École nationale
de la statistique
et de l'administration
économique

université
PARIS-SACLAY

RAPPORT DU PROJET DE *Compressed sensing*

Reconnaissance faciale par représentation *sparse*

Sommaire

Benjamin Phan
Jean-Michel Roufosse

Mars 2018

1 Introduction

La reconnaissance faciale est un problème très compliqué avec les techniques d'apprentissages classiques. En effet, une image est un objet de très grande dimension ; par exemple une image de taille 1000×1000 contient 10^6 pixels, qui contiennent chacun l'information de la couleur qui peut se résumer à un nombre dans le cas d'une image en échelle de gris ou d'une combinaison de plusieurs nombres (généralement 3 ou 4) dans le cas d'une image en couleur. Dans une telle situation, la réduction des dimensions semblent essentielle, notamment d'un point de vue computationnel. Nous nous appuyons ici des travaux de Arvind Ganesh, Andrew Wagner, Zihan Zhou, Allen Y. Yang, Yi Ma et John Wright sur la reconnaissance faciale par représentation *sparse*. En particulier, nous allons présenter la méthode dite *parse representation-based classification (SRC)* qu'ils ont introduite. Elle est construite à l'aide de minimisation ℓ_1.

Cette méthode peut être utilisée pour reconnaître un visage parmi une base de plusieurs visages connus à l'aide d'images d'entraînement. Elle tente de pallier à certains problèmes telles que l'occultation d'une partie du visage, une luminosité non consistante, positions du visage et expressions variées, ou encore une corruption de l'image, qui font de la reconnaissance faciale une tâche spécialement ardue. Il peut y avoir diverses applications pratiques à cette méthode notamment pour la recherche d'image, le *tag* de personnes sur des photos (réseaux sociaux) et la sécurité (identification de personnes ou bien accès contrôlé par reconnaissance faciale).

Le *compressed sensing* semble adapté à notre problème puisqu'on cherche un seul sujet parmi la base de données qui correspond le mieux à l'image. L'article fait appel à différents outils du *compressed sensing* et l'algorithme proposé est inspiré en particulier de la minimisation ℓ_1 et des projections aléatoires. Cependant la reconnaissance faciale s'écarte un peu des cas habituels puisque souvent les données ne vérifient pas quelques hypothèses comme l'isométrie restreinte (*restricted isometry property*). De plus, par la nature du problème (surtout en ce qui concerne l'alignement imparfait) il faudra parfois résoudre des problèmes de représentation *sparse* sous des contraintes non linéaires.

2 Présentation théorique du problème

2.1 *Sparse Representation-based Classification*

Nous traitons en premier lieu le cas où le visage est bien aligné de face mais où il peut y avoir de grandes variations d'illumination. Dans ce cas de figure, la représentation *sparse* ainsi que la minimisation ℓ_1 peut se faire assez aisément.

On suppose qu'on dispose d'une base d'entraînement constituée d'images labellisées : $\{\phi_i, l_i\}$ des C sujets d'intérêt (par exemples des personnes). On suppose que $\phi_i \in \mathbb{R}^m$ (on concatène les colonnes de la matrice des pixels afin d'avoir un seul grand vecteur) et que $l_i \in \{1, \dots, C\}$ désigne le sujet représenté. L'objectif est de pouvoir déterminer à partir d'une nouvelle image $y \in \mathbb{R}^m$ si l'un des C sujets est représenté et l'identifier dans le cas échéant.
Certains travaux montrent que si on dispose de suffisamment d'images pour un sujet, on peut trouver un sous-espace linéaire de petite dimension (≈ 9) près duquel se trouvent les images de ce sujet. Il est alors possible de bien représenter une nouvelle image y du sujet i par une combinaison linéaire des images d'entraînement :

$$y \approx \sum_{j|l_j=i} \phi_j c_j = \Phi_i c_i \tag{1}$$

Où Φ_i est la concaténation des ϕ_j avec les images j représentant l'individu i. D'après l'article, il est raisonnable de faire cette hypothèse d'approximation par un sous-espace linéaire en considérant le visage humain comme un objet convexe et Lambertien (ou orthotrope) c'est à dire que la luminance

