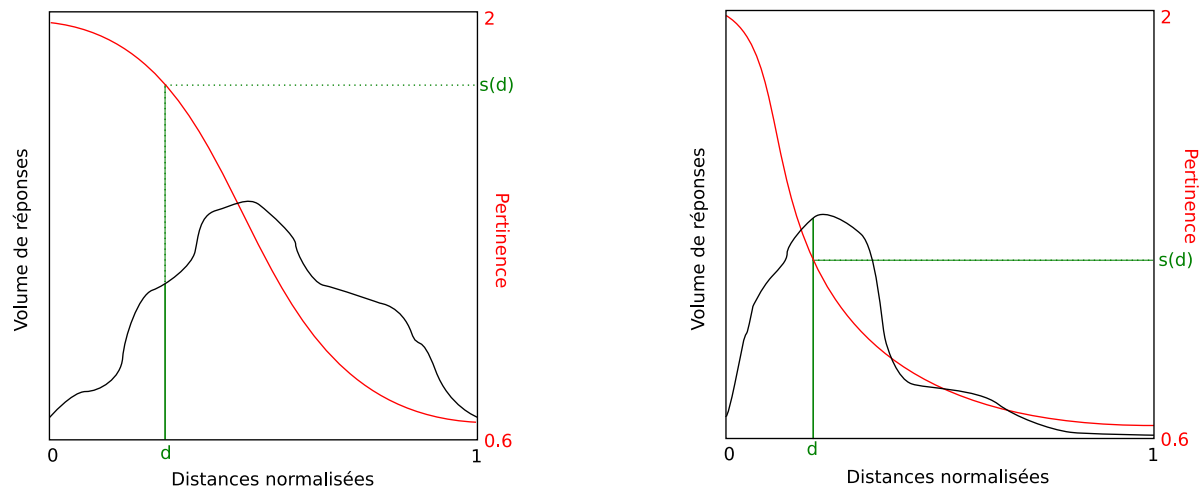


Fig. 8 – Calcul des pertinences. Gauche : Visible, Droite : Infrarouge

sonne au cours du temps. Cependant, nous présentons une méthode pour fusionner les scores provenant du visible et de l'infrarouge, la fusion surclassant chacune des deux modalités prise seule. Nous menons actuellement des expériences pour corréler les projections des deux modalités pour étendre les possibilités de la reconnaissance (par exemple, enrôler les personnes avec la modalité infrarouge et reconnaître avec le modalité visible).

Références

- [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] X. Chen, P. J. Flynn, and K. W. Bowyer. IR and visible light face recognition. *CVIU*, 99(3) :332–358, September 2005.
- [6] S. Duffner and C. Garcia. Face recognition using non-linear image reconstruction. In *i-LIDS : Bag and vehicle detection challenge*, pages 459–464, 2007.
- [7] L. Wiskott et J. M. Fellous et N. Kruger et C. von der Malsburg. Face recognition by elastic bunch graph matching. *PAMI*, 19(7) :775–779, July 1997.
- [8] X. Chen et P.J. Flynn et K.W. Bowyer. PCA-based face recognition in infrared imagery : Baseline and comparative studies. In *AMFG*, pages 127–134. IEEE Computer Society, 2003.
- [9] R. Brunelli et T. Poggio. Face recognition : Features versus templates. *PAMI*, 15(10) :1042–1052, 1993.
- [10] 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. *TSMC, Part B*, 37(4) :847–862, 2007.
- [11] S-W. Jung, Y. Kim, A. Jin Tech, and K-A. Toh. Robust identity verification based on infrared face images. In *ICCIT*, 2007.
- [12] 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.
- [13] 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.
- [14] Y. LeCun, L. Bottou, G. Orr, and K. Muller. Efficient backprop. In G. Orr and Muller K., editors, *Neural Networks : Tricks of the trade*. Springer, 1998.
- [15] D. Petrovska-DelaCrétaz, G. Chollet, and B. Dorizzi, editors. *Biometric Reference Systems and Performance Evaluation*. Springer, 2009.
- [16] H. Sahbi. Kernel PCA for similarity invariant shape recognition. *Neurocomputing*, 70(16-18) :3034–3045, 2007.
- [17] H. Schwenk. The diabolo classifier. *Neural Computation*, 10(8) :2175–2200, 1998.
- [18] D.A. Socolinsky and A. Selinger. Thermal face recognition in an operational scenario. In *CVPR*, pages 1012–1019, 2004.
- [19] M. A. Turk and A. P. Pentland. Face recognition using eigenfaces. In *CVPR*, pages 586–590, Hawaii, June 1992.
- [20] 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.