Réseau de Neurones Convolutionnels pour la Reconnaissance Faciale Infrarouge

Pierre Buyssens[1, 2], Marinette Revenu[1], Olivier Lepetit[2]

[1]Laboratoire GREYC - CNRS UMR6072 - Caen - France

[2]Orange Labs - 42 Rue des Coutures - Caen - France

pierre.buyssens@orange-ftgroup.com,
marinette.revenu@greyc.ensicaen.fr,
olivier.lepetit@orange-ftgroup.com

Résumé – Nous présentons une technique de reconnaissance faciale pour la modalité infrarouge basée sur un type particulier de réseaux de neurones convolutionnels. Le réseau a été entraîné pour extraire des caractéristiques faciales d'images de visages infrarouges et les projeter sur un espace de faible dimension à des fins de comparaison. Nous montrons expérimentalement que notre approche obtient de bons résultats sur des individus nouveaux et *inconnus* (i.e. n'ayant pas été utilisés lors de l'apprentissage).

Abstract – We present a face recognition technique in infrared modality based on a special type of convolutional neural network. The network was trained to extract facial features from infrared faces automatically and to project them in a new low–dimensional space. The faces are then compared in this new space. We show experimentally that our approach obtains good results on new and *unseen* subjects (i.e. subjects that have not been seen during the training phase.)

1 Introduction

La reconnaissance des personnes à l'aide de données biométriques est un sujet dont l'intérêt n'est plus à démontrer : biométrie, vidéo–surveillance, IHM avancées ou encore indexation d'images/vidéos. De nombreux types de données biométriques font l'objet d'études, parmi elles la reconnaissance faciale présente de nombreux intérêts : méthode de reconnaissance universelle, sans contact, voire même à l'insu de la population. Cependant, elle se heurte encore à de nombreux problèmes, dont le plus caractéristique est lié aux changements d'illumination. Une possibilité pour pallier ce problème est l'utilisation d'autres modalités, telles que l'infrarouge qui n'est pas sujet aux changements d'illumination. 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, rendant ainsi une cartographie thermique des visages.

1.1 Approches existantes

La plupart des travaux réalisés dans le domaine de la reconnaissance faciale infrarouge reprennent les approches linéaires classiques utilisées dans le domaine visible. Ainsi l'Analyse en Composantes Principales (ACP) est utilisée par Flynn *et al.* dans [3], ou par Jung *et al.* dans [7] conjointement avec une analyse des contours du visage. Une autre technique classique est l'Analyse Discriminante Linéaire (LDA) où l'espace des visages est divisé en classes selon le critère de Fisher (i.e. minimiser la variance intra–classe et maximiser la variance inter–classe). Elle a par exemple été utilisée par Socolinsky *et al.* dans [10]. Une Analyse en Cosinus Discrets est également testée par Wu *et al.* dans [11]. Huang *et al.* [5] proposent une méthode utilisant l'algorithme Adaboost pour la sélection d'ondelettes de Gabor pertinentes, une LDA les classifiant ensuite. D'autres méthodes, en cours de réalisation, sont exclusives à la modalité infrarouge, comme les travaux publiés par Akhloufi *et al.* dans [2] où des caractéristiques physiologiques sont extraites du visage infrarouge via les réseaux veineux sous la surface de la peau.

Un des principaux problèmes de ces approches est qu'elles se basent souvent sur des approches linéaires (cf. ACP, LDA). Cependant, comme le fait la luminosité dans le domaine visible, la température modifie la cartographie thermique des visages de façon non linéaire, ce qui rend ces méthodes non robustes.

Des extensions de ces approches linéaires ont été proposées comme l'Analyse en Composantes Principales à noyaux (kernel–PCA) [9], ou l'Analyse Discriminante Linéaire à noyaux (kernel–LDA) [6] pour la reconnaissance

faciale. Le défaut de ces méthodes est qu'elles n'offrent pas d'invariance à certaines transformations sauf si celles–ci sont prises en compte dans le noyau, ce qui induit une construction du noyau *à la main*. C'est également le défaut d'autres méthodes d'apprentissage comme les Machines à Support de Vecteurs (SVM).