est la même dans toutes les directions (ce qui ne serait pas raisonnable pour une surface métallique par exemple). Dans ces conditions une base d'entraînement avec 9 images de luminosités bien choisies pour chaque sujet suffirait.
Il reste quand même un problème : pour une nouvelle image y, on ne sait pas quel sujet i est représenté *a priori*. On peut toutefois réécrire l'équation (1) de la manière suivante :

$$y = [\Phi_1; \dots; \Phi_C]\, c_0 = \Phi c_0 \ \in \mathbb{R}^m \tag{2}$$

où

$$c_0 = \left[\dots; 0^T; c_i^T; 0^T; \dots\right]^T \ \in \mathbb{R}^n$$

Si on arrive à trouver un coefficient c concentré sur une seule classe, cela donnerait une forte indication que l'image y représenterait le sujet correspondant à cette classe.
Le but du SRC est de trouver un tel c_0. À noter qu'en général, le vecteur c_0 est très *sparse* avec seulement une fraction $\frac{1}{C}$ des coefficients qui sont non nuls et est alors la solution la plus *sparse* de l'équation $y = \Phi c_0$. Il est alors possible dans la plupart des cas de trouver la solution par minimisation ℓ_1. Le problème d'optimisation est le suivant :

$$\min_c \|c\|_1 \ \ s.c. \ \ \|y - \Phi c\|_2 \leq \epsilon \tag{3}$$

où $\epsilon \in \mathbb{R}$ représente le niveau de bruit de l'observation.
Cette méthode peut même fonctionner avec des images à très basse résolution (12×10) pour lesquelles l'équation $y = \Phi c$ est sous-déterminée.
Une fois le vecteur c trouvé par minimisation, on peut par exemple définir

$$\alpha_i \doteq \|c_i\|_1 \,/\, \|c\|_1$$

Il suffit alors de choisir le i qui maximise α_i ou bien de considérer qu'aucun des sujets n'est représenté si la valeur maximale des α_i est inférieure à un seuil prédéterminé (dans ce cas, aucun des sujets i ne se démarque suffisamment dans les coefficients du vecteur c trouvé).
Nous allons ensuite voir comment étendre le problème à des cas moins idéaux avec d'éventuels occultations ou mauvais alignements ainsi que comment réduire la complexité de l'algorithme d'optimisation (3).

2.2 Réduction de dimension

Le problème majeur est que la dimension des images peut être dans les millions ce qui est énorme alors que généralement le nombre d'observations est proportionnel au nombre de sujets ce qui peut être bien inférieur. De plus, la taille des données affecte directement la complexité de l'algorithme. C'est pourquoi on cherche à réduire les données à un espace plus petit : $\mathbb{R}^d$ avec $d \ll m$ de telle sorte que les données projetées gardent leurs propriétés essentielles. Il existe diverses méthodes permettant de faire ce genre de réduction de dimension comme l'analyse en composante principale ou l'analyse discriminante linéaire. On peut représenter une projection linéaire de la facon suivante :

$$\tilde{y} \doteq Ay \approx A\Phi c = \Psi c \in \mathbb{R}^d \tag{4}$$

L'équation $A\Phi c = \tilde{y}$ peut avoir plusieurs solutions. Cependant on cherche la solution la plus *sparse*. On a donc le programme de minimisation suivant :

$$\min_c \|c\|_1 \ \ s.c. \ \ \|A\Phi c - \tilde{y}\|_2 \leq \epsilon \tag{5}$$

On peut alors se demander si le choix de la projection A affecte notre capacité à retrouver c_0. En particulier les projections aléatoires sont intéressantes en *compressed sensing*. Si c est *sparse* dans une base orthonormé connue alors la minimisation ℓ_1 permet de retrouver c avec relativement peu d'observations avec grande probabilité. Bien que Φ n'est pas orthonormale, il est intéressant de se pencher sur les performances des projections aléatoires. La réduction de dimension entraîne généralement une baisse du taux de reconnaissance mais la baisse n'est pas très importante quand d est suffisamment grand. Subir cette perte permet toutefois de réduire le coût computationnel de l'algorithme.

