Projet EFIGI : Classification automatique de galaxies

Anthony BAILLARD

Rapport de stage
Master IAD 2ème année
Spécialité *Fouille de données*

Résumé

Nous proposons une classification morphologique automatique des images de galaxies. Ce système a été conçu dans le but de gérer des images de résolution variable, un problème courant des observations astronomiques effectuées depuis le sol. La classification des galaxies à partir de leur forme apparente est traditionnellement une tâche difficile, basée sur des critères subtils. Ceci est d'autant plus vrai pour les images de galaxies lointaines qui souffrent particulièrement du "flou" causé par la réponse impulsionnelle instrumentale et atmosphérique ainsi que d'un faible rapport signal sur bruit.

Nous discutons les contraintes propres posées par ce type de classification et présentons notre implémentation logicielle, dont nous décrivons en détail les 3 étapes : prétraitement, mise en forme et réduction dimensionnelle, et enfin classification par apprentissage supervisé.

Nous présentons des résultats préliminaires obtenus à partir d'un sous-échantillon de 774 galaxies du *Principal Galaxies Catalog* (PGC), imagées dans le domaine visible par le *Sloan Digital Sky Survey* (SDSS). Les performances du classificateur automatique sont comparées à celles d'une classification visuelle par des astronomes.

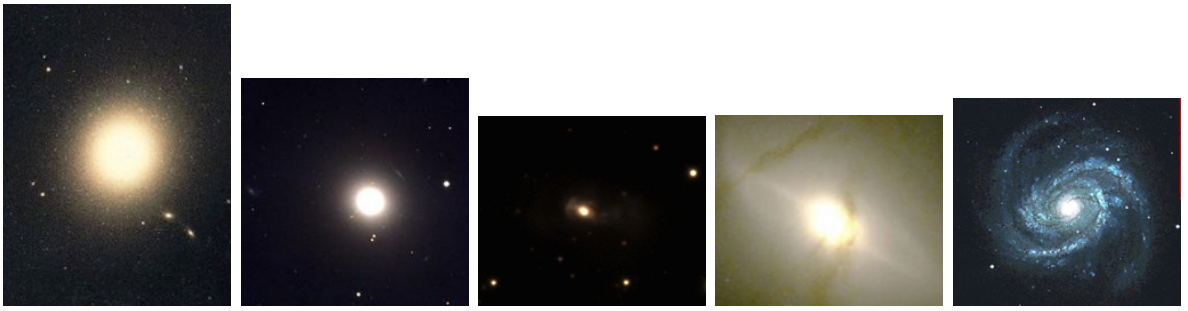
Enfin nous discutons des diverses améliorations envisagées avant la mise à disposition de l'outil à la communauté, notamment sous forme de services.

Mots clés : galaxies : classification - galaxies : type de Hubble - méthodes : analyse de données - méthodes : réseaux de neurones - méthodes : ACP

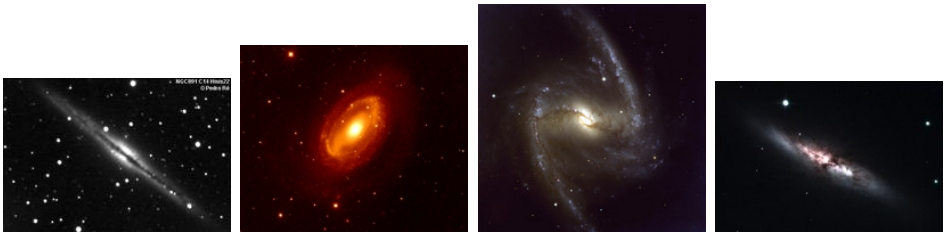
1 Introduction

Ce stage s'inscrit dans le cadre du projet ACI-MDA (Action Incitative Masse de Données en Astronomie) nommé EFIGI[1] (Extraction des Formes Idéalisées de Galaxies en Imagerie). Le but du projet est de fournir un classifieur totalement automatique et autonome d'images variées de galaxies. Le travail réalisé durant le stage constitue une première étape exploratoire visant à initier des pistes de recherche mais également à poser

[1] http ://www.efigi.org



(a) Elliptique (b) Elliptique (c) Lenticulaire (d) Lenticulaire (e) Spirale de face



(f) Spirale par la tranche (g) Spirale (h) Spirale barree (i) Irreguliere

FIG. 1 – Exemples de galaxies.

une base logicielle. Etant données les quantités d'images et les contraintes de calcul inhérentes, le développement de tous les outils décrits s'est effectué dans le langage de programmation $C++$.

La classification morphologique des galaxies est un centre d'intérêt de l'astronomie extragalactique depuis de nombreuses années car l'analyse de la morphologie des galaxies donne des informations précieuses pour comprendre les processus d'évolution de ces objets et, par extension, de l'univers. Les galaxies sont des ensembles auto-gravitants en trois dimensions composés principalement d'étoiles, de gaz et de poussière, mais aussi de matière noire, invisible. Ces constituants sont répartis selon des proportions et des organisations variables d'une galaxie à une autre. On distingue des structures communes dans toutes les galaxies, une composante sphéroïdale qui est la partie centrale ; et un disque, la partie extérieure, dans laquelle on peut trouver des structures cohérentes ou non, traduisant des ondes de densité au sein de la galaxie. Le bulbe est la composante sphéroïdale des galaxies spirales, il a un profil dit "de de Vaucouleurs", c'est-à-dire qu'il suit une loi radiale en $\exp(-r^{1/4})$. Les composantes sphéroïdales sont généralement constituées d'étoiles agées, plus rouges que notre Soleil et pauvres en gaz et en poussière. Le disque est fortement applati et son profil de brillance radial est exponentiellement décroissant. La population stellaire des disques est jeune, souvent plus bleue que notre Soleil et riche en gaz et en poussière (sauf pour les galaxies lenticulaires). Le rapport entre la composante sphéroïdale (appelé bulbe pour les galaxies ayant un disque) et le disque peut varier de facon importante, donnant un panel de profils très différents. Observées depuis la Terre, elles sont semi-transparentes dans le spectre visible et se présentent sous différents angles de position et d'inclinaison ce qui contribue à la variation de leur brillance de surface.

Pour illustrer cette grande diversité, la figure 1 montre des images en couleur de quelques galaxies typiques de l'univers local. Les galaxies elliptiques présentent une composante sphéroïdale proéminente et un disque inexistant. Les lenticulaires quant à elles possèdent un bulbe important mais également un disque sans structure apparente. Les spi-

rales ont des bulbes moins importants mais présentent des structures en formes de bras courbés au sein du disque. Il existe également des galaxies irrégulières, qui peuvent être par exemple de petits amas stellaires ou le résultat de l'interaction de plusieurs galaxies.

La morphologie des galaxies étant fortement liée à leurs propriétés physiques, il est rapidement apparu nécessaire de déterminer une façon de les distinguer qui ait une signification physique.

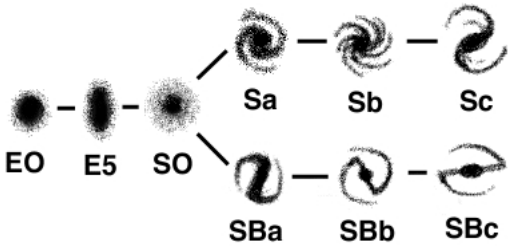


FIG. 2 – Le "diapason" de Hubble

La première classification largement utilisée était le "diapason" de Hubble (*tuning fork*, 1936) mais de nombreuses autres classifications plus objectives et plus exhaustives ont été créées. Par exemple Morgan (1958, 1959) proposa un système basé sur la concentration centrale de lumière. Le système DDO (Van Den Bergh 1960) et sa révision (Van Den Bergh 1976) également basés sur la luminosité s'intéressent davantage à la structure des bras spiraux (longueur, continuité, largeur relative). De Vaucouleurs (1959) complète la classification de Hubble en ajoutant une différenciation entre les galaxies avec ou sans anneau mais aussi les types mixtes entre les galaxies barrées et non barrées. Le système de Hubble révisé (*revised Hubble system*, RHS) offre également une échelle continue au lieu de valeurs entières pour prendre en considération des états intermédiaires. L'échelle de Hubble illustrée par la figure 2 place les galaxies avancées (*late type*) sur les branches de droite en deux groupes, les spirales non barrées en haut et les spirales barrées en bas. Les deux branches se rejoignent pour former le groupe des galaxies lenticulaires (S0), ces dernières ayant un disque mais pas de bras visibles. Enfin, à l'extrême gauche du diagramme, on trouve les galaxies elliptiques, les E0 étant les plus sphériques et les E5 les plus allongées. A noter que les galaxies irrégulières qui ne sont pas représentées sur la figure 2 se trouvent normalement à l'extrême droite, après les galaxies spirales.

En 1936, Hubble pensait que les galaxies précoces (*early type*), les elliptiques, évoluaient progressivement vers une structure plus complexe, passant au stade elliptique puis au stade avancé des spirales. Dès 1976, Van Den Bergh suggérait que c'était plutôt le processus inverse qui se produisait. Aujourd'hui, il s'avère que l'hypothèse de Van Den Bergh est la plus réaliste, cependant la terminologie est restée la même.

Il n'existe pas de système de classification unique accepté par tous les astronomes, c'est pourquoi nous avons choisi de travailler avec les types morphologiques révisés dont la dernière version est proposée par le *Goddard Space Flight Center* de la NASA (voir la table 1). Le plan révisé propose de mettre en relation différents schémas (Hubble, Morgan, Van Den Bergh) pour obtenir un système unique et global. D'après la documentation du *Second Reference Catalogue* (RC2, de Vaucouleurs 1976), le nombre de symboles utilisés pour décrire le type donne une indication sur le degré de résolution et sur la fiabilité de la classification.

Dans ce travail préliminaire, notre méthode ne traite que le type de Hubble converti dans le schéma révisé (avec un petit nombre de symboles). En effet, même si l'échelle de Hubble est subjective et dépend de la classification d'experts, c'est une des classifications les plus compréhensibles pour l'être humain. A l'heure actuelle, il n'existe pas de logiciel automatique universel capable de classifier subtilement les galaxies, si bien que des humains doivent effectuer ce travail à l'œil. Cette opération manuelle introduit systématiquement des biais humains propres à chaque expert comme nous allons le vérifier, d'où l'intérêt de confier cette tâche à une machine plus objective.

