HAL
open science

Cascade boostée de forêts aléatoires pour la reconnaissance d'émotions faciales

Léonard Benedetti, Joris Garnier, Irwin Girard, Lionel Prevost

To cite this version:

Léonard Benedetti, Joris Garnier, Irwin Girard, Lionel Prevost. Cascade boostée de forêts aléatoires pour la reconnaissance d'émotions faciales. RFIAP, Jun 2018, Marne-la-Vallée, France. hal-01860000

HAL Id: hal-01860000

https://hal.science/hal-01860000v1

Submitted on 22 Aug 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Cascade *boostée* de forêts aléatoires pour la reconnaissance d'émotions faciales

Léonard BENEDETTI Joris GARNIER Irwin GIRARD Lionel PREVOST

Laboratoire Learning, Data & Robotics
ESIEA, Paris, France

benedetti@et.esiea.fr

Résumé

Nous explorons ici l'application du boosting *aux forêts aléatoires, ces dernières combinant* bagging *et* attribute bagging. *La capacité à focaliser sur les hyperplans séparateurs du* boosting *peut permettre d'améliorer, localement, les performances d'une cascade* boostée *de forêts. Nous montrons expérimentalement l'efficacité de cette solution pour la reconnaissance d'émotions faciales et analysons l'impact du biais d'identité sur les systèmes dédiés à cette tâche. Les performances, évaluées sur une base largement utilisée, se situent très favorablement par rapport à l'état de l'art.*

Mots-clés

Forêts aléatoires, *boosting*, cascade de classifieurs, reconnaissance d'émotions faciales.

Abstract

In this article, we explore the application of boosting to random forests, the latter combining bagging and attribute bagging. The ability to focus on the boosting separating hyperplanes can help to improve, locally, the performance of a boosted cascade of forests. We experimentally demonstrate the effectiveness of this solution for the recognition of facial emotions and analyze the impact of identity bias on systems dedicated to this task. Performance, evaluated on a widely used dataset, can be very favourably compared with the state of the art.

Keywords

Random forests, boosting, cascading classifiers, facial emotions recognition.

1 Introduction

Les émotions faciales sont des facteurs clés de la communication humaine qui nous aident à comprendre les intentions des autres. En général, les gens perçoivent instinctivement les états émotionnels (comme la joie, la tristesse ou la colère) d'autres personnes, à travers leurs expressions faciales, leur posture et le ton de leur voix.

L'intérêt pour la reconnaissance des émotions faciales varie d'une étude à l'autre ; mais il ne cesse de croître avec le développement rapide de l'apprentissage artificiel et des interfaces homme-machine. La caméra utilisée pour capter l'information a aussi l'avantage d'être peu invasive, peu coûteuse et ne nécessite pas d'instrumenter le sujet.

Parmi les descriptions possibles de l'expression faciale, la plus ancienne est celle des émotions basiques d'Ekman [10]. Plus récemment, une description plus objective est apparue : les micro-mouvements ou unités d'action faciales [11]. Enfin, les dimensions affectives offrent une autre description en termes de positivité et d'éveil du sujet en interaction [24].

Bien que la reconnaissance automatique des émotions basiques soit considérée comme un succès dans des conditions contraintes, les performances diminuent quand les conditions se dégradent. De plus, le biais d'identité dû aux variabilités morphologiques (formes des traits du visage) et comportementales (façon que chaque individu a d'exprimer ses émotions) peut encore mettre les algorithmes de l'état de l'art en échec, en particulier dans le cas des expressions subtiles.

Parmi les algorithmes de l'état de l'art (section 2), les forêts aléatoires [5] sont une méthode d'ensembles qui présentent de nombreux avantages. Elles utilisent peu d'hyperparamètres, sont naturellement multi-classe, robustes, ont d'excellentes capacités de généralisation et sont très rapides, tant en phase d'apprentissage qu'en phase d'évaluation. De plus, elles sont incrémentales en classes et en données. Les arbres de décision qui les composent ont un biais faible mais une variance importante. En parallélisant de nombreux arbres et en combinant leur décision, la forêt parvient à réduire cette dernière.

Dans la catégorie des ensembles de classifieurs, le *boosting* [13] a une approche diamétralement opposée en combinant des classifieurs faibles au biais élevé et à la variance faible pour construire un classifieur fort. Les capacités de généralisation du *boosting* sont dues à l'apprentissage séquentiel des classifieurs faibles. En effet, le classifieur courant est entraîné sur un ensemble de données constitué, pour partie, des erreurs du classifieur précédent ; l'objectif étant que le classifieur courant ne répète pas ces erreurs.

Nous proposons ici de combiner les deux approches en entraînant une cascade de forêts aléatoires avec un ensemble

de données s'adaptant au fur et à mesure aux données les plus difficiles à discriminer. Après une analyse détaillée du processus d'apprentissage de la cascade, nous montrons expérimentalement qu'elle permet souvent d'augmenter les performances d'une forêt unique, dès que la complexité de la solution le nécessite.
L'article est organisé comme suit. La section 2 présente un bref état de l'art. La section 3 décrit en détail l'algorithme d'apprentissage de la cascade de forêts. La section 4 est dédiée à la description des jeux de données et aux expérimentations. Finalement, la section 5 propose quelques conclusions et travaux futurs.