2.3 Reconnaissance avec corruption et occultation

En pratique, une partie du visage peut être occultée (lunettes de soleil, écharpe, obstacle, ...). De plus, d'autres conditions peuvent ne pas être optimales avec par exemple de l'ombre sur une partie du visage ou autre particularité due aux conditions réelles. On peut même imaginer une corruption de l'image. Dans ce genre de conditions le modèle linéaire $y \approx \Phi c$ peut être un peu optimiste. On peut à la place poser :

$$y = \Phi c + e \tag{6}$$

où e est un terme d'erreur dont les entrées non nulles correspondent aux pixels corrompus.
De par la nature des nuisances, e peut être de grande amplitude ce qui n'est pas très adapté aux méthodes conçues pour des petites erreurs, comme les moindres carrés. Cependant, l'erreur est bien souvent *sparse* (seule une portion $\rho < 1$ des pixels est corrompue). On peut alors résoudre le problème en cherchant des solutions *sparses* de c et de e. On étend alors le modèle précédent :

$$\min_{c,e} \|c\|_1 + \|e\|_1 \;\; s.c.\; y = \Phi c + e \tag{7}$$

L'article indique que ce modèle marche bien pour des occultations allant jusqu'à 20% du visage ou bien bien une corruption aléatoire de 70% des pixels. Il y est également souligné qu'il est surprenant que cette méthode fonctionne. En effet, on peut poser $B = [\Phi I] \in \mathbb{R}^{m*(m+n)}$. Le programme d'optimisation (7) est alors équivalent à :

$$\min_{y=Bw} \|w\|_1 \tag{8}$$

avec $w = [c^T e^T]^T$.

Puisque les colonnes de Φ sont toutes des images, elles sont toutes similaires dans l'espace $\mathbb{R}^m$. La matrice B viole les conditions classiques de récupération *sparse* uniforme (*uniform sparse recovery*) comme les critères d'incohérence ou la propriété d'isométrie restreinte. Contrairement aux cas classiques de *compressed sensing*, B n'est pas très homogène : les colonnes de Φ sont cohérentes entre elles alors que celles de I ne le sont pas du tout. Toutefois, malgré cette configuration quelque peu inhabituelle, les résultats sont plutôt bons.

2.4 Alignement du visage

Le modèle précédant permet de prendre en compte à la fois des problèmes d'occultations modérées et de variation d'illumination. Cependant, pour avoir un système de reconnaissance faciale satisfaisant, il faut pouvoir gérer un mauvais alignement du visage, c'est à dire les cas où la pose n'est pas parfaitement droite ou frontale.

On suppose qu'on a une image déformée $y = y_0 \circ \tau^{-1}$ d'une image originale y_0 par une déformation τ en 2D. La déformation τ bouge souvent la région détectée du visage de la zone optimale ce qui fausse la représentation *sparse* de y lorsqu'elle est effectuée à l'aide d'images d'entraînement parfaitement alignées. Cependant, s'il est possible de trouver de manière efficace la vraie perturbation τ, alors on peut trouver une représentation *sparse* c pour y_0. Le modèle devient alors :

$$y \circ \tau = \Phi c + e \tag{9}$$

et on souhaiterait résoudre le problème d'optimisation suivant :

$$\min_{c,e,\tau} \|c\|_1 + \|e\|_1 \;\; s.c.\; y \circ \tau = \Phi c + e \tag{10}$$

Cependant, estimer à la fois c, e et τ est un problème difficile qui n'est pas linéaire. En particulier, comme il y a plusieurs classes dans la matrice Φ, le problème d'optimisation a plusieurs minima

locaux correspondant chacun à aligner l'image avec l'un des sujets. Pour pallier à cette difficulté, on cherche d'abord un alignement avec chacun des sujets k :

$$\min_{c,e,\tau_k} \|e\|_1 \;\; s.c.\; y \circ \tau_k = \Phi_k c + e \tag{11}$$

(On ne pénalise plus la norme de c_1 puisque Φ_k ne contient que des images du sujet k)

Si on dispose d'une première estimation de la vraie transformation τ_k assez bonne (avec un détecteur de visage par exemple), on peut résoudre de manière itérative les équations linéarisées :

$$\min_{c,e,\Delta\tau_k} \|e\|_1 \;\; s.c.\; y \circ \tau_k^i + J_k^i \Delta\tau_k = \Phi_k c + e \tag{12}$$

où τ_k^i est l'estimation de τ_k à l'étape i, $\Delta\tau_k$ est un pas de mise à jour de τ_k pour obtenir τ_k^{i+1} et $J_k^i = \nabla_{\tau_k}(y \circ \tau_k^i)$ est la jacobienne de $y \circ \tau_k^i$ par rapport à τ_k.
Cet algorithme peut être vu comme une généralisation de la méthode de Gauss-Newton pour la minimisation de la composition d'une fonction objectif pas *smooth* (ici la norme ℓ_1) avec une transformation différentiable. Ce genre de problème a déjà été étudié et l'article spécifie que dans notre cas, la convergence devrait se faire en 10 à 15 étapes. Par ailleurs, on peut améliorer l'algorithme en normalisant $y \circ \tau_k$ ce qui empêche un cas dégénéré où l'algorithme zoomerait sur une zone sombre de l'image.