Indépendamment du système de classification choisi, l'étude morphologique des galaxies est un vaste champ de recherche car elle permet de contraindre les scénarii de formation des galaxies. De nombreuses caractéristiques morphologiques correspondant à des propriétés physiques particulières qui interviennent dans les processus qui aboutissent à la formation des différentes galaxies. Le rapport bulbe/disque et le profil de brillance de surface sont souvent utilisés comme des paramètres discriminants le long de la séquence de Hubble. Mais d'autres caractéristiques que la forme globale d'une galaxie, comme la presénce d'anneaux, de bandes de poussière ou bien la couleur sont intéressantes pour comprendre la façon dont les galaxies évoluent. Buta et Combes [4] proposent une revue de ces caractéristiques. La morphologie automatique repose sur différentes techniques. Le coefficient de Gini a été introduit par Abraham, Van Den Bergh et Nair [1] pour fournir une mesure quantitative de l'inégalité de la distribution de lumière des galaxies locales. Ils créent un espace à trois dimensions défini par le coefficient de Gini, la concentration centrale et la brillance de surface moyenne afin de définir toutes les galaxies proches sur un plan. Goderya et al. [8] proposent une classification basée sur les caractéristiques de forme globale afin de distinguer les galaxies elliptiques, spirales et spirales barrées. Des méthodes récentes sont elles basées sur des décompositions de Fourier, en ondelettes, [17] ou en "shapelettes" elliptiques [18], [12]. Certaines recherches s'intéressent à des problèmes spécifiques, par exemple Naim tente d'identifier les galaxies singulières (*peculiar*) [15]. Certains programmes d'observation produisant des images multi-bandes, il est également intéressant d'utiliser la couleur comme une nouvelle information, c'est ce que proposent Kelly et McKay [12].

Malheureusement, ces méthodes sont intrinsèquement sensibles aux conditions d'observation, car elle définissent des modèles dépendant d'un contexte précis en terme de résolution, de longueur d'onde et de rapport signal sur bruit. Notre démarche a été guidée par le besoin de gérer ce problème et d'intégrer la réponse impulsionnelle des images dans le processus pour prendre en considération des conditions d'observation quelconques. Notre idée est donc de créer une base analytique de fonctions capable de classer des galaxies selon le RHS mais également de détecter la présence de caractéristiques particulières comme les bandes de poussière. Une décomposition linéaire sur une telle base constitue une réduction de dimensionnalité pouvant intégrer une résolution variable.

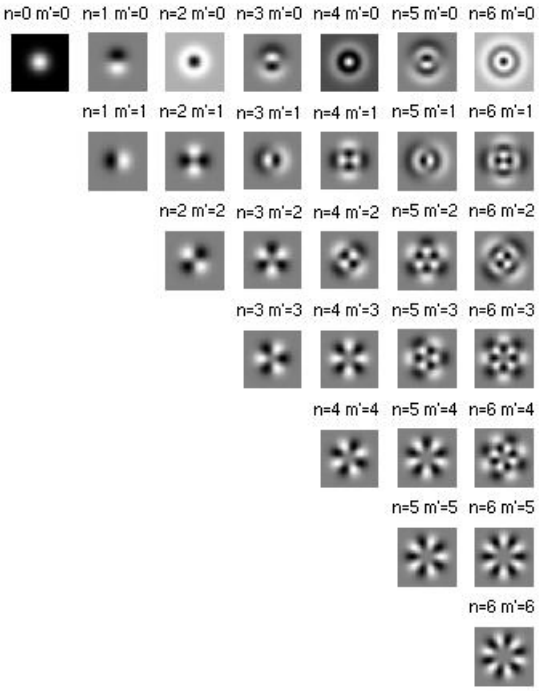


FIG. 3 – Les 26 premiers vecteurs de base de shapelettes polaires non redondants

Une base de vecteurs appropriée a été proposée indépendamment par Refregier [18] et Jarvis [2]. Les premières fonctions de cette base sont illustrées sur la figure 3. Ces fonctions ne sont autres que les vecteurs propres de l'oscillateur harmonique quantique à deux dimensions. Malheureusement, comme nous allons le voir, cette base ne s'avère guère plus efficace que des

bases plus génériques comme une base de cosinus (DCT) ou de Fourier pour représenter les images de galaxies.

Nous présentons dans une première partie notre méthode, en séparant les trois étapes principales, le "nettoyage", la réduction de dimensionnalité et la classification. Dans une seconde partie, nous décrivons les données exploitées durant le stage et dans une troisième partie, nous analysons les résultats obtenus.

2 Méthodologie

Notre méthode peut être globalement divisée en trois étapes consécutives :

1. Nettoyage des images ;
2. Réduction de dimensionnalité ;
   - Calcul d'invariants ;
   - Analyse en Composantes Principales ;
3. Classification supervisée ;

Disposant d'images de galaxies et de leurs types de Hubble associés, nous avons donc choisi d'assigner la tâche de classification à un perceptron multi-couches (PMC) avec une couche cachée, la couche de sortie ayant pour unique sortie le type de Hubble. Ce choix est motivé par deux raisons principales. Tout d'abord la subtilité des critères discriminants. En effet, les perceptrons multi-couches présentent l'intérêt de déterminer automatiquement les critères de segmentation de classes qui sont fixés par des seuils de décision dans d'autres classifieurs statistiques. De plus, ils possèdent de bonnes capacités de généralisation et constituent des estimateurs du classifieur bayésien optimal [9]. En second lieu, la nécessité de traiter plusieurs dizaines d'images par seconde nous a poussé à choisir ce modèle, car les perceptrons présentent l'avantage, une fois entrainés, de calculer rapidement les valeurs de classification.

Pour parvenir à classer les galaxies efficacement grâce à un apprentissage supervisé par réseau de neurones, il est préférable d'extraire de nos données un petit nombre d'attributs. Cela afin de réduire le nombre d'entrée du système d'apprentissage et par conséquent le nombre de neurones et de connexion interneuronales. La taille réduite de l'échantillon d'apprentissage est aussi une contrainte. Il s'agit donc de réduire la dimensionnalité de nos données tout en conservant le maximum d'information pertinente pour la classification. Nous avons choisi d'écarter des estimateurs de forme (basés sur l'analyse des contours ou des contrastes par exemple) trop sensibles à la qualité des images. Nous avons préféré nous diriger vers une décomposition linéaire sur une base de fonctions dont les vecteurs peuvent ensuite être convolués par la réponse impulsionnelle. Etant donnés les mauvaises performances des bases déjà étudiées dans la littérature comme les shapelettes, notre choix pour cette première étude s'est reporté vers l'Analyse en Composantes Principales (connue comme la transformée de Karhunen-Loeve en théorie de la communication).

L'ACP est une opération sensible aux composantes non gaussiennes du bruit, c'est-à-dire aux étoiles brillantes, aux galaxies et aux divers défauts optiques qui se superposent aux galaxies à classer. Il apparaît donc important de réduire ce bruit non gaussien en effaçant autant que possible ces contaminations. Pour ce faire, nous appliquons successivement des opérateurs simples issus de la morphologie mathématique : ouverture d'aire, *white-top hat*, étiquettage, dilatation.

L'ACP construit une base telle que ses axes maximisent la variance de l'échantillon (maxima et minima) et que ses vecteurs soient décorrélés. Par conséquent, les galaxies se présentant sous toutes les combinaisons d'angle de position et d'inclinaison, la base créée s'appuiera sur des composantes à même de représenter la façon dont celles-ci varient en taille, en angle, en position, autant de paramètres qui ne sont pas du tout pertinents pour notre classification. Nous voulons que notre base trouve les variances maximales uniquement sur des informations structurales, c'est pourquoi nous effectuons préalablement un calcul d'invariants. Par calcul d'invariants,

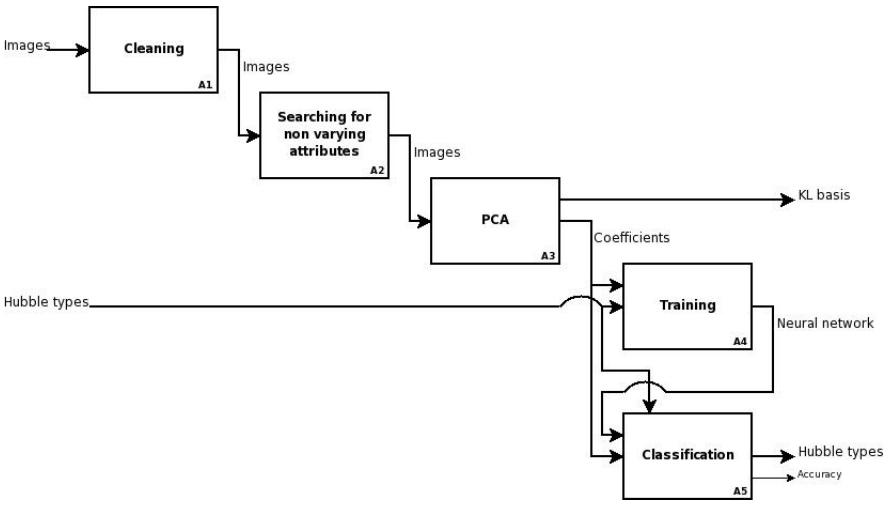


FIG. 4 – Processus global (formalisme SADT)

nous entendons ici un processus de remise en forme des images permettant d'obtenir des vecteurs ne dépendant que de la forme des galaxies et non de leur présentation. En effet transformer les images pour que des caractéristiques comme la direction de l'axe principal, l'échelle de la galaxie, la position du centre ou la position de la bande de poussière soient similaires d'une image à une autre, améliorera l'efficacité de l'ACP.