2 État de l'art

Le domaine de la reconnaissance d'émotions à partir d'images ou de vidéos est un domaine très actif et de nombreuses méthodes ont été étudiées pour parvenir à des résultats satisfaisants.
Parmi les différentes techniques élaborées pour aborder ce problème, il convient d'abord de distinguer celles qui utilisent des images fixes de celles traitant des données temporelles (disponibles par exemple dans une vidéo).
Les systèmes se basant exclusivement sur des images statiques (photographie unique) fonctionnent en extrayant ou en inférant, sur l'image en question, différents descripteurs qui serviront comme entrée du classifieur. Une première approche, la plus conventionnelle, pour obtenir un système de reconnaissance d'émotions enchaîne séquentiellement les étapes suivantes :

1. détection des visages ;
2. extraction des descripteurs ;
3. entraînement du classifieur.

Une seconde utilise l'apprentissage profond pour réaliser conjointement les étapes 2 et 3, apprenant en même temps les descripteurs optimaux et la fonction discriminante.
Ces descripteurs extraits sont constitués la plupart du temps d'informations géométriques comme les points d'intérêts du visage (52 points pour Ghimire et Lee [14]), les distances euclidiennes entre ces différents points, ou les angles que l'on retrouve dans la maille constituée avec ces mêmes points. À ces descripteurs de formes géométriques s'ajoutent des descripteurs d'apparence qui encodent une texture : *histogram of oriented gradient* (HOG), *scale-invariant feature transform* (SIFT), *local binary patterns* (LBP) ou descripteur de Gabor [8, 25].
Quant au choix du modèle pour un tel classifieur, on observe une évolution au cours du temps. Dans les années 2000, suite aux travaux de Vapnik [28], la plupart des systèmes étaient des classifieurs tels que des machines à vecteurs de support (*Support Vector Machine*, SVM) [14, 16, 20, 21].
Par la suite, de nouvelles approches ont émergé. Le système présenté dans [15] repose par exemple sur des masques faciaux (*facial patches*). Pour obtenir une amélioration des résultats, beaucoup de projets se sont tournés vers des modèles de plus en plus complexes comme les réseaux convolutifs profonds [23, 27, 7]. Un avantage notable de ce type de modèles est leur capacité à utiliser directement l'image : il devient alors inutile d'effectuer une phase d'extraction de descripteurs, puisque ces derniers sont appris par le réseau. Cependant, de tels systèmes nécessitent un très grand nombre d'images pour être entraînés. De plus, il est très difficile d'effectuer un apprentissage incrémental sur un très petit jeu de données. Pour plus de détails sur ces techniques, [19] décrit les différentes méthodes utilisées. Il conclut d'ailleurs sur la « supériorité » des réseaux profonds dans le domaine de la reconnaissance d'émotions.
Pour les systèmes basés sur des vidéos, on cherchera plus à mesurer le déplacement des descripteurs entre chaque image pour obtenir le vecteur d'entrée du classifieur, par exemple à partir de la distance euclidienne entre les coordonnées du point de l'image précédente et de l'image courante. D'autres systèmes étendent les descripteurs bidimensionnels sur la dimension temporelle (2D + t) pour créer des descripteurs spatio-temporels comme les LBP-TOP [1]. Une solution intéressante a été proposée dans [9]. Ils considèrent des paires d'images exprimant des émotions différentes et utilisent comme descripteur l'écart entre les descripteurs statiques des deux images.

3 Méthodologie

3.1 Détection de visage et extraction de descripteurs

Afin de construire un système pour la reconnaissance d'émotions faciales, il faut d'une part être en mesure de détecter le visage sur lequel portera la prédiction, et d'autre part produire à partir de ce visage un vecteur de descripteurs qui servira d'entrée au classifieur.
La bibliothèque *OpenFace* [3] que nous avons utilisée permet, entre autres, de réaliser une détection d'un ou plusieurs visages sur des photos ou des vidéos et d'extraire des caractéristiques faciales. Nous formons ensuite à partir de ces dernières, d'une façon que nous détaillons ci-dessous, le vecteur d'entrée pour notre système.
Pour ce faire, nous avons choisi de combiner des points caractéristiques faciaux, des descripteurs géométriques et des descripteurs d'apparence. Il a été démontré (notamment dans [26]) que ce type de descripteur mixte améliorait les performances des classifieurs d'émotions faciales.
Notons également que nous cherchons à déterminer l'émotion sur des images statiques, indépendantes les unes des autres, nous traitons donc indifféremment les images et les vidéos (que l'on peut considérer comme une suite d'images indépendantes). Ces images peuvent contenir des visages frontaux ou non, avec des rotations dans le plan (roulis) et hors plan (lacet et tangage). De plus, ces visages ont une taille et une position dans l'image qui sont arbitraires. De ce fait, il est indispensable de normaliser les descripteurs qui pourraient être affectés par ces transformations pour obtenir une certaine invariance face à ces dernières.