Une fois tous les τ_k trouvés, on peut appliquer leurs inverses aux Φ_k pour que la base d'entraînement soit alignée avec y. On peut alors se ramener au problème d'optimisation (7). Ensuite, pour déterminer le sujet représenté par y, on choisit la classe k qui minimise la distance l_2 entre y et l'approximation $\hat{y}_k = \Phi_k \delta_k(\hat{c})$ avec $\delta_k(\hat{c})$ étant le vecteur $\hat{c}$ en remplaçant les coefficients ne correspondant pas à la classe k par 0.

3 Résultats

Pour tester notre algorithme, nous avons eu recours à une base de données de visages utilisée aussi par les auteurs de l'article : AR Database. Il s'agit d'une base de donnée de 50 hommes et 50 femmes, chacun pris en photos en deux sessions de 13 photos. À chaque session, 4 photos ont été prises de face, avec des expressions du visage différentes, puis 3 photos correspondent à la même personne mais avec une variation de la lumière dans la photo, puis 3 autres photos ont été prises avec des lunettes de soleil et enfin 3 dernières photos prises avec une écharpe. Les lunettes de soleil cachent 20% du visage, l'écharpe quasiment 40%.

Nous avons testé plusieurs algorithmes de *feature reduction*, notamment ceux mentionnés dans l'article (*Eigenfaces*, *Fisherfaces*, *Randomfaces*). Ces algorithmes permettent de gagner énormément de temps, car une version réduite de l'image est analysée, et non pas toute l'image de taille 120x165. Nous avons remarqué que les résultats étaient plus satisfaisants en utilisant une réduction *Fisherfaces*, alors toutes nos analyses ont été faites avec cette méthode.

Nous avons plusieurs résultats à illustrer. D'abord, il faut comparer ce que nous obtenons avec les résultats de l'article. L'article arrive à reconnaître à 87 % la bonne personne malgré les lunettes de soleil, puis à 59,5 % si cette personne a une écharpe. Ces chiffres pour nous sont respectivement 50% et 40 %. Nous y reviendront plus en détail dans les section dédiées. Nous nous sommes aussi intéressés à ajouter dans la base de test les photos prises sous différentes luminosités, même avec des lunettes de soleil et une écharpes. Les résultats sont nettement moins bons, 34,3 % pour les lunettes de soleil et 25,3 % pour les photos où les sujets portent une écharpe. Nous présentons ici une synthèse de nos résultats :

Type d'images	Nos résultats	Résultats de l'article
Visages de face à luminosité constante	90,5 %	-
Avec une variation de luminosité	65,3 %	-
Avec des lunettes de soleil	50 %	87 %
Avec une écharpe	40 %	59,5 %

FIGURE 1 – Résumé des résultats

3.1 Photos de face - sans occultation ni variation de lumière

Nous avons donc ici en base de test les 4 images de la première session où les sujets sont de face et n'ont aucun accessoire. La lumière est également la même pour chacune des photos. On teste sur les 4 images correspondantes de la deuxième session. Notre algorithme va nous sortir trois images : d'abord la photo test, ensuite la combinaison linéaire des photos de la base de train qui se doit d'être assez *sparse* ainsi que de minimiser l'erreur, et enfin l'affichage de l'erreur même.