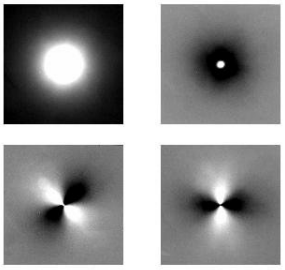


FIG. 5 – Les quatre premières composantes d'une base de KL obtenue sant prétraitement

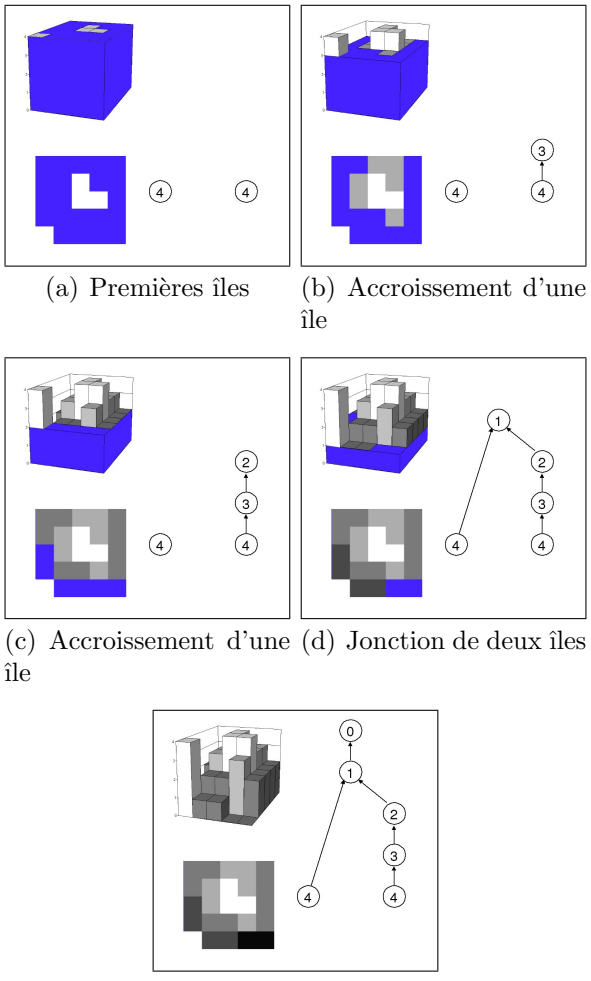
L'ACP et la classification automatique sont les deux étapes principales de notre méthode mais le prétraitement est essentiel pour obtenir de bons résultats. En effet, escamoter l'étape de calcul d'invariants produit une base de projection composée d'harmoniques circulaires inadaptées pour représenter une galaxie. La figure 5 le montre, la composante en haut à gauche étant parfaitement circulaire et les deux composantes du bas etant des dipôles en quadrature de phase.

2.1 Nettoyage des images

Notre premier but est de nettoyer les images de telle façon que les objets parasites soient éliminés. Techniquement, nous voulons effectuer de la segmentation d'images pour retirer certaines composantes. Comme l'expliquent Salembier dans [20] et Heijmans dans [10], une technique efficace pour le traitement d'image est la représentation sous la forme d'un arbre. Salembier explique également comment utiliser des filtres morphologiques sur un tel arbre. La morphologie mathématique est un outil d'extraction de composantes d'image qui est utile pour la représentation et la description, elle a été originellement développée par Matheron et Serra [22] mais les opérateurs d'ouverture ont été définis par Vincent [25] et la granulométrie a été introduite par Breen et Jones [3]. La morphologie peut fournir les limites des objets, leurs squelettes, et d'autres informations de forme. Elle est également utile pour de nombreuses techniques de pré- et de post-traitement, et c'est ce qui nous intéresse.

De facon générale, la plupart des filtres morphologiques sont basés sur de simples opérations d'extension et réduction. Ces filtres

sont particulièrement intéressants parce qu'ils sont faciles à comprendre et rapide à utiliser sur des images binaires ou en niveau de gris. Nous utilisons des opérateurs d'attributs qui conserve la forme afin de ne pas altérer les galaxies.



(a) Premières îles (b) Accroissement d'une île

(c) Accroissement d'une île (d) Jonction de deux îles

(e) Résultat

FIG. 6 – Construction de l'arbre des composantes. La hauteur de l'eau diminue d'un niveau de gris sur chaque image de gauche à droite et de haut en bas. Le niveau maximal en haut à gauche fait apparaitre les maxima. Le niveau minimal en bas laisse apparaitre la totalité de l'image et l'arbre entièrement construit.

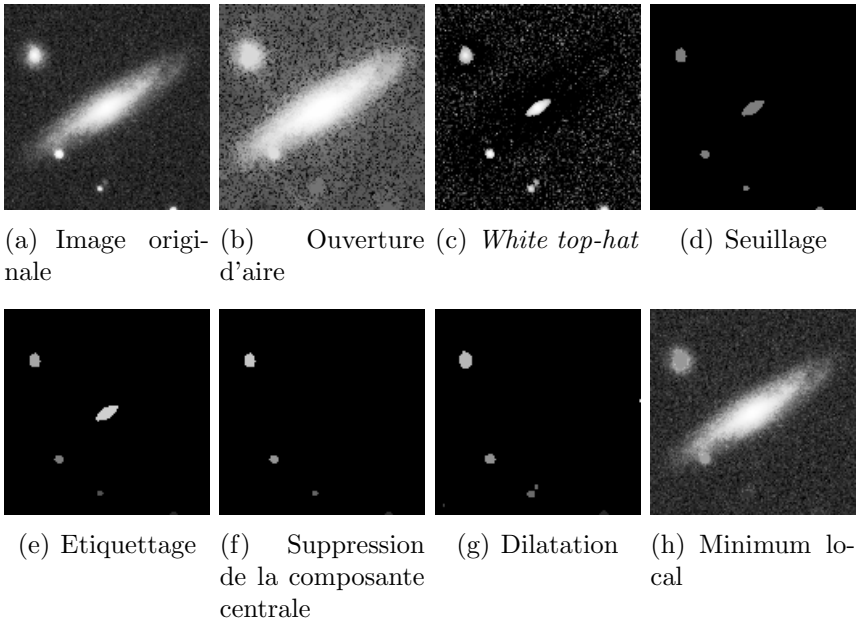
Notre construction du *max tree* est basée sur la méthode proposée par Najman et Couprie [16]. C'est un algorithme quasi-linéaire basé sur le principe d'*union-find* de Tarjan [24]. Construire l'arbre des composantes est facilement compréhensible lorsque l'on effectue une analogie avec la topographie. Nous pouvons considérer une image comme un relief, le niveau de gris d'un point correspondant à son altitude. La surface est submergée, puis le niveau de l'eau diminue. Des îles (maxima) apparaissent. Ces îles sont les feuilles de l'arbre. Avec la diminution du niveau de l'eau, les îles grandissent, construisant les branches de l'arbre. Lorsque plusieurs îles se rejoignent, elles créent une ramification de l'arbre. La racine est le plus faible niveau de gris (le fond) et représente la totalité de l'image. La figure 6 décrit la construction d'un arbre des composantes sur une image simple en utilisant les analogies topographiques.

Pendant la construction de l'arbre, nous pouvons ajouter des informations diverses à utiliser ultérieurement. Actuellement, nos nœuds incluent le plus faible niveau de gris et l'aire globale de la composante correspondante. On peut voir qu'il devient facile d'appliquer une ouverture d'aire par exemple, puisqu'il suffit de couper les branches dont le nœud racine a une aire inférieure à un paramètre donné.

La principale différence avec la structure de l'arbre de Najman est que nous n'utilisons pas de listes de fils mais une architecture fils-gauche frère-droit. Comme nos images sont des images en niveau de gris codées sur des flottants, nous devons d'abord quantifier les valeurs en construisant l'arbre car il est beaucoup plus rapide de calculer des opérateurs morphologiques sur des entiers.

Nous construisons l'arbre des composantes de l'image puis nous appliquons les opérateurs basiques suivants :

1. Ouverture d'aire avec une taille donnée pour repérer les zones claires ;
2. *White top-hat* pour conserver uniquement les pixels de l'ouverture d'aire ;
3. Seuillage pour créer un masque qui contient uniquement les points les plus lumineux ;
4. Etiquettage par 4-connexité pour obtenir des composantes distinctes ;



(a) Image originale (b) Ouverture d'aire (c) *White top-hat* (d) Seuillage

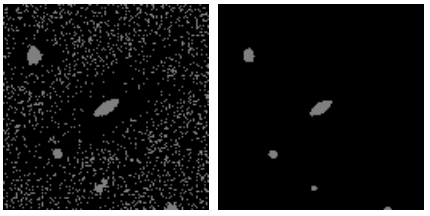
(e) Etiquettage (f) Suppression de la composante centrale (g) Dilatation (h) Minimum local

FIG. 7 – Processus global de nettoyage. L'image en haut à gauche est la galaxie d'origine et celle en bas à droite est le résultat final. Les opérateurs sont appliqués de gauche à droite et de haut en bas.

5. Suppression de la composante centrale correspondant au cœur de la galaxie ;
6. Dilatation par 8-connexite pour récupérer une valeur de remplacement pour chaque composante étiquettée ;
7. Mimimum local sur l'image d'origine pour niveler les objets parasites en fonction des composantes dilatées ;

L'étape 5 est vraiment spécifique aux galaxies. Elle implique que les galaxies soient centrées sur les images, ce que nous obtenons en calculant un barycentre itératif comme l'explique la section suivante. Comme nous ne voulons pas modifier le centre de la galaxie, nous retirons la composante qui lui correspond de l'arbre. La figure 7 illustre le processus sur une image classique.

L'ouverture d'aire et le seuillage sont caractérisés par un paramètre. Evidemment, l'ouverture d'aire demande une taille d'aire (400 pour les images de 128x128 pixels de la figure 7). Si la taille est trop faible, les étoiles brillantes ne seront pas totalement sélectionnées, seul leur cœur sera nivelé. Si la taille est trop grande, différentes composantes peuvent être jointes. Le seuil doit aussi être fixé pour calculer un masque significatif. La figure 8 montre à quel point il est important de choisir une bonne valeur pour le seuil afin de ne pas altérer l'image en traitant des maxima locaux qui font partie d'une galaxie singulière ou simplement du fond ou du bruit. A terme, cette valeur sera déterminée automatiquement à partir du calcul de la variance du fond de ciel.



(a) Seuil = 5 (b) Seuil = 30

FIG. 8 – Résultats de seuillage pour différentes valeurs. La première valeur est trop faible si bien que les opérateurs morphologiques suivants altèreraient l'image.

2.2 Invariants

Le calcul d'invariants est basé sur des calculs simples et bien connus. Notamment les moments du deuxième et troisième ordre qui sont utilisés pour appliquer des transformations linéaires aux images. Soit μ_{20} le moment centré du deuxième ordre sur l'axe des x, μ_{02} le moment centré du deuxième ordre sur l'axe des y, μ_{30} le moment centré du troisième ordre sur l'axe des x, μ_{03} le moment centré du troisième ordre sur l'axe des y et μ_{11} le moment centré croisé du deuxième ordre.