1.2 Notre approche

Nous proposons une approche pour la reconnaissance faciale dans le domaine infrarouge basée sur un type de réseau de neurones convolutionnels, approche que nous avons déjà appliquée dans le domaine visible.

Le réseau de neurones convolutionnels utilisé ici, appelé *réseau de reconstruction*, réalise une projection non linéaire du visage présenté en entrée sur un sous–espace puis reconstruit un visage de référence choisi au préalable. Cette approche, inspirée des travaux de Duffner et Garcia [4] peut être vue comme une ACP non linéaire, dans le sens où un visage est reconstruit grâce à un ensemble de vecteurs de reconstruction.

Les réseaux de neurones convolutionnels offrent l'avantage d'apprendre automatiquement l'extraction des caractéristiques dans les premières couches, puis de les classifier dans les dernières. Ainsi, aucun choix d'algorithmes d'extraction n'est effectué. Ils sont de plus conçus pour être plus robustes aux changements de pose ainsi qu'aux changements thermiques des zones d'un visage liés aux conditions extérieures (comme la température), étant données leurs fonctions d'activation non linéaires.

2 Architecture du réseau

Le réseau de reconstruction prend une image de taille 56×46 en entrée et la passe à travers une succession de couches de convolution C_i, de sous–échantillonnage S_i et de neurones complètement connectés F_i. Chaque couche calcule son opération spécifique (convolution, sous–échantillonnage) avec son vecteur de poids, ajoute un biais puis passe le résultat dans une fonction d'activation Φ de la forme :

$$\Phi(x) = 1.7159 \times tanh(\frac{2}{3}x) \qquad (1)$$

Cette fonction présente les particularités suivaintes : $\Phi(1) = 1$, $\Phi(-1) = -1$ et sa dérivée seconde atteint un maximum en 1 et un minimum en -1, ce qui assure une meilleure convergence lors de la phase d'apprentissage. La sortie du réseau est une image, de même taille que l'entrée, qui est reconstruite par la dernière couche F_7. L'architecture interne du réseau (voir Fig.1), a été inspirée par les travaux de Lecun *et al.* [8], et adaptée à notre problème.

Plus précisément :

- 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 : 100; Couche complètement connectée à C_5.
- F_7. Nombre de cartes : 2576; Couche complètement connectée à F_6.

Notons que lors des tests, ce n'est pas l'état de la dernière couche qui est prise en compte mais l'avant–dernière (couche F_6, soit un vecteur de dimension 100).

Cette architecture a déjà été testée sur la base visible de visages ORL/AT&T [1] qui contient 10 images pour chacune des 40 personnes de la base. Cette base de données contient des variations de luminosités et de poses (i.e. rotations hors plan). Des tests sur 50 images de personnes inconnues (n'ayant pas été utilisées lors de l'apprentissage) sont présentés au tableau 1.

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

Table 1: Correspondances cumulées sur des visages *inconnus* de la base ORL/AT&T

Le tableau 1 présente les résultats obtenus par le réseau pour les 50 images de tests de personnes inconnues fonctionnant en identification. Pour une image test I_p de la personne p, sa projection G_I (le vecteur calculé à la couche F_6) est comparé à chacun des 40 modèles de la base. Le modèle pour une personne q est calculé comme étant le vecteur moyen des projections des images de la personne q. Cependant, pour le calcul du modèle d'une personne dont une image I_p est testée, l'image I_p n'est pas prise en compte. Il est ainsi nécessaire lors de chaque test d'une image I_p de recalculer le modèle de la personne p. La norme L_1 est ensuite utilisée pour calculer les distances du projeté G_I de l'image test I_p aux modèles. Les distances sont