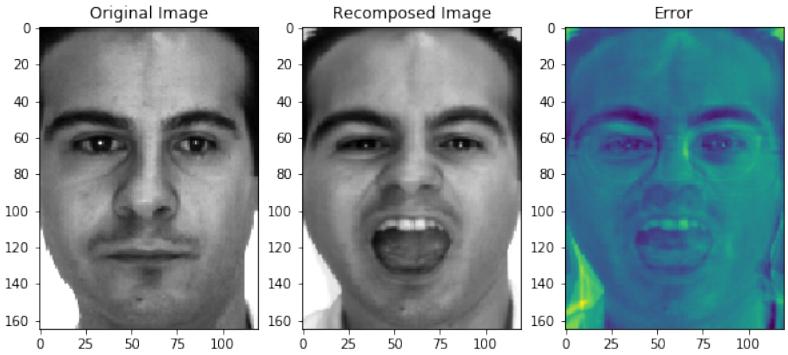
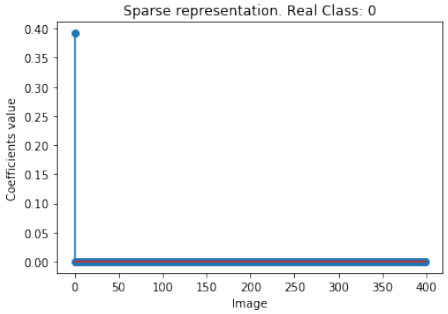
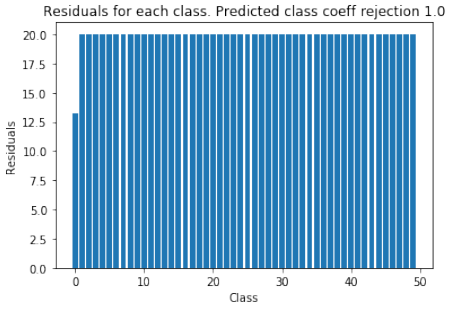


FIGURE 2 – Sujet bien identifié, recomposé avec plusieurs images de la base d'entraînement

Ensuite, notre algorithme nous montre la représentation du vecteur ainsi trouvé par minimisation en norme ℓ_1. On remarque en Figure 3a un pic correspondant bien à un fort coefficient dans les images du train venant du même sujet. Ainsi, en regardant Figure 3b c'est bien l'individu 0 qui minimise les résidus.



(a) Coefficients du vecteur sparse



(b) Graphe des résidus pour chaque personne

FIGURE 3 – Résultats pour une image frontale - Concentration sue les images du premier sujet et résidus minimisés pour ce même sujet

Le taux de succès est à 90,5%. Il y a parfois des erreurs, et cela concerne les personnes ayant des attributs physiques similaires (à peu près la même barbe, teint de couleur de peau etc ...) Voici un exemple :

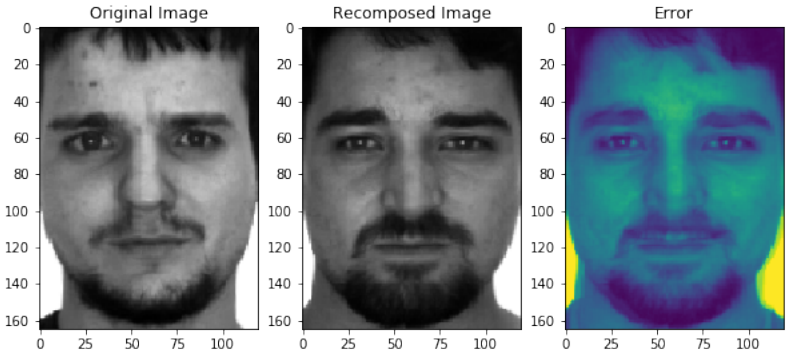


FIGURE 4 – Un exemple où deux personnes sont confondues

3.2 Avec des variations de lumières

Pour étudier les photos avec des variations de lumières, on utilise les 8 photos frontales venant des deux session (4 dans la première session et 4 dans la deuxième). Les résultats sont moins bons, mais méritent d'être commentés, avec d'abord un cas où notre algorithme a donné un bon résultat :