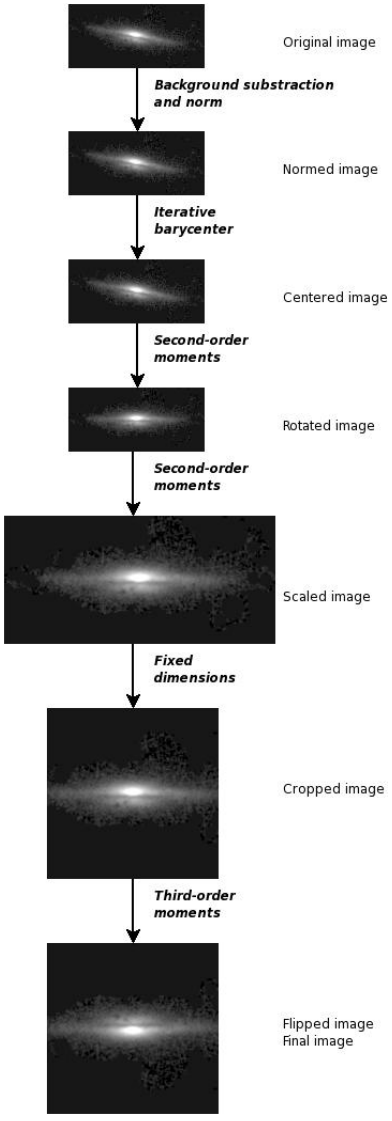


FIG. 9 – Processus global de calcul d'invariants

Tout d'abord, le fond est soustrait des images qui sont ensuite normées en variance pour le bon fonctionnement de l'ACP. De toute façon, même si la brillance de surface est corrélée au type de Hubble, la variation d'intensité dépend des conditions d'observation, les images ne sont donc pas calibrées et l'information de brillance n'est pas exploitable.

Centrer les galaxies est nécessaire pour étudier les symétries (une des informations visuelles les plus pertinentes), pour appliquer les transformations de deuxième et troisième ordre. Ce centrage est également nécessaire pour éviter l'apparition de modes oscillatoires en x et en y durant le calcul de l'ACP. Un calcul itératif du barycentre est effectué pour opérer des translations. Le barycentre est calculé à travers un fenêtrage de l'image par une gaussienne. A chaque étape de l'algorithme, l'image est translatée de la différence entre le centre courant de l'image et le barycentre calculé. L'opération est répétée jusqu'à ce que la différence soit nulle dans les deux directions.

Les galaxies ne sont pas sphériques (exceptées certaines elliptiques compactes et E0) et les spirales vues par la tranche sont très allongées. L'angle de position des galaxies ne représente pas une information morphologique pertinente. Les moments du deuxième ordre sont utilisés pour calculer un angle de position θ du grand axe de chaque image de galaxie :

$$\theta = \frac{1}{2} arctan(\frac{2\mu_{11}}{\mu_{02} - \mu_{20}})$$

La taille visuelle d'une galaxie n'est pas non plus une information pertinente pour la classification. Deux galaxies ayant le même type de Hubble ont quasiment le même aspect quelque soit leur distance (en faisant abstraction du décalage vers le rouge dû à l'effet Doppler) et quelque soit leur taille réelle. On donne donc à toutes les galaxies la même taille apparente. Les moments du second ordre permettent de calculer un facteur d'échelle $\frac{\bar{f}}{f_i}$ tel que :

$$f_i = \sqrt{\mu_{02} + \mu_{20}} \quad pour \quad chaque \quad image \quad i$$

$$\bar{f} = \frac{1}{n} \sum_{i=1}^{n} f_i$$

Au sein des galaxies, on trouve des asymétries systématiques telles que les bandes de poussière. Les moments du troisième ordre sont utilisés pour procéder ou non à des retournements verticaux ou horizontaux des images permettant de toujours représenter les asymétries dans la même direction. Si μ_{30} est négatif, l'image est inversée verticalement. Si μ_{03} est négatif, l'image est inversée horizontalement.

Comme les images seront transformées en vecteurs d'une matrice, elles sont toutes rognées aux mêmes dimensions, généralement 128x128 ou 256x256 pixels.

Les transformations comme les rotations et les recalages d'échelle sont interpolées en utilisant une fenêtre *lanczos 2* (voir [26]) qui préserve au mieux la structure des images à petite échelle. Toutes les mesures des moments employés pour les transformations sont également fenêtrées par une gaussienne dont le σ vaut 1/6ème de la plus petite dimension. Ceci est important afin de ne pas donner trop d'importance au fond et aux bords bruités de l'image mais bien aux régions centrales de l'image.

2.3 ACP

L'ACP est calculée sur ces images afin d'obtenir une base de projection. Seule une partie k des composantes calculées est utilisée pour décomposer les galaxies. Dans notre travail, k est compris entre 10 et 50, comme cela est expliqué dans l'analyse des résultats. Chaque image est représentée par un vecteur de coefficients de la base de Karhunen-Loeve obtenue par l'ACP.

Une ACP calcule une base à travers la transformée de Karhunen-Loeve décrite dans [11]. Dans notre cas, chaque image est transformée en un vecteur et ajoutée comme une nouvelle ligne de la matrice x. Ainsi, pour d images de $w \times h$ pixels, on obtient une matrice $d \times s$ avec $s = w \times h$.

Le modèle de l'ACP est

$$u = Wx$$

où u est le vecteur projeté de dimension m, x le vecteur original de dimension d et W la matrice de passage (la base de Karhunen-Loeve).

On peut montrer que les m vecteurs projetés qui maximisent la variance de u, i.e., les axes principaux, sont les vecteurs propres $e_1, \ldots, e_m$ de la matrice de covariance C des données, correspondant aux m valeurs propres non nulles les plus grandes, $\lambda_1, \ldots, \lambda_m$.

La matrice de covariance C est calculée grâce à l'équation :

$$C = \frac{1}{n-1} \sum_{i=1}^{n} (x - \mu)(x - \mu)^T$$

Les valeurs et vecteurs propres sont ensuite obtenus en résolvant le système d'équations :

$$(C - \lambda_i I)e_i = 0, i = 1, \ldots, d$$

Les vecteurs propres sont ensuite classés par valeur propre décroissante et les m premiers sont sélectionnés comme étant les composantes principales. La matrice de projection est alors $W = E^T$, les colonnes de E étant les vecteurs propres.

Chaque image est décomposée comme un vecteur c de coefficients pour chaque composante de la base de KL. Ces coefficients sont obtenus par un simple produit matriciel :

$$c = xW$$

Evidemment, les images reconstruites sont obtenues par le calcul :

$$\hat{x} = cE$$

A noter que l'ACP peut être calculée dans l'espace des données, ici les images, ou dans l'espace des attributs, les pixels. Dans l'espace des données, l'ACP maximise la variance inter-individus alors que dans l'espace des attributs, elle s'intéresse à la corrélation entre les variables. Calculer les composantes principales dans un espace ou l'autre peut être plus rapide et plus facile si leurs dimensions sont très differentes. Dans l'espace des données, la taille de la matrice de covariance C dépend du nombre

d'attributs s alors que dans l'espace des attributs, elle dépend du nombre d'individus d. Il convient donc de choisir dans quel espace calculer l'ACP pour minimiser la taille de la matrice de covariance et donc les temps de calculs. On peut facilement passer d'un espace à l'autre en transposant la matrice x. Concernant la matrice de passage W obtenue dans un espace, il convient de la transposer mais également de normer chaque vecteur propre par sa valeur propre pour obtenir la matrice de passage dans le second espace.

2.4 Classification

Nous utilisons deux sortes de classificateurs différents, le premier séparant simplement les galaxies en type *précoce* et en type *avancé*, le deuxième attribuant un véritable type de Hubble à chaque image. Les deux systèmes ont de nombreux points communs à commencer par leur architecture globale décrite par la figure 10. Ce sont des perceptrons multi-couches avec une couche cachée. Tous les neurones possèdent une sigmoïde comme fonction d'activation. L'apprentissage d'un perceptron s'effectue grâce à un algorithme de rétropropagation de gradient. L'algorithme choisi est *rprop*, proposé en 1993 par Riedmiller [19], un des plus performants pour ce style de problème parmi ceux fonctionnant en mode *batch*.

Les deux réseaux utilisent actuellement le type de Hubble continu. Ils sont entrainés avec en entrée les coefficients obtenus par la projection d'images de galaxies sur les k premières composantes de la base de KL. Le premier système n'a besoin que des deux premières composantes. On compte également deux neurones de sortie, le premier donnant un score pour le type précoce et le deuxième donnant un score pour le type avancé. Le deuxième réseau possède un nombre variable de neurones d'entrée, généralement 30 (voir la section *Résultats*) et un seul neurone de sortie qui donne le type T de Hubble entre -6 et 10 (voir table 1).

Le seul paramètre libre restant est le nombre de neurones sur la couche cachée. Les chercheurs travaillent toujours sur des méthodes permettant de déterminer le nombre optimal de neurones cachés. Donoho et Johnstone [6] proposent un *benchmark* complet basé sur des critères statistiques. Elisseeff et Paugam-Moisy [7] préfèrent une approche analytique pour obtenir des bornes inférieure et supérieure pour un apprentissage exact. Selon leur article, le nombre de nœuds cachés N_H dépend de N_P, le nombre d'exemples ; N_I, le nombre de neurones sur la couche d'entrée et N_S, le nombre de neurones sur la couche de sortie de telle sorte que :

$$\frac{N_P N_S}{N_I + N_S} \leq N_H \leq 2\frac{N_P N_S}{N_I + N_S}$$

Typiquement, pour le deuxième système d'apprentissage, nous avons $N_I = 30$, $N_S = 1$ et $N_P = 2000$ si bien que $64 \leq N_H \leq 128$. Ces limites sont données pour un apprentissage exact sur le jeu d'apprentissage ce qui peut facilement déboucher sur de la suradaptation (*overfitting*).

De nombreux articles s'intéressent à la suradaptation, et différentes solutions sont proposées comme l'entrainement interrompu (*stopped training*) ou la régularisation [21]. Nous préférons la première solution car elle est simple à mettre en place.Il suffit de calculer l'erreur de validation régulièrement pour savoir si le réseau est en passe de suradapter les données d'apprentissage. La régularisation, testée sous Matlab, donne des résultats équivalents.