Points caractéristiques faciaux. Nous utilisons les 68 points d'intérêts (*landmarks*) en deux dimensions fournis

par *OpenFace*. Ils correspondent aux traits du visage : les sourcils, les yeux, le nez, la bouche et la forme de la mâchoire ; ils apportent des informations sur la forme et la déformation de ces traits.
Pour la normalisation de ces premiers descripteurs, nous centrons et réduisons le nuage de points dans une boîte englobante et nous le redressons pour que le segment entre le nez et le menton soit parallèle à l'axe vertical. Remarquons qu'en procédant de cette façon, seule la rotation dans le plan du visage est normalisée. Nous limitons en effet notre étude à des visages relativement frontaux, c'est-à-dire qui présentent des rotations hors plan qui ne sont pas trop importantes ($\pm$ 20°).

Descripteurs géométriques. Les descripteurs géométriques proposés par [9] sont les distances euclidiennes entre deux points d'intérêts et les angles formés entre trois de ces points. Ces descripteurs viennent compléter l'information donnée par les coordonnées, notamment sur la déformation des traits du visage due à l'expression émotionnelle. En ce sens il s'agit de descripteurs de formes.
Le nombre de distances croît de façon quadratique, et le nombre d'angles de façon cubique, par rapport au nombre de points. Avec les 68 points d'intérêts que nous utilisons, cela représente plus de 10^5 distances et angles. Afin de conserver un nombre raisonnable de descripteurs géométriques, il est nécessaire de procéder à une sélection des valeurs à intégrer au vecteur d'entrée.
Nous proposons de sélectionner les descripteurs pour lesquels la variance expliquée par la classe est élevée par rapport aux autres. Cette valeur, notée V_Q pour le descripteur géométrique Q, est égale à la variance inter-classe (la variance des moyennes conditionnelles sachant la classe C) divisée par la variance totale :

$$V_Q = \frac{\mathrm{Var}(\mathbb{E}_C[Q])}{\mathrm{Var}(Q)}, \tag{1}$$

de plus, la variance totale étant aussi égale à la somme de la variance intra-classe (la moyenne des variances conditionnelles sachant la classe C) et de la variance inter-classe, d'après le théorème de la variance totale, on a également :

$$V_Q = \frac{\mathrm{Var}(\mathbb{E}_C[Q])}{\mathrm{Var}(\mathbb{E}_C[Q]) + \mathbb{E}[\mathrm{Var}_C(Q)]}. \tag{2}$$

En procédant ainsi, nous avons sélectionné 122 distances et angles, calculés après l'étape de normalisation sur les points décrite plus haut, qui constitueront des descripteurs géométriques dans le vecteur d'entrée.

Descripteurs d'apparence. Les histogrammes de gradients orientés (*Histogram of Oriented Gradients*, HOG), présentés par Dalal et Triggs [8] correspondent à des informations d'apparence, de contours et de textures. Ils présentent par ailleurs l'intérêt d'être peu influencés par la variation de luminosité dans l'image.
Le calcul des gradients se fait localement dans des fenêtres, de petite taille par rapport à la taille du sujet, qui recouvrent la totalité du visage (après centrage et réduction). On peut ensuite utiliser les histogrammes des directions prises par ces gradients dans les différentes fenêtres et les combiner pour obtenir nos descripteurs.
Nos descripteurs HOG sont calculés en utilisant la méthode et les paramètres présentés dans [12] et implémentés dans [18], nous utilisons donc des fenêtres de taille $k = 8$ (produisant ainsi des fenêtres de 12×12 pixels en l'espèce), sur 31 canaux différents. Cela représente dans le vecteur d'entrée 4 464 valeurs réelles comprises entre 0 et 1.

3.2 Cascade *boostée* de forêts aléatoires

Forêts aléatoires. La forêt aléatoire (*Random Forest*) est un ensemble d'arbres de décision aléatoires. Ce nom leur vient du processus d'apprentissage qui se fait de façon doublement aléatoire. D'une part, chaque arbre est entraîné sur un sous-ensemble S de données d'apprentissage tirées avec remise suivant le principe de *bagging* proposé par Breiman [4]. En outre, chaque nœud de l'arbre construit une règle de décision dans un sous-espace de dimension réduite (par rapport à la dimension initiale des données). Le sous-ensemble de descripteurs utilisés étant lui aussi tiré aléatoirement, on parle alors d'*attribute bagging* [6]. Dans le cadre de notre système, le critère de segmentation utilisé pour juger de la qualité de la séparation est l'indice de diversité de Gini. Chaque arbre de décision aléatoire étant entraîné sur S, il dispose naturellement d'un ensemble de validation (appelé *out-of-bag*) constitué des exemples exclus de S. Dans le cas des arbres de classification, chaque arbre produit une probabilité d'appartenance de la donnée à chaque classe. Une moyenne des probabilités est calculée, résultant en une prédiction unique pour la forêt. Le nombre d'arbres constituant la forêt constitue un premier hyperparamètre.

Cascade de forêts aléatoires. La multiplicité des facteurs morphologiques et comportementaux impactant l'expression d'une émotion fait qu'il est très peu probable qu'une unique forêt (de dimension raisonnable) puisse gérer toute la variabilité du problème. Une solution est de distribuer le problème sur plusieurs forêts. L'entraînement sera réalisé en cascade, en spécialisant les forêts successives sur des données de plus en plus « difficiles » (proches de l'hyperplan séparateur). Une fois l'entraînement des forêts réalisé, les prédictions de ces dernières sur une donnée seront combinées. On se rapproche ici du *boosting* en raison de la stratégie d'apprentissage séquentielle des classifieurs et de la génération des ensembles d'entraînement des différents étages de la cascade (décrite ci-dessous). D'aucuns pourraient regretter l'utilisation de forêts aléatoires comme classifieurs « faibles » et nous reviendrons sur ce point lors des expérimentations (voir sous-section 4.2).