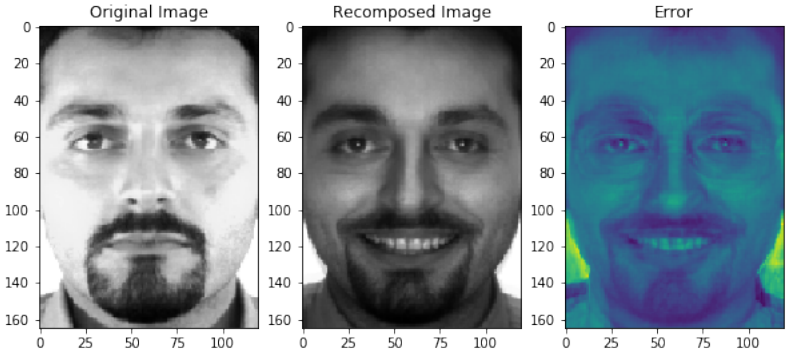
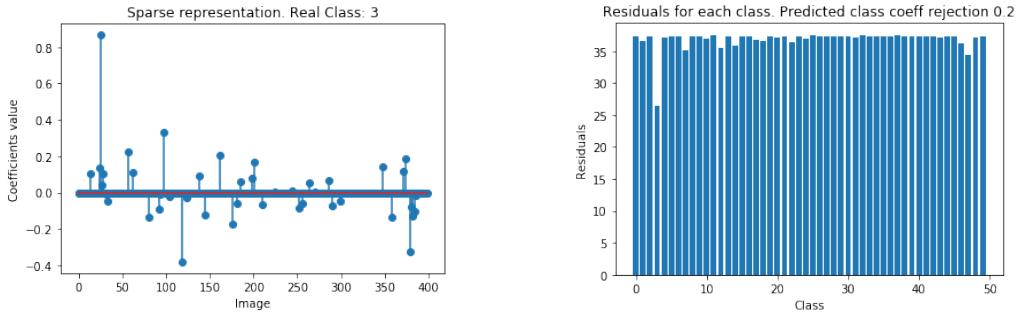


FIGURE 5 – Personne bien identifiée malgré une forte luminosité dans la photo

Le vecteur des coefficients a beau être moins *sparse*, la personne ayant le plus petit des résidus reste identifiable.



(a) Coefficients du vecteur sparse

(b) Graphe des résidus pour chaque personne

FIGURE 6

Le taux de succès est cependant plus bas qu'avec une luminosité ordinaire. On se retrouve à 65,3 %.

3.3 Avec les lunettes de soleil

On s'intéresse maintenant aux photos prises avec des lunettes de soleil. Il y a donc deux sources de difficultés principales : la luminosité mais aussi à cause les yeux qui ne peuvent plus être identifiés. Les lunettes couvrent environ 20% de l'image. On se retrouve quand bien même à pouvoir détecter la moitié des individus si on ne compte pas les photos où il y a des variations de luminosités.

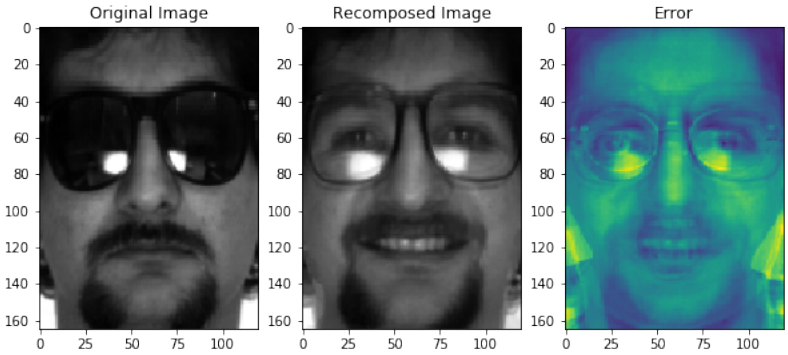


FIGURE 7 – Personne bien identifiée avec les lunettes de soleil

Les erreurs sont souvent causées pas des personnes ayant des lunettes qui ressemblent par leur forme aux lunettes de soleil utilisées (cf notebook pour plus d'exemples d'images).