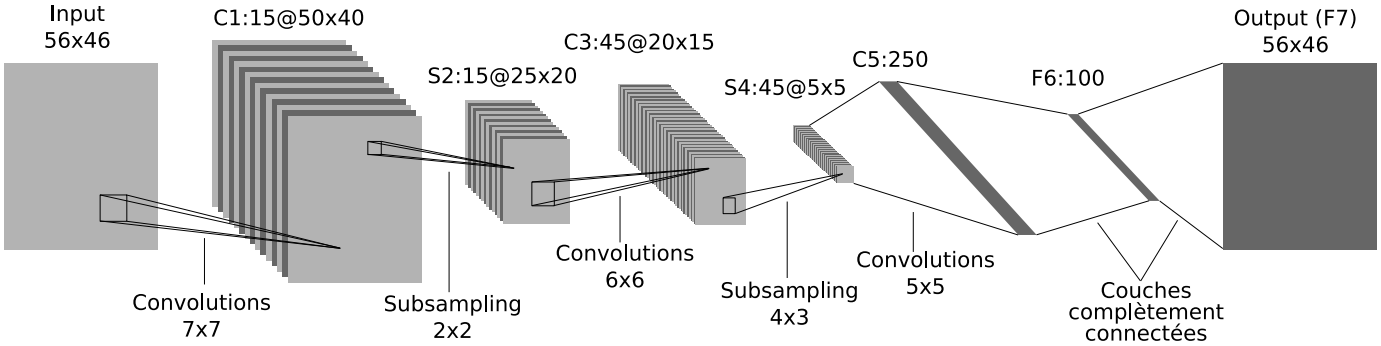

Figure 1: Architecture du réseau de reconstruction

ensuite rangées par ordre croissant, et la distance correspondant au modèle p détermine le rang auquel l'image I_p est trouvée. Si il s'agit de la plus faible, le rang vaut 0, si il s'agit de la deuxième plus faible alors le rang vaut 1 etc.

Le tableau 1 montre que la méthode utilisant un réseau de reconstruction de visages surpasse la méthode classique des visages propres (ACP). La dernière correspondance pour la méthode des visages propres (ACP) s'effectue au rang 23.

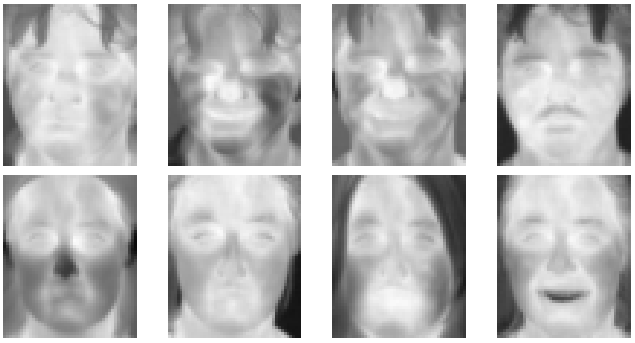

Table 2: Echantillons d'images de deux personnes différentes (une par ligne) de la base de données.

3 Apprentissage

Dans notre étude sur la modalité infrarouge, la base de référence Notre–Dame [3] est utilisée pour l'apprentissage ainsi que pour les tests. Celle–ci présente des variations de poses, d'expressions faciales ainsi que parfois de fortes variations de la distribution thermique des visages, certaines personnes de la base ayant en effet été capturées plusieurs fois avec un laps de temps entre les captures parfois important (voir le tableau 2 pour un échantillon de visages de la base). Nous nous sommes limités à un sous–ensemble de celle–ci contenant 870 images infrarouges de 26 personnes différentes, avec une grande variation du nombre d'images disponibles par personne, allant de 4 à 40. Les images présentent de plus des variations de pose ainsi que des changements de chaleur pour certaines parties du visage (typiquement les oreilles ou le nez).

Les images ont été centrées manuellement par rapport aux yeux de sorte que tous les visages aient leurs yeux approximativement au même endroit dans l'image. Elles ont été réduites à une taille de 56×46 (la taille de la rétine du réseau). Les valeurs des images ont été normalisées de sorte que leur moyenne soit approximativement de 0 et leur variance de 1, ceci pour assurer une meilleure convergence lors de l'apprentissage.

3.1 Partitionnement de la base

La base de données a été divisée en deux parties disjointes SET1 et SET2. SET1 contient les images de 20 personnes (totalisant 736 images), tandis que SET2 contient les images restantes (134 images de 6 personnes).

L'apprentissage a été réalisé avec SET1, et SET2 a été utilisé pour les tests. En partitionnant la base de la sorte, on peut ainsi tester la capacité de généralisation du réseau à des individus n'ayant pas été utilisés lors de l'apprentissage.