Apprentissage de la cascade. Il s'agit d'un processus récursif que nous décrivons en détail ci-dessous. On appellera ensemble d'entraînement S_k l'ensemble à partir duquel on extrait aléatoirement, dans des proportions fixées à 60 %/40 %, les ensembles d'apprentissage $S_{k,\mathrm{APP}}$ et de validation $S_{k,\mathrm{VAL}}$ d'une forêt aléatoire.

Le processus d'apprentissage de l'étage k de la cascade est le suivant :

1. Une k^{e} forêt aléatoire est créée à partir de S_k en utilisant une méthode de validation croisée simple : la forêt aléatoire est entraînée sur $S_{k,\text{APP}}$ tandis que $S_{k,\text{VAL}}$ pourra être utilisé pour évaluer le classifieur ainsi produit.

2. Une fois cette forêt entraînée, il est ainsi possible de l'utiliser pour classer les données de l'ensemble d'entraînement S_k (qui réunit donc les données utilisées pour l'entraînement et les données de validation). Suivant la prédiction, nous différencions :

   — l'ensemble $S_{k,\text{ERR}}$ des images mal classées (dont l'émotion n'a pas été correctement prédite) qui doit impérativement intégrer l'ensemble d'entraînement de l'étage $k+1$ de la cascade ;

   — les images bien classées que l'on va sélectionner en fonction de leur proximité à la frontière de décision. Cette information n'étant pas directement disponible dans le cas des forêts aléatoires, nous utiliserons la probabilité associée à la classe prédite calculée par la forêt : plus cette probabilité est élevée, moins le classifieur est incertain. Ainsi, et afin de déterminer quels sont les cas problématiques sur lesquels les étages suivants doivent se concentrer, ces données bien classées sont triées selon cette probabilité dans l'ordre croissant. Les premiers exemples de cette liste triée seront alors réunis dans $S_{k,\text{AMB}}$, l'ensemble des d'exemples ambigus.

   Les deux ensembles $S_{k,\text{ERR}}$ et $S_{k,\text{AMB}}$ sont ensuite réunis pour former l'ensemble d'entraînement de l'étage $k+1$.

   Afin de se focaliser davantage sur les exemples difficiles, proches de la frontière, au fur et à mesure de la cascade, la taille de l'ensemble d'entraînement de la forêt à l'étage $k+1$ sera pondérée d'un facteur de réduction $r \in]0\,;1[$ par rapport à l'étage précédent :

$$\text{card}(S_{k+1}) = \lfloor r \times \text{card}(S_k) \rfloor. \tag{3}$$

   Sachant de plus que :

$$S_{k+1} = S_{k,\text{ERR}} \cup S_{k,\text{AMB}}, \tag{4}$$

   on en déduit aisément $\text{card}(S_{k,\text{AMB}})$, le nombre d'exemples ambigus nécessaires. Enfin, dans le cas où $\text{card}(S_{k+1}) < \text{card}(S_{k,\text{ERR}})$, on réentraîne l'étage k (pour éviter de boucler indéfiniment, on renvoie une erreur si le problème persiste après deux entraînements du même étage).

3. Les étapes précédentes sont répétées autant de fois que le nombre maximum d'étages de la cascade (fixé par l'utilisateur) ou jusqu'à ce que $\text{card}(S_{k+1})$ soit trop faible et ne permette plus l'entraînement d'une nouvelle forêt.

FIGURE 1 – Images provenant de CK+

Processus de décision. À l'issue de l'apprentissage de notre système, nous disposons de N forêts, notées F_k avec $k \in \{1, 2, \ldots, N\}$. Nous cherchons à prédire, à partir de l'observation q dont le vecteur est noté X_q, l'émotion exprimée P_q (parmi les n classes d'émotions, notées C_i avec $i \in \{1, 2, \ldots, n\}$).
Pour une observation, chaque forêt calcule les probabilités d'appartenance à chaque classe. Nous sélectionnons comme prédiction finale la prédiction de la forêt la moins incertaine (la forêt décisionnaire), celle pour laquelle la probabilité d'appartenance à la classe est maximale :

$$P_q = \arg\max_i \mathbb{P}(C_i \mid F_k \mid X_q). \tag{5}$$

En procédant de cette façon, la position des forêts dans la cascade n'influe pas directement la prédiction, la forêt qui présente l'incertitude la plus faible est sélectionnée, indépendamment de sa position. Les différents étages étant entraînés sur des ensembles d'apprentissages différents, la forêt la plus adaptée (au sens du degré de certitude exprimé par la probabilité d'appartenance) sera décisionnaire.
Il est aussi intéressant de remarquer qu'ainsi, le processus d'entraînement se fait en cascade, tandis que celui de décision a lieu en parallèle.