Nous utilisons le procédé standard de validation croisée. Le PMC est entrainé avec environ 80% du jeu de données et validé avec les 20% restant. Ceci est fait 5 fois, en divisant le jeu de données en 5 parties et en fusionnant 4 d'entre elles, la dernière étant l'échantillon de test. La validation croisée est nécessaire pour l'entrainement interrompu et donne également des informations sur l'homogénéité du jeu de données.

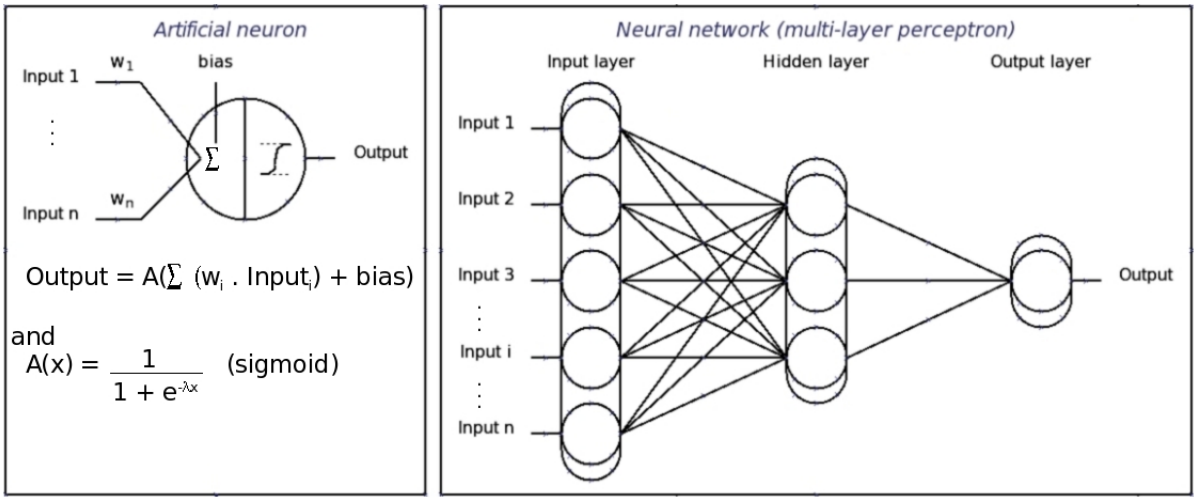


FIG. 10 – Modèle du perceptron multi-couches.

3 Données

3.1 Images

Dans cet article, nous n'utilisons qu'un seul jeu de données composé de 774 galaxies du *Sloan Digital Sky Survey* (SDSS). Les images originales autour des galaxies du catalogue *Principal Galaxies Catalog* (PGC) ont été obtenues directement depuis le serveur SDSS DAS[2] dans les bandes u (ultraviolet), g (bleu-vert), r (orange), i (très proche infrarouge) et z (proche infrarouge) (voir figure 11). Notre choix s'est porté sur ce programme d'observation car il est homogène, récent, de bonne qualité et multi longueur d'ondes. Les images correspondent à des objets du *Principal Galaxies Catalog*. Le choix du PGC est également justifié pour notre étude car les objets de ce catalogue sont limités en diamètre. De cette façon, l'échantillon doit représenter équitablement toutes les classes de galaxies, ce qui n'aurait pas été le cas avec une sélection selon la brillance de surface par exemple.

Ces images ont été nettoyées par le *deblender* du SDSS [14], [13] si bien que nous n'avons pas utilisé notre technique de nettoyage mais seulement les calculs d'invariants. Pendant tout le traitement, nous n'avons exploité que les 3 bandes centrales g, r et i car les galaxies ont des profils similaires. A l'opposé, les bandes u et z sont plus bruitées et présentent des différences, ce qui est bien visible sur la figure 13. En effet, les éléments qui composent les galaxies absorbent différemment la lumière aux longueurs d'ondes utilisées si bien que l'apparence du même objet n'est pas la même selon le filtre.

Finalement, nous avons un jeu de 2322 images. Une des limitations de ce jeu est qu'il ne contient aucune galaxie ayant un type de Hubble de -6, c'est-à-dire des elliptiques compactes.

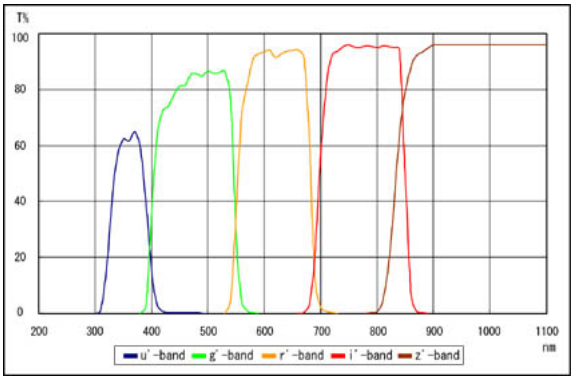


FIG. 11 – Longueurs d'onde des 5 bandes d'observation.

[2] http ://das.sdss.org/DR3-cgi-bin/DAS

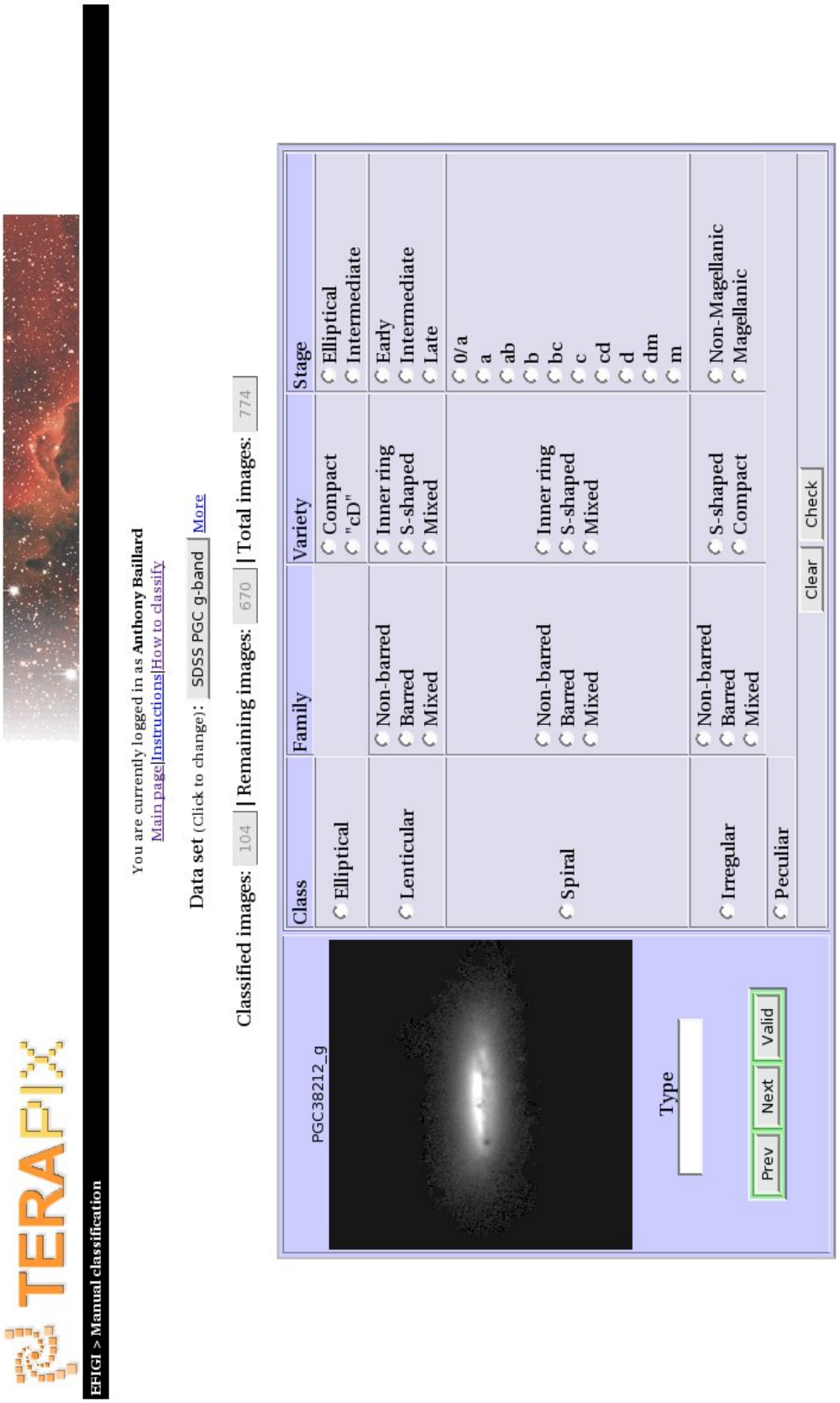
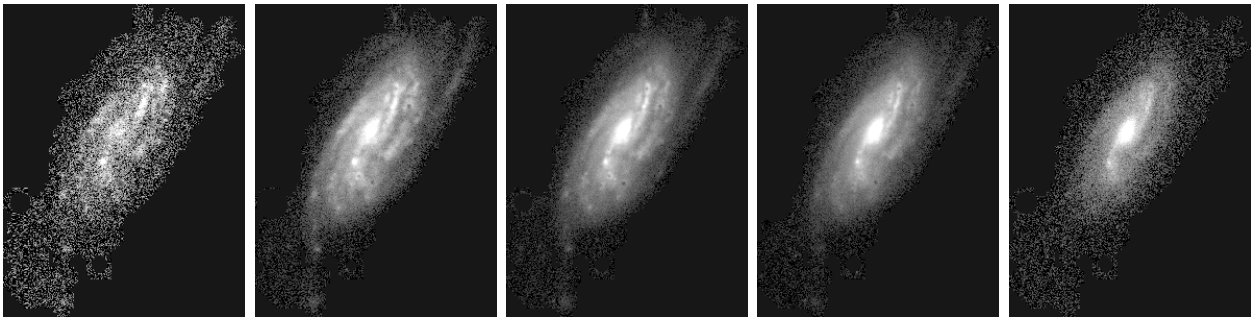


FIG. 12 – Copie d'écran du classifieur manuel.



(a) u (b) g (c) r (d) i (e) z

FIG. 13 – Image de la galaxie PGC70765 dans les 5 bandes du SDSS.