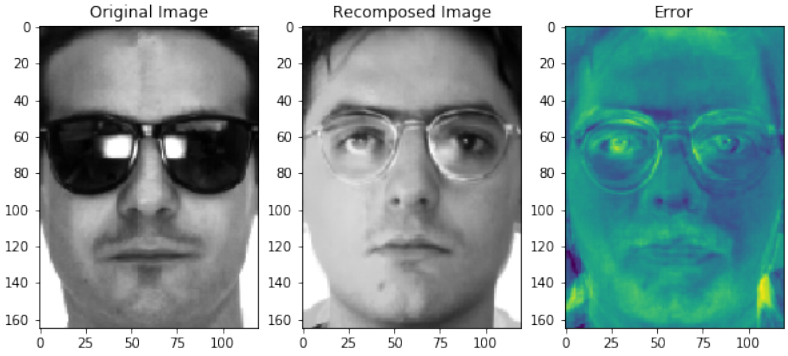


FIGURE 8 – Personne mal identifiée avec les lunettes de soleil

3.4 Avec une écharpe

L'occultation est ici à plus de 40%. On se rend compte qu'il y a beaucoup d'erreurs pour reconnaître les personnes qui portent une écharpe, avec en plus des variations dans la luminosité. Ceux qui sont reconnus ont un signe distinctif : lunettes, teint de peau plus foncé ou plus clair. Mais encore, ce chiffre est presque réduit de moitié lorsque l'on ajoute les photos de personnes avec différents luminosités. En effet, nous avons vu que la luminosité réduisait le taux de succès de notre algorithme, puis qu'avec les lunettes de soleil obstruant 20% de l'image ce score diminuait également. Ainsi, nous obtenons un résultat de 25,3 %.

Cependant, en regardant les erreurs, on remarque que c'est dû à notre algorithme qui souhaite identifier à chaque fois un barbu. En effet, l'occultation noircissant les pixels, une barbe noire aurait le même effet. D'ailleurs, les hommes barbus sont mieux reconnus. Nous retrouvons également cette conclusion dans l'article.

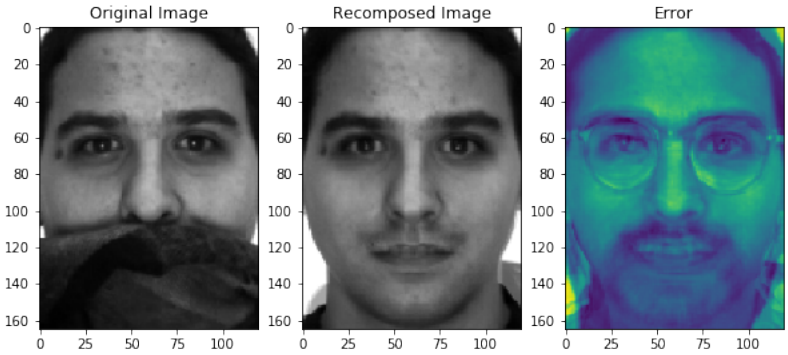


FIGURE 9 – Bien reconnu

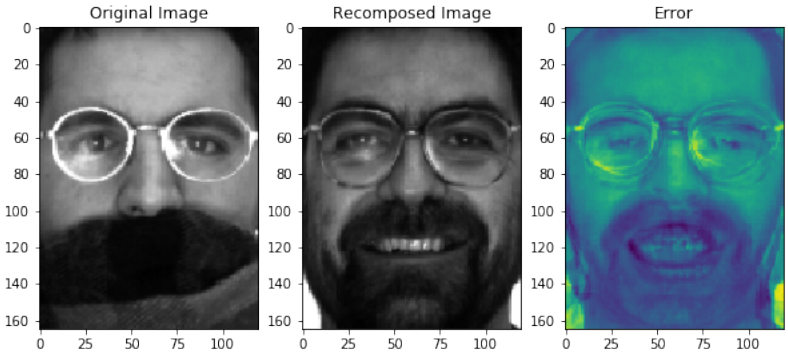
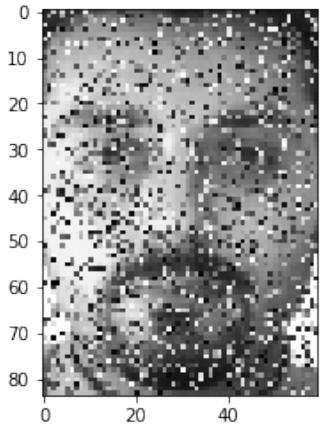


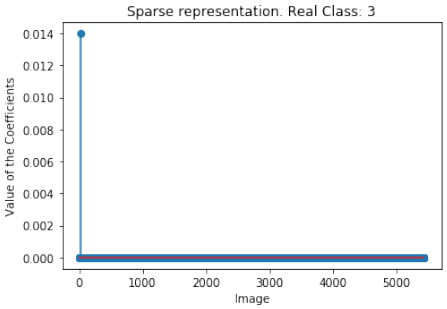
FIGURE 10 – Mauvaise reconnaissance, à cause de la barbe