4 Expérimentations et résultats

4.1 Jeux de données

Cohn-Kanade étendue (CK+). Publié en 2010 par Lucey et al [21] comme une évolution de CK [17], ce jeu de données est une référence dans le domaine de la reconnaissance d'émotions.
Il se présente sous la forme de 593 séquences d'images. Pour chaque séquence, il a été demandé au sujet de représenter une émotion — parmi la colère, le dégout, la peur, le mépris, la joie, la tristesse et la surprise — en partant d'une expression neutre pour arriver jusqu'à l'apex de l'émotion demandée. On trouve dans ce jeu de données 210 sujets âgés de 18 à 50 ans de genres et d'origines diverses (voir figure 1).
Pour ce jeu de données, nous utiliserons les images de la deuxième moitié de chaque séquence comme exemples de visages présentant l'une des sept émotions citées. En plus de ces émotions, il est possible de sélectionner les images qui

FIGURE 2 – Images provenant de *10k US Adults*

représentent une absence d'émotion (neutre). En effet, les séquences allant du neutre vers l'expressif, il est possible de considérer la première image de chacune de ces séquences comme un exemple de visage neutre.

10k US Adults. Présenté en 2013 par Bainbridge et Olivia [2], ce jeu de données est constitué de 10 168 images de visages détourés dont 2 222 sont labellisés (voir figure 2). Il convient de noter que chaque image est indépendante des autres et présente un sujet différent.
Ce jeu de données a été constitué — en recherchant aléatoirement des images sur un moteur de recherche à partir de noms générés procéduralement — de façon à obtenir des visages qui suivent la distribution de la population états-unienne adulte en matière de genre, d'âge, et d'origine ethnique. Chaque image a ensuite été labellisée par une dizaine de personnes non expertes qui, via un formulaire, votaient pour attribuer les caractéristiques. En ce qui concerne l'émotion, c'est le mode de ces votes qui a été retenu pour attribuer le label parmi six émotions : les mêmes que CK+ auxquelles on soustrait le mépris.
Cela dit, et contrairement à CK+, ce jeu de données présente un très fort déséquilibre de classes : le neutre et la joie représentent à eux seuls plus de 96 % des exemples. Cet état de fait est intéressant, en effet le neutre et la joie semblent être à l'origine des expressions les plus archétypiques alors que les autres émotions sont peut-être plus subtiles. En tout état de cause, ce déséquilibre sera pris en compte lors des expérimentations.

4.2 Analyse de sensibilité aux hyperparamètres

Dans cette sous-section, nous reviendrons sur les différents hyperparamètres de notre système en essayant de détailler leur influence. Pour arriver à cette synthèse, nous avons procédé à un *grid search* pour observer l'évolution du score de précision, calculé sur un ensemble de test avec un processus de validation croisée simple, en faisant varier les hyperparamètres.

Facteur de réduction. Le facteur de réduction r permet de déterminer quelle sera la taille de l'ensemble d'entraînement d'un étage de la cascade par rapport au précédent. Dans la mesure où ce nouvel ensemble d'entraînement sera constitué des exemples mal classifiés et ambigus, on peut considérer que plus r est proche de 0 plus l'étage suivant sera spécialisé sur ces derniers, cela dit le nombre d'exemples à disposition pour l'entraînement d'une nouvelle forêt diminue rapidement d'un étage à l'autre. En fixant une valeur proche de 1, il est possible d'entraîner un nombre plus important d'étages qui seront peu à peu spécialisés sur les exemples proches de la frontière.
Pour les différents jeux de données que nous avons testés, une valeur de r inférieure à 0,5 conduit à une diminution trop importante du nombre d'exemples à disposition et ne permet d'entraîner un nombre suffisant d'étages. La figure 3 montre l'évolution de la précision en fonction de r à partir des résultats obtenus, nous remarquons que la dispersion des scores obtenus semble diminuer quand le facteur de réduction augmente.

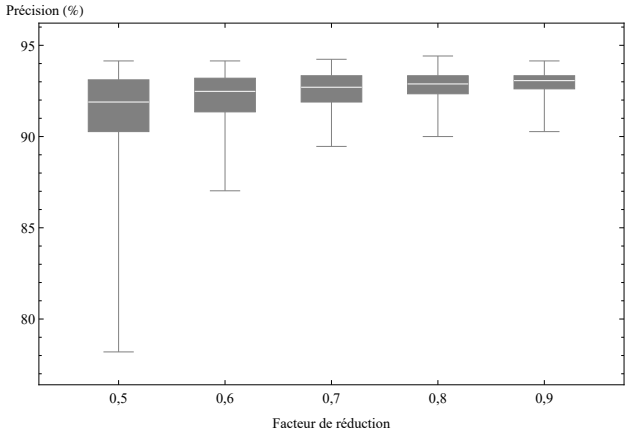

FIGURE 3 – Influence du facteur de réduction r sur la précision

Nombre de forêts aléatoires. Comme nous l'avons noté, le nombre d'étages (c'est-à-dire le nombre de forêts aléatoires) qu'il est possible d'entraîner au maximum pour un jeu donné dépend du facteur de réduction. Il semble rai-