3.2 Classification

Les classifications sont également extraites du *Principal Galaxies Catalog* (PGC). Les types morphologiques et les types de Hubble sont listés pour chaque galaxie sous la forme des types morphologiques révisés (voir la table 1). Nous extrayons seulement quatre colonnes du fichier comme nous l'indiquons ci-dessous : l'identification de la galaxie dans le catalogue PGC, la chaîne de caractères du type révisé et le type de Hubble.

```
PGC   243     PLA.0*. -2.0
PGC   255     .S..9*.  9.3
PGC   281     .S..1?.  1.0
PGC   451     .S..8*.  8.0
PGC   635     .SBR1..  1.0
```

Cet échantillon de galaxies bénéficie d'une des classifications les plus sûres, issue du RC3 (*Third Reference Catalog*), compilé par un groupe d'astronomes experts en morphologie des galaxies. Les données extraites sont donc a priori fiables.

Parallèlement à ce catalogue, nous voulions établir des statistiques sur les biais humains, sur la qualité des classifications des experts et comparer aux performances de notre système. Nous avons donc demandé à des astronomes de classer les 774 galaxies à partir des images en bande *g* uniquement. Cette restriction a été imposée aux utilisateurs afin qu'ils ne disposent pas d'informations inaccessibles au système d'apprentissage, comme la couleur de la galaxie. Les images ont été normées pour la même raison, afin que les experts n'estiment pas la classe d'une galaxie en fonction de sa brillance de surface mais par rapport à ses propriétés morphologiques. Nous avons donc fourni aux astronomes une interface graphique couplée à une base de données proposant les différents classes, familles, variétés et stages prévus par la classification révisée du tableau 1. Seuls ont été omises les précisions apportées pour toutes les classes (*Uncertain, Doubtful, Spindle, Outer ring,* et *Pseudo-outer*). L'interface interactive est accessible à partir d'un navigateur Internet (voir 12) et a permis aux astronomes de classer rapidement et à distance un grand nombre de galaxies sans devoir entrer manuellement la suite de symboles de la classification révisée.

4 Résultats

Il faut être prudent en étudiant ces résultats pour plusieurs raisons. Tout d'abord, la méthode n'est pas complète car nous n'utilisons que le type de Hubble. Ensuite, le jeu de données n'est pas assez grand. Enfin, l'architecture globale du classificateur n'est pas encore définie. Nous essayons simplement de mettre en relief les avantages et les intérêts de notre technique.

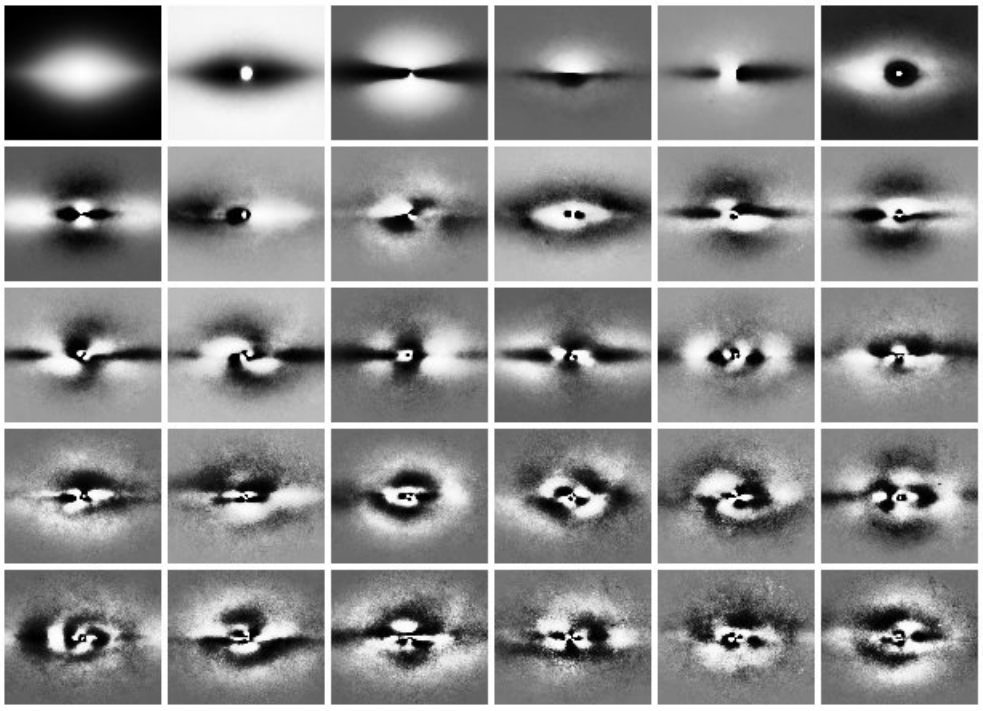


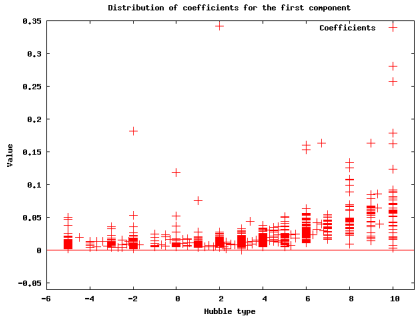
FIG. 14 – Exemple des 30 premières composantes d'une base de KL. Elles sont ordonnées de gauche à droite et de haut en bas.

4.1 Bases de projection

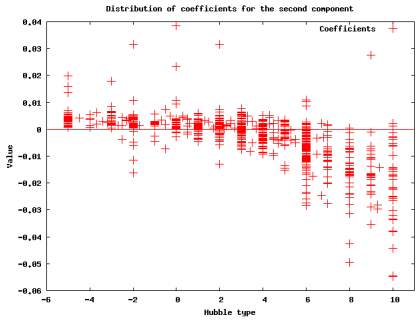
La base d'exemple proposée sur la figure 14 montre des composantes intéressantes et même si l'ACP n'est basée que sur des statistiques du deuxième ordre, on peut tenter d'interpréter leur aspect visuel. Une caractéristique morphologique ne s'appuie généralement pas sur une seule composante mais sur plusieurs si bien que l'on ne peut pas non plus attribuer une composante à une caractéristique de forme. La première composante de la base représente le profil moyen d'une galaxie alors que la seconde est sensible au rapport bulbe/disque. La quatrième conviendrait peut-être à la détection de bandes de poussière. Les composantes comme la huitième mettent en valeur l'importance de l'utilisation des moments du troisième ordre dans le calcul d'invariants car elle est totalement asymétrique. A partir de la 14ème composante, on trouve des structures qui pourraient éventuellement représenter une spirale grossière. Les autres composantes sont plus difficiles à interpréter pour un être humain mais les reconstructions d'image à partir d'une telle base sont plus convaincantes (voir la figure 16).

Il est intéressant d'étudier la distribution de certains coefficients en fonction du type de Hubble. Particulièrement les deux premières composantes de la base de KL de la figure 14. La distribution des coefficients est illustrée par la figure 15.

La deuxième composante semble être un bon paramètre discriminant pour les types précoce et avancé, illustrant la prédominance du rapport bulbe/disque dans la classification de Hubble (Simien et de Vaucouleurs [23]). En effet, la plupart des galaxies dont le type de Hubble est inférieur à 2 ont un coefficient positif alors que la plupart de celles dont le type de Hubble est supérieur à 2 ont un coefficient négatif. La zone difficile se trouve entre 1 et 4. On peut aussi noter que le premier coefficient est toujours positif, ce qui était tout à fait prévisible pour le profil moyen de la galaxie.

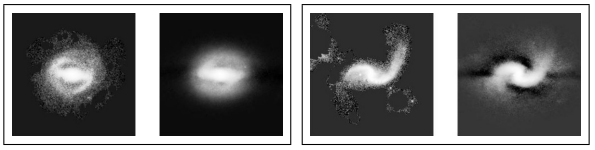


(a) Premiere composante

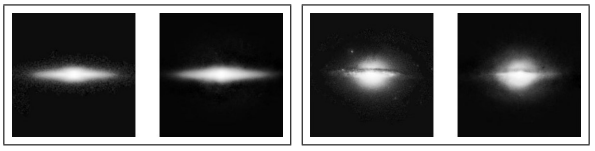


(b) Deuxieme composante

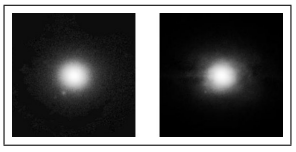
FIG. 15 – Distributions des coefficients en fonction du type de Hubble



(a) Anneau (b) Spirale de face



(c) Spirale par la tranche (d) Bande de poussière



(e) Elliptique

FIG. 16 – Exemple d'images reconstruites. Sur la gauche l'image originale. Sur la droite, l'image reconstruite avec 30 composantes.

Les images reconstruites avec un petit nombre de coefficients sont relativement belles. On peut facilement y reconnaître des galaxies elliptiques, des galaxies spirales, des anneaux et des bandes de poussière par exemple. Il semble plus difficile de représenter des galaxies vues de face dont les bras sont fins et très incurvés.

Comparer nos meilleurs bases aux shapelettes est intéressant. Pour reconstruire une image de galaxie qui puisse être interprétée par un humain, nous avons besoin d'environ 30 composantes alors que une décomposition en shapelettes demande environ 160 composantes, un nombre à peine inférieur à celui nécessaire pour une base de Fourier ou DCT. Voir [18] ou [12] pour des exemples de shapelettes appliquées à la morphologie des galaxies. De plus, nos composantes sont visuellement compréhensibles tout en restant orthogonales.

4.2 Performances de classification

Nous utilisons deux classifieurs différents pour tester la pertinence de la base de Karhunen-Loeve. Le premier système discrimine les galaxies précoces et avancées, c'est-à-dire qu'il sépare les galaxies en fonction de leur type de Hubble, précoce pour un type inférieur à 2 et avancé pour un type supérieur ou égal à 2. Le deuxième système assigne un type T de Hubble à chaque galaxie.

Le premier résultat intéressant est obtenu en n'utilisant que les deux premières composantes. Les résultats sont raisonnables à la vue de la distribution de la figure 15. En effet, un perceptron très simple (2-2-2) obtient 85% d'exactitude sur le jeu d'apprentissage et environ 70% sur le jeu de test. L'entrainement n'est pas interrompu. En comparaison, nos astronomes obtiennent des performances comprises entre 84% et 86%.