3.5 Avec du bruit dans la photo

On a aussi analysé les photos lorsqu'elles étaient pixelisées. Cela signifie que nous avons aléatoirement changé la valeur de plusieurs pixels dans la photo originale. Une photo qui était bien reconnue à la base pouvait toujours être reconnue malgré 65 % des pixels de la photo ainsi transformées comme ci-dessous :



(a) Image ainsi pixelisée



(b) Vecteurs des coefficients

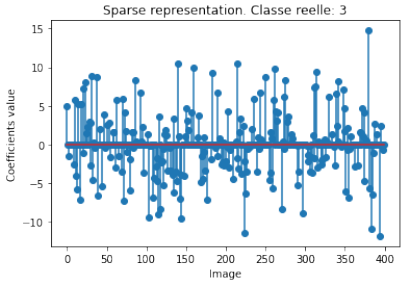
FIGURE 11 – Résultat pour une image Pixelisée

Ceci est très intéressant puisqu'un être humain aurait du mal à reconnaître un visage ainsi transformé.

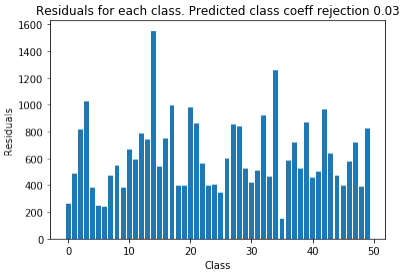
3.6 Avec une image choisie au hasard sur Internet

Tout comme dans l'article, nous avons voulu vérifier qu'une image n'ayant rien à faire dans la base de donnée puisse être détectée, grâce à leur indicateur SCI qui devrait être faible.

(a) Image choisie au hasard sur internet



(b) Sa représentation spectrale



(c) Les résidus ainsi formés

FIGURE 12

Au vue des résultats, le spectre n'est plus du tout sparse et les résidus sont anarchiques.

Avec un score SCI de 0,03 , l'image doit bien être rejetée si on prend comme seuil 0,05.

4 Conclusion

Les auteurs ont présenté une méthode permettant une de faire une classification supervisée d'images à l'aide d'une représentation *sparse*. Elle permet de reconnaître des visages connus et rejeter des visages inconnus. La représentation *sparse* permet de pouvoir traiter aisément des images alors que ce sont des données en très grande dimension, ce qui peut être problématique pour l'usage de certains outils computationels.

Nous obtenons des résultats plutôt bon pour reconnaître un visage en conditions optimales (visage de face à luminosité constante). La méthode exposée permet aussi de reconnaître des visages dans des conditions plus difficiles, comme des variations de luminosités, une partie du visage occultée ou encore une corruption d'une partie des pixels. Dans ces conditions, les performances de notre algorithme baissent fortement. Ces baisses sont souvent compréhensibles : confusion d'une écharpe avec une barbe, recherche de personne portant des lunettes lorsque l'image de test contient des lunettes de soleil, ...

On note toutefois que la baisse est plus forte que dans les résultats exposés par les auteurs. Ceci est peut être dû au fait que nous ne disposons pas d'énormément d'images dans notre base d'entraînement. En effet, l'article indique qu'on peut se contenter de 9 images pour chaque personne avec une luminosité bien choisie alors que nous ne disposons que de 4 variations de luminosité pour chaque sujet. Ainsi, les sous-espaces linéaires associés à chaque sujet ne sont peut être pas bien identifiés.

Références

[1] A.M. Martinez and R. Benavente. The AR Face Database. CVC Technical Report #24, June 1998.

[2] J. Wright, A. Yang, A. Ganesh, S. Sastry, and Y. Ma. Robust face recognition via sparse representation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 31(2) :210 – 227, 2009

[3] Ganesh, A., Wagner, A., Zhou, Z., Yang, A., Ma, Y., & Wright, J. (2012). Face recognition by sparse representation. In Y. Eldar & G. Kutyniok (Eds.), Compressed Sensing : Theory and Applications (pp. 515-539). Cambridge : Cambridge University Press. doi :10.1017/CBO9780511794308.013

[4] MicrosoftResearch. "Robust Face Recognition via Sparse Representation." YouTube, YouTube, 6 Sept. 2016, www.youtube.com/watch ?v=ZoCwNIXU-Hk.