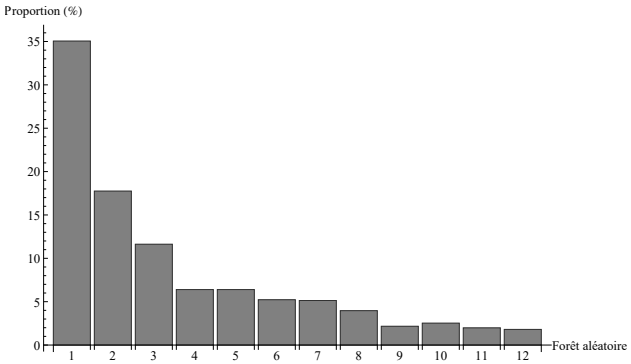

FIGURE 4 – Proportion du caractère décisionnaire pour chaque étage

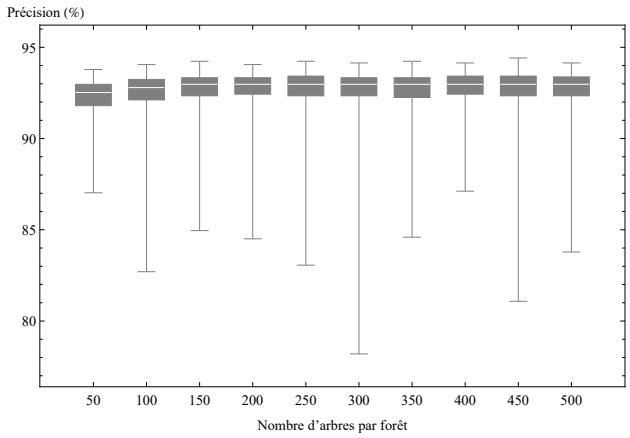

FIGURE 5 – Influence du nombre d'arbres de décisions par forêt sur la précision

sonnable de fixer cet hyperparamètre à cette limite haute : c'est-à-dire que l'on entraîne autant d'étages que possible tant que le nombre d'exemples reste suffisant. En effet, les résultats du *grid search* montrent que l'ajout d'une nouvelle forêt — quand elle a été entraînée avec suffisamment d'exemples — ne vient pas diminuer la précision.

Enfin, puisque la procédure de décision consiste à sélectionner la classe dont la probabilité associée à cette prédiction est la plus élevée, il est également possible d'étudier la répartition de la forêt décisionnaire. En d'autres termes, d'étudier à quelle fréquence chaque forêt s'est trouvée être celle qui a pris la décision. La figure 4 permet de voir que toutes les forêts sont utilisées, mais qu'elles le sont d'autant moins à mesure qu'elles ont été entraînées tard dans cascade. Cette constatation est cohérente avec l'idée selon laquelle plus on progresse dans les étages, plus les classifieurs produits sont spécialisés sur les exemples problématiques (proches de l'hyperplan séparateur).

Nombre d'estimateurs par forêt. Chaque forêt sera composée d'un certain nombre d'estimateurs (arbres de décision). La figure 5 montre l'évolution de la précision en fonction de ce paramètre. Il semblerait que son influence soit limitée ; on peut tenter de l'expliquer en avançant que l'augmentation de ce paramètre permet de diminuer la variance de l'estimation d'une forêt aléatoire unique, mais notre système étant constitué de plusieurs forêts spécialisées, l'utilisation de classifieurs plus faibles (des forêts avec moins d'estimateurs) sera contrebalancée par la sélection de l'estimation la moins incertaine.

Valeurs retenues. À l'issue de cette analyse, les valeurs finalement retenues pour le système présenté sont :

— facteur de réduction : 0,9 ;

— nombre de forêts aléatoires : 12 ;

— nombre d'estimateurs par forêts : 150.

4.3 Évaluation du système

CK+. Nous évaluons notre système sur le jeu de données CK+ avec une procédure de validation croisée de type *10-fold*.

D'abord, on tire aléatoirement les données des ensembles d'entraînement et de test sans tenir compte du sujet. On pourra trouver des images d'un même sujet indifféremment dans les deux ensembles. Ici, on diminue artificiellement le biais d'identité et on mesure les performances du système sur un nombre limité d'individus.

Puis, pour juger des capacités de généralisation du système sur de nouveaux sujets, en effectuant un *leave-many-subjects-out* (LMSO) : les images d'un même sujet seront regroupées soit dans l'ensemble d'entraînement, soit dans celui de test. Ici, on mesure les performances indépendamment du sujet.

Pour être en mesure de nous comparer avec d'autres systèmes, nous avons exclu de cette évaluation le cas de l'absence d'émotion (neutre). La table 1 est tirée de [7], nous y avons ajouté les performances de notre système et celles d'une forêt aléatoire similaire au premier étage de notre cascade.

Système	Précision *10-fold*	Précision *10-fold* LMSO
Happy et al. [15]	94,09 %	-
Lucey et al. [21]	$\geq$ 80 %	-
Song et al. [27]	99,2 %	-
DeXpression [7]	99,6 %	-
Forêt aléatoire	99,20 % ($\sigma = 0{,}8$)	85,95 % ($\sigma = 3{,}7$)
Système présenté	99,60 % ($\sigma = 0{,}4$)	90,76 % ($\sigma = 5{,}8$)

TABLE 1 – Résultats obtenus pour le jeu de données CK+. σ correspond à l'écart-type des scores obtenus sur les différents *folds*.