Ce système pourrait être une pré-classifieur intéressant pour calculer une nouvelle base de KL avec uniquement les galaxies de type *early* ou *late*. Si nous augmentons la taille du vecteur d'entrée en ajoutant des coefficients des composantes principales suivantes, l'exactitude n'augmente pas pour le jeu de valida-

tion comme l'indique la figure 17. Cela semble indiquer une suradaptation. De plus, la figure met en avant l'importance de la validation croisée car un des trois jeux d'apprentissage est plus efficace sur le jeu de validation. Si nous étudions le comportement du classifieur lorsque le nombre de nœuds d'entrée augmente, il apparaît clairement que les 30 premiers paramètres sont utiles pour déterminer le type de Hubble pour les cas les plus délicats, entre 1 et 4. Mais cela est vrai uniquement pour les jeux d'apprentissage, ce qui prouve que le réseau est surentrainé. Avec des échantillons plus grands et plus exhaustifs, le résultat pourrait être meilleur. Cette figure montre également que 30 coefficients sont suffisants, en utiliser plus n'augmente pas significativement l'exactitude de la classification.

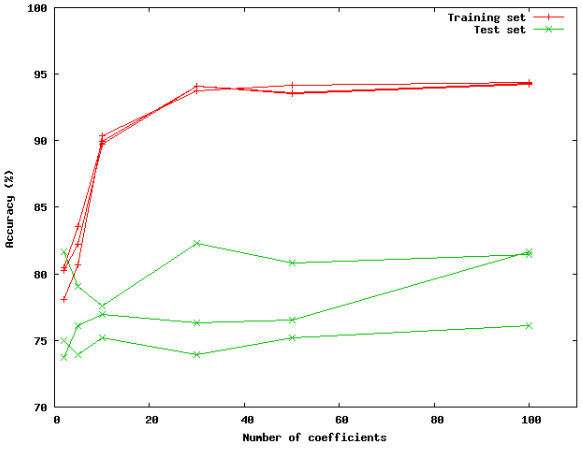
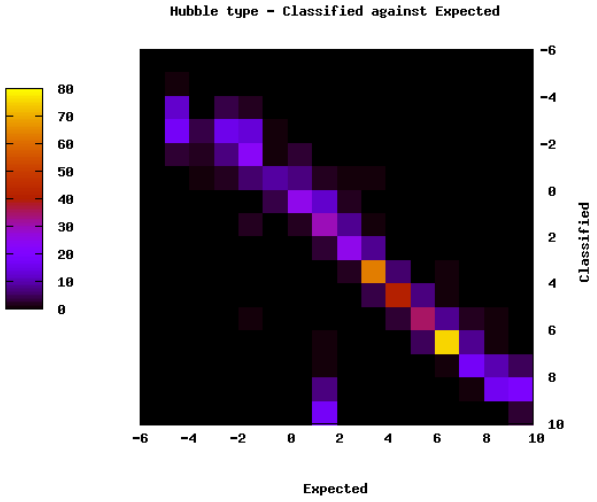
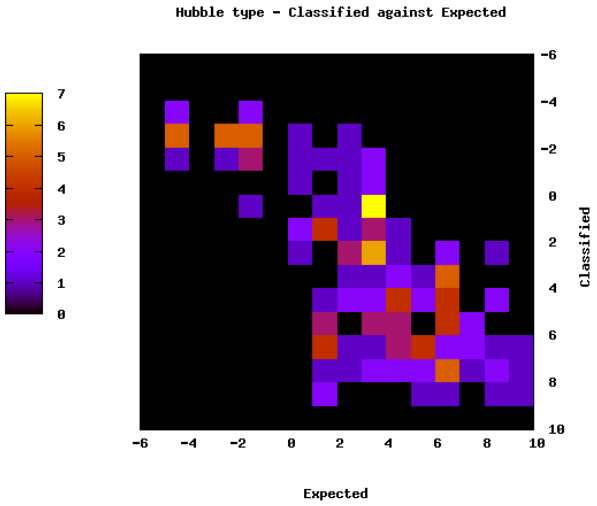


FIG. 17 – Exactitude de la classification *early/late* pour plusieurs nombres de coefficients en entrée. Les courbes rouges correspondent à trois jeux d'apprentissage et les trois courbes vertes aux trois jeux de validation correspondants.

Le deuxième classifieur donne des résultats relativement bons. En utilisant le jeu d'apprentissage qui donnait les meilleurs résultats sur le premier système, nous entrainons un réseau de neurones avec 30 neurones d'entrée, un nombre variable de nœuds cachés et un seul nœud de sortie donnant le type de Hubble. Les classifications obtenues avec ce système sont données à la figure 18. L'exactitude à 2 types de Hubble près est de 89% pour le jeu d'apprentissage et de 82% pour le jeu de test. Les résultats sont encourageants car le jeu de données est un peu trop petit et ne contient pas tous les types de Hubble. Dans tous les cas, l'exactitude du réseau de neurone n'est pas si éloignée de celle d'un classifieur humain (du fait de sa subjectivité).



(a) Jeu d'apprentissage



(b) Jeu de validation

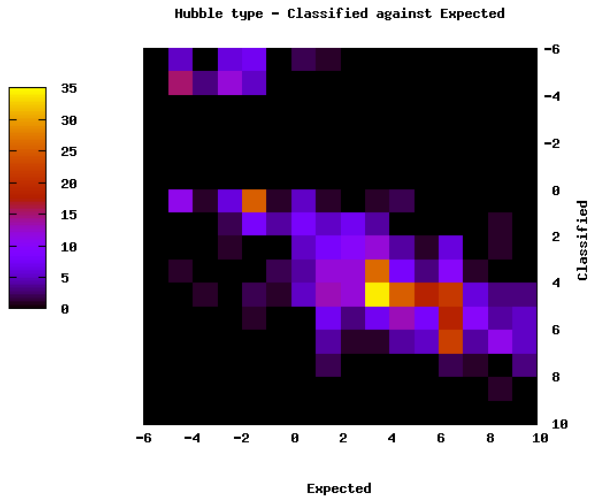
FIG. 18 – Résultats de classification du second système donnant un type T de Hubble.

Nous utilisons l'entrainement interrompu avec ce classifieur car, avec 2000 images, nous

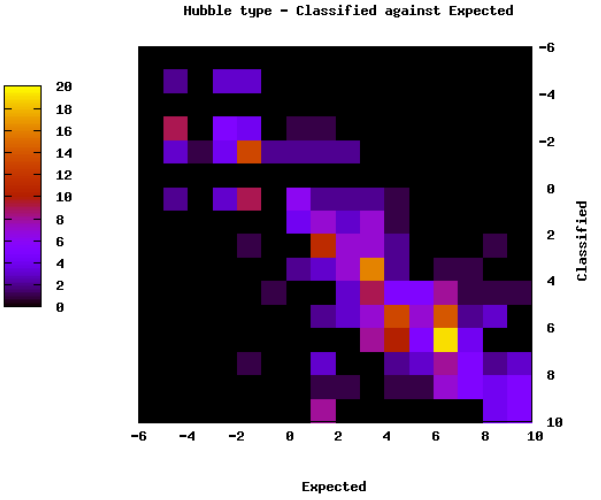
surentrainons facilement le réseau si bien que les résultats sont très bons pour l'échantillon d'apprentissage (0.4% d'erreur de classification à 2 types près) et très mauvais pour l'échantillon de validation (environ 40% d'erreur de classification à 2 types de Hubble près). Nous ne devrions pas rencontrer ce problème avec des bases de données plus grandes. Même avec un petit nombre de neurones cachés, nous obtenons des résultats équivalents car l'interruption de l'entrainement a lieu très tôt durant le processus d'apprentissage. Cela montre que notre jeu de données n'est pas homogène et par conséquent trop petit pour être exhaustif.

En observant la figure 18, on peut étudier la distribution de l'erreur. Les galaxies correctement classifiées se trouvent sur la diagonale principale de l'image puisque le type de Hubble attendu correspond à celui qui a été calculé. On observe clairement 4 pics (les points oranges et jaunes) correspondant aux galaxies les plus courantes dans l'univers proche c'est-à-dire, de la gauche vers la droite, aux elliptiques, aux lenticulaires, aux spirales Sb et aux spirales Scd. Entre ces pics, la classification est plus floue. Ce profil illustre l'hétérogénéité du jeu de données qui contient de nombreuses galaxies pour les types de Hubble -5 (elliptiques), -2 (lenticulaires intermédiaires), 3 (Sb) et 6 (Scd). La capacité de généralisation du réseau est bonne puisque le diagramme de l'échantillon de validation a la même forme globale que celui de l'échantillon d'apprentissage.

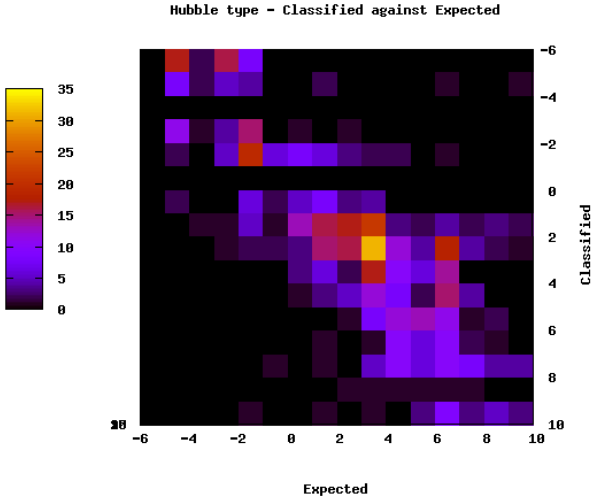
Afin d'estimer plus précisément la qualité de notre classifieur, nous avons demandé à des astronomes de la collaboration EFIGI de trier manuellement les 774 galaxies de notre échantillon à l'aide de l'outil illustré à la figure 12. Trois d'entre eux ont répondu et nous avons ensuite pu comparer leurs résultats à ceux de notre système en traçant les mêmes diagrammes. La figure 19 montre à quel point les résultats peuvent être différents, même pour des experts des galaxies. On peut voir que le premier astronome ne classifie aucune galaxie lenticulaire, ce qui provoque un biais évident dans sa classification.



(a) Premier astronome



(b) Deuxième astronome



(c) Troisième astronome

FIG. 19 – Résultats de classifications faites par des experts.