3.2 Phase d'apprentissage

Pour chaque individu i de SET1, une image de référence r_i a été choisie pour l'apprentissage. Elle représente l'image que le réseau va reconstruire à chaque fois que sera présentée une image de l'individu i. Cette image de référence est l'image de i qui est la plus proche de la moyenne des images de i, ainsi elle est sémantiquement la plus représentative de l'individu i.

Une image différente de l'image de référence a ensuite été extraite de SET1 pour chaque individu, afin de former l'ensemble de validation.

L'ensemble d'apprentissage est finalement composé de 716 images (incluant les 20 images de référence), tandis que l'ensemble de validation est composé de 20 images (une par individu de SET1).

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

$$E = \frac{1}{2}\|\boldsymbol{o_p} - \boldsymbol{t_p}\|^2 \qquad (2)$$

où $\boldsymbol{o_p}$ et $\boldsymbol{t_p}$ sont les valeurs de la sortie ($\boldsymbol{o}utput$) et de la

cible (*target*) respectivement pour un visage p. Minimiser cette fonction de coût revient à forcer le réseau à apprendre conjointement les vecteurs de reconstruction (poids de la dernière couche) ainsi qu'une projection invariante aux transformations de la base d'apprentissage (couches C_1 à C_5).

Une phase de validation est réalisée après chaque itération lors de l'apprentissage sur l'ensemble de validation précédemment créé. Cette phase a pour but d'éviter un sur–apprentissage, et assure ainsi une meilleure généralisation du réseau.

4 Protocole de test et résultats

Pour tester notre réseau, nous utilisons le protocole de test suivant : Pour une image test I_t de SET2, celle–ci est présentée en entrée du réseau et le résultat $G_W(I_t)$ (i.e. le vecteur de caractéristiques extrait de F_6) est comparé à chacun des 26 modèles m_p de la base. Le modèle m_p de l'individu p est calculé comme étant le vecteur moyen des projetés des images de l'individu p. Si le modèle correspondant à I_t est le modèle le plus proche de $G_W(I_t)$, alors I_t est trouvée au rang 0. Si c'est le deuxième modèle le plus proche, alors I_t est trouvée au rang 1, etc. Les résultats obtenus pour les 134 images de SET2 sont regroupés dans le tableau 3.

Rangs	Correspondances cumulées	%
0	115	85,8%
1	129	96,2%
2	132	98,5%
3	134	100%

Table 3: Rangs cumulatifs obtenus pour les images de SET2

Le tableau 3 montre donc que dans environ 85% des cas, une image test est reconnue du premier coup. Pour notre jeu de test, le réseau ne présente au pire que 3 faux positifs.

5 Conclusion

L'approche proposée pour la reconnaissance faciale dans le domaine infrarouge fonctionne pour de basses résolutions (les images sont de taille 56×46), intéressantes dans le cadre de futures applications télécoms. L'approche utilise une architecture particulière de réseau de neurones convolutionnels. Celle–ci projette un visage infrarouge dans un espace de plus faible dimension, où la reconnaissance proprement dite est effectuée. Les tests effectués sur une base restreinte étant encourageants, une extension pour la fusion des modalités visible et infrarouge est en cours de réalisation.

References

[1] www.cl.cam.ac.uk/research/dtg/attarchive/ face-database.html.

[2] M. Akhloufi and A. Bendada. Thermal faceprint: A new thermal face signature extraction for infrared face recognition. In *Computer and Robot Vision*, pages 269–272, 2008.

[3] 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.

[4] S. Duffner and C. Garcia. Face recognition using non-linear image reconstruction. In *i-LIDS: Bag and vehicle detection challenge*, pages 459–464, 2007.

[5] Di Huang, Yun-Hong Wang, and Yi-Ding Wang. A robust infrared face recognition method based on adaboost gabor features. *International Conference on Wavelet Analysis and Pattern Recognition*, 2007.

[6] 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.

[7] 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.

[8] 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.

[9] Hichem Sahbi. Kernel PCA for similarity invariant shape recognition. *Neurocomputing*, 70(16-18):3034–3045, 2007.

[10] Diego A. Socolinsky and Andrea Selinger. Thermal face recognition in an operational scenario. In *Computer Vision and Pattern Recognition (2)*, pages 1012–1019, 2004.

[11] 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.