On constate que les résultats obtenus sont très satisfaisants pour un nombre limité d'individus (colonne 1). Cela confirme l'état de l'art : dans des conditions assez contraintes et pour un nombre limité de sujets, la reconnaissance des émotions faciales est très performante.

En présence du biais d'identité (LMSO), la solution proposée montre tout son intérêt avec une erreur de prédiction diminuée d'un tiers (colonne 2).

Nous observons malgré tout une détérioration des performances, comparativement au système présenté dans [9]. Toutefois, ce dernier est évalué sur les trois dernières images de la séquence qui montre une émotion d'intensité maximale (apex). Nous testons sur la seconde moitié de la séquence et nous évaluons donc des expressions subtiles, voire une absence d'expression.

Notre système est par ailleurs relativement léger : l'entraînement de la cascade est rapide (environ une minute sur un processeur Intel Xeon Skylake cadencé à 2,0 GHz) et le stockage du modèle, c'est-à-dire de tous les étages de la

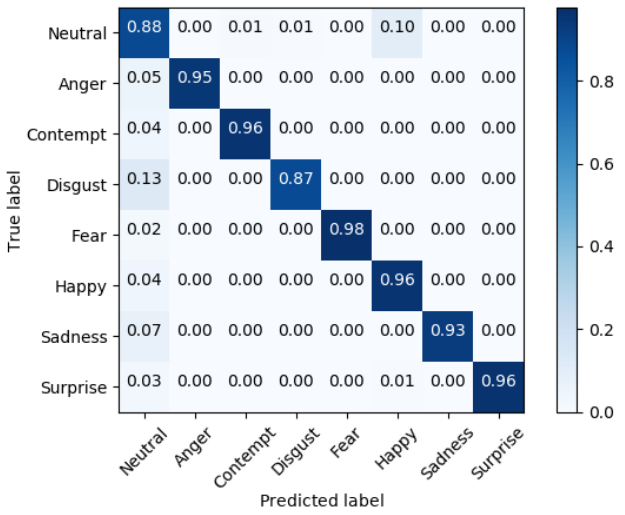

FIGURE 6 – Matrice de confusion obtenue par notre système pour la fusion de CK+ et de 10k, ici avec une procédure de validation croisée simple

cascade, semble tout à fait acceptable ($\approx$ 170 Mio).

CK+ et 10k. Afin d'évaluer notre système dans une situation présentant un déséquilibre en faveur des émotions archétypiques, il est intéressant d'utiliser le jeu de données 10k. Nous utilisons donc les 2 222 images labellisées auxquelles il nous a semblé pertinent d'adjoindre CK+. En effet, nous voulons également vérifier la capacité de notre système à gérer un déséquilibre de données dans les classes tout en étant à même de distinguer les émotions subtiles.
Nous utilisons une procédure de validation croisée de type *10-fold*, sans recourir au principe LMSO puisque les images issues de 10k présentent toutes un sujet distinct et que les performances de CK+ en la matière ont déjà été évaluées. Les résultats obtenus par notre système et par une forêt aléatoire similaire au premier étage de notre cascade sont présentés dans la table 2. La figure 6 permet de voir que la plupart des erreurs commises sont des cas d'expressions émotives (que l'on peut raisonnablement supposer subtiles) classés comme neutre.

Système	Précision *10-fold*
Forêt aléatoire	92,65 % ($\sigma = 0{,}90$)
Système présenté	94,02 % ($\sigma = 0{,}59$)

TABLE 2 – Résultats obtenus pour la fusion de CK+ et de 10k. σ correspond à l'écart-type des scores obtenus sur les différents *folds*.

Nous n'avons pas été en mesure de trouver dans la littérature d'autres systèmes auxquels nous comparer. Il n'en reste pas moins qu'on observe, cette fois encore, que le système présenté permet de diminuer l'erreur (d'environ un cinquième ici) par rapport à notre forêt aléatoire unique.

5 Conclusion et perspectives

Nous avons exploré dans cet article l'application du *boosting* aux forêts aléatoires qui combinent *bagging* et *attribute bagging*. Nous avons montré expérimentalement qu'un apprentissage séquentiel de forêts, en se concentrant sur les exemples proches des hyperplans séparateurs, permettait d'améliorer significativement les performances pour certaines classes de problèmes complexes. Dans le domaine de la reconnaissance d'émotions faciales, nous avons vérifié l'impact du biais d'identité et montré que la solution proposée était efficace.
Il serait maintenant intéressant de comparer cette approche avec celle intégrant directement le *boosting* [22] dans l'apprentissage des forêts afin de réduire le nombre d'estimateurs. Néanmoins, l'impact du nombre d'estimateurs sur les performances semble limité dans la cascade que nous proposons.

6 Remerciements

Ce travail a été réalisé avec le soutien des centres de recherches du CRREF de l'université des Antilles et du LICEF de l'université TÉLUQ, ainsi qu'avec le soutien de l'ANR et du FRQSC dans le cadre du programme conjoint pour les projets France-Québec 2016-2019.

Références

[1] T. R. Almaev and M. F. Valstar, "Local Gabor Binary Patterns from Three Orthogonal Planes for Automatic Facial Expression Recognition," in *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, sept 2013, pp. 356–361.

[2] W. A. Bainbridge, P. Isola, and A. Oliva, "The intrinsic memorability of face photographs," *Journal of Experimental Psychology: General*, vol. 142, no. 4, pp. 1323–1334, 2013.

[3] T. Baltrušaitis, P. Robinson, and L. P. Morency, "OpenFace: An open source facial behavior analysis toolkit," in *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, march 2016, pp. 1–10.

[4] L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, no. 2, pp. 123–140, aug 1996.

[5] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[6] R. Bryll, R. Gutierrez-Osuna, and F. Quek, "Attribute bagging: improving accuracy of classifier ensembles by using random feature subsets," *Pattern Recognition*, vol. 36, no. 6, pp. 1291–1302, 2003.

[7] P. Burkert, F. Trier, M. Z. Afzal, A. Dengel, and M. Liwicki, "DeXpression: Deep Convolutional Neural Network for Expression Recognition." *CoRR*, vol. abs/1509.05371, 2015.

[8] N. Dalal and B. Triggs, "Histograms of Oriented Gradients for Human Detection," in *Proceedings of the 2005 IEEE Computer Society Conference on*

Computer Vision and Pattern Recognition (CVPR'05) - Volume 1 - Volume 01, ser. CVPR '05. Washington, DC, USA: IEEE Computer Society, 2005, pp. 886–893.

[9] A. Dapogny, "A Walk Through Randomness for Face Analysis in Unconstrained Environments," Thèse de doctorat, Université Pierre et Marie Curie, Paris, France, 2016.

[10] P. Ekman, *Basic Emotions*. John Wiley & Sons, Ltd, 2005, pp. 45–60.

[11] P. Ekman, W. V. Freisen, and S. Ancoli, "Facial signs of emotional experience." *Journal of Personality and Social Psychology*, vol. 39, no. 6, pp. 1125–1134, 1980.

[12] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object Detection with Discriminatively Trained Part-Based Models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1627–1645, Sep. 2010.

[13] Y. Freund and R. E. Schapire, "Experiments with a New Boosting Algorithm," in *Proceedings of the Thirteenth International Conference on International Conference on Machine Learning*, ser. ICML'96. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1996, pp. 148–156.

[14] D. Ghimire and J. Lee, "Geometric Feature-Based Facial Expression Recognition in Image Sequences Using Multi-Class AdaBoost and Support Vector Machines," *Sensors*, vol. 13, no. 12, pp. 7714–7734, jun 2013.

[15] S. L. Happy and A. Routray, "Automatic facial expression recognition using features of salient facial patches," *IEEE Transactions on Affective Computing*, vol. 6, no. 1, pp. 1–12, jan 2015.

[16] B. Jiang, M. F. Valstar, and M. Pantic, "Action unit detection using sparse appearance descriptors in space-time video volumes," in *Face and Gesture 2011*, march 2011, pp. 314–321.

[17] T. Kanade, J. F. Cohn, and Y. Tian, "Comprehensive database for facial expression analysis," in *Proceedings Fourth IEEE International Conference on Automatic Face and Gesture Recognition (Cat. No. PR00580)*, 2000, pp. 46–53.

[18] D. E. King, "Dlib-ml: A machine learning toolkit," *Journal of Machine Learning Research*, vol. 10, pp. 1755–1758, 2009.

[19] B. Ko, "A Brief Review of Facial Emotion Recognition Based on Visual Information," *Sensors*, vol. 18, no. 2, p. 401, jan 2018.

[20] S. H. Lee, W. J. Baddar, and Y. M. Ro, "Collaborative Expression Representation Using Peak Expression and Intra Class Variation Face Images for Practical Subject-independent Emotion Recognition in Videos," *Pattern Recogn.*, vol. 54, no. C, pp. 52–67, Jun. 2016.

[21] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*, june 2010, pp. 94–101.

[22] Y. Mishina, M. Tsuchiya, and H. Fujiyoshi, "Boosted random forest," in *2014 International Conference on Computer Vision Theory and Applications (VISAPP)*, vol. 2, jan 2014, pp. 594–598.

[23] A. Mollahosseini, D. Chan, and M. H. Mahoor, "Going deeper in facial expression recognition using deep neural networks," in *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, march 2016, pp. 1–10.

[24] J. A. Russell, "A circumplex model of affect." *Journal of Personality and Social Psychology*, vol. 39, no. 6, pp. 1161–1178, 1980.

[25] E. Sariyanidi, H. Gunes, and A. Cavallaro, "Automatic Analysis of Facial Affect: A Survey of Registration, Representation, and Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 6, pp. 1113–1133, june 2015.

[26] T. Senechal, V. Rapp, H. Salam, R. Seguier, K. Bailly, and L. Prevost, "Facial Action Recognition Combining Heterogeneous Features via Multikernel Learning," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 42, no. 4, pp. 993–1005, aug 2012.

[27] I. Song, H. J. Kim, and P. B. Jeon, "Deep learning for real-time robust facial expression recognition on a smartphone," in *2014 IEEE International Conference on Consumer Electronics (ICCE)*, jan 2014, pp. 564–567.

[28] V. N. Vapnik, *The Nature of Statistical Learning Theory*. New York, NY, USA: Springer-Verlag New York, Inc., 1995.