Un autre détail intéressant, sur le diagramme du deuxième astronome, est de trouver une zone de classification qui diverge par rapport à la classification du PGC aux coordonnées (1,10). On retrouve cette divergence sur le diagramme du réseau de neurones. Cela suggère qu'il y a une dégénérescence d'aspect pour certains types de galaxies.

La figure 20 met en relief les différences de classification obtenues pour deux astronomes, les diagrammes leurs correspondant sur la figure 19 étant le deuxième et le troisième. Ceci illustre bien à quel point la classification est une tâche subtile et difficile.

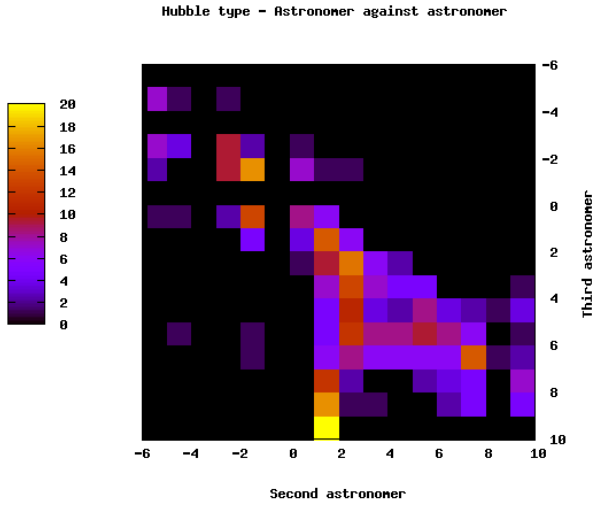


FIG. 20 – Comparaison des classifications de deux astronomes (deuxième et troisième astronomes).

La figure 21 illustre un comportement prévisible du classifieur. En effet, si l'on utilise le réseau de neurones sur les coefficients obtenus pour la décomposition d'images en z, on constate un décalage vers les types plus précoces. En bande z, les bulbes sont plus visibles, plus proéminents en comparaison du disque car ce filtre est plus sensible aux populations agées. Le classifieur les considère donc naturellement comme plus précoces. Un décalage en sens inverse se produirait avec des images en bande u.

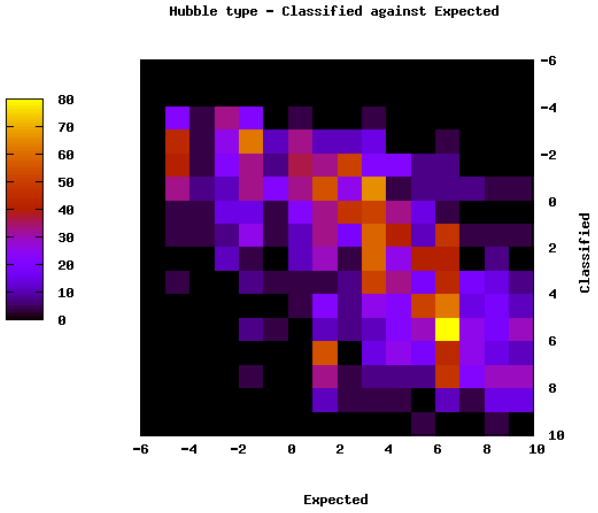


FIG. 21 – Classification des images en bande z.

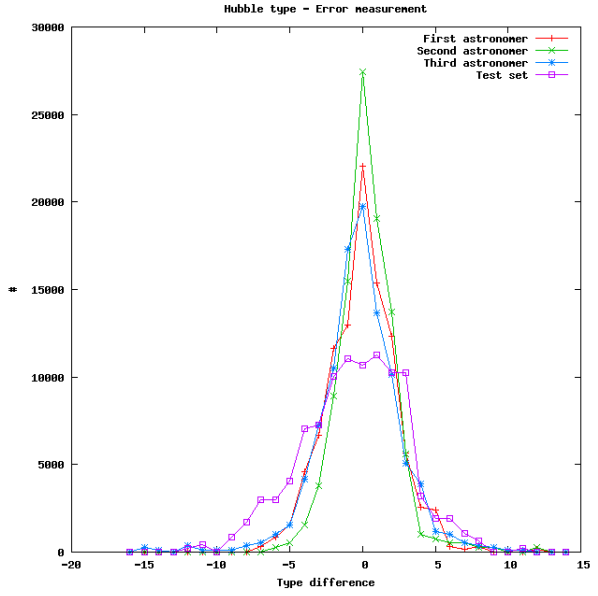


FIG. 22 – Profils d'erreur. La courbe rouge correspond au premier astronome, la verte au deuxième, la bleue au troisième et la violette au système automatique.

Pour estimer les performances de classification indépendemment du type de Hubble, on construit des profils d'erreur en calculant la différence entre le type donné par le classifieur et le type donné par le PGC. Ces profils sont calculés pour les astronomes et pour

notre système sur le jeu de validation. La figure 22 illustre les différences entre les astronomes mais aussi les résultats moins bons du classifieur automatique. Le calcul des demi-largeurs à mi-hauteur sur ces courbes donne les résultats suivants qui confirment les performances inférieures de notre système :

- Premier astronome : 3.2285
- Deuxième astronome : 2.5880
- Troisième astronome : 3.1685
- Système automatique sur le jeu de validation : 4.5630

4.3 Estimations de temps

Opération	Temps pour 2322 images (s)	Temps moyen par image (ms)
Soustraction du fond, norme	16	6.9
Centrage	69.84	30.0
Rotation	183.08	78.8
Mise a l'echelle	368.49	(158.7)
Rognage	24.28	10.5
Inversion	105.83	45.6
Processus global des invariants	758.32	326.6
ACP	38793.49 (11h)	-
Entrainement 2 noeuds d'entree	1.75	-
Entrainement 5 noeuds d'entree	207.30	-
Entrainement 10 noeuds d'entree	1450.40	-
Entrainement 30 noeuds d'entree	823.27u	-
Entrainement 50 noeuds d'entree	1091.60	-
Entrainement 100 noeuds d'entree	2497.13	-
Classification	< 0.23	instantane
Processus global de classification	758.55	326.6

La partie du processus demandant le plus de temps pour la fabrication du classificateur est le calcul de la base de Karhunen-Loeve. Typiquement, nos matrices se composent de 700 à 2000 images 256x256, si bien que nous devons traiter des matrices 2000x65536. Heureusement, nous bénéficions de l'espace dual de l'ACP pour réduire le temps de calcul en cherchant la plus petite matrice de covariance possible. Entrainer les réseaux de neurones demande également du temps, surtout si le nombre de neurones cachés est élevé. La phase de prétraitement est rapide sur chaque image mais comme elle doit être effectuée sur la totalité des données, elle est globalement assez longue. Si la création de la base de KL et l'entrainement des réseaux demandent du temps, la classification d'une galaxie présente l'avantage d'être ensuite très rapide. Nous avons uniquement besoin d'appliquer les calculs d'invariants, calculer la projection sur la base de Karhunen-Loeve et nourrir le réseau de neurones pour obtenir le résultat. Cependant, entrainer un réseau avec un grand nombre de neurones cachés demande du temps pour finalement obtenir des résultats relativement pauvres, nous préférons donc un petit perceptron.

Les estimations de temps sont données par *GNU time 1.7* sur un Athlon AMD 64 3500+ cadencé à 2200MHz avec 2Go de mémoire vive. L'entrainement et la classification correspondent au premier réseau de neurones, c'est-à-dire le classifieur *early/late*. La troisième colonne est obtenue en divisant les résultats de la seconde par le nombre d'images, 2322 soit 774 galaxies dans les bandes g, r et i. Les temps entre parenthèses sont donnés à titre informatif mais ne sont pas réellement significatifs car certaines étapes de calcul dépendent de la totalité des images.

Même s'ils ne sont pas encore optimaux, ces temps permettent de classer les galaxies beaucoup plus rapidement qu'un être humain. En effet, il a fallu aux astronomes environ 16 heures par tranches de 2 heures (à cause de la fatigue et du caractère répétitif de la tâche) pour classer les 774 galaxies alors que le système entraîné demande environ 5 minutes, calcul d'invariants compris.

5 Conclusions et perspectives

A l'issu de ce travail, nous obtenons un système capable de classer les galaxies mais dont les performances restent légèrement insuffisantes. La limitation principale vient vraisemblablement de la réduction de dimensionnalité, car les astronomes sont incapables de classer les images reconstruites avec 30 composantes principales alors que ce sont ces images qui sont utilisées pour l'apprentissage.

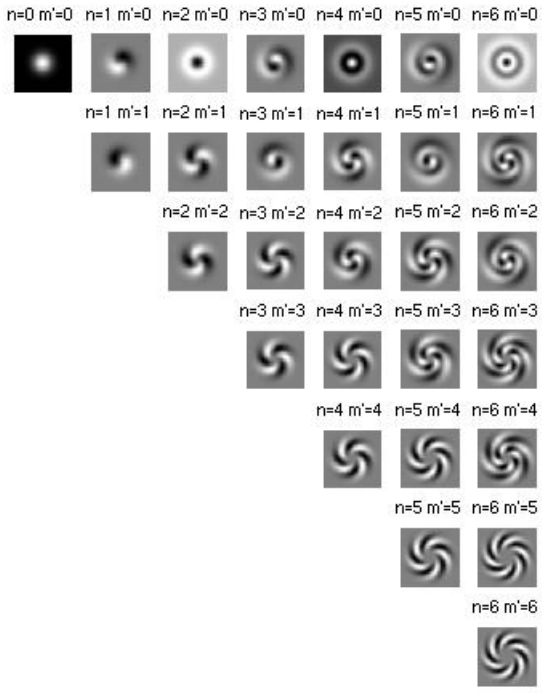


FIG. 23 – Les 26 premiers vecteurs de base de shapelettes polaires non redondants ayant une dépendance radiale de la phase de la composante harmonique circulaire.

La transformée de Karhunen-Loeve fournit un bon point de départ pour la conception d'une future base de vecteurs, possiblement analytique. Les toutes premières composantes sont très intéressantes, visuellement mais également pour la classification. Certaines composantes de notre base semblent particulièrement adaptées pour trouver certaines propriétés indiquées dans le type morphologique révisé, par exemple, la quatrième composante de la figure 14 pourrait facilement indiquer la présence d'une bande de poussière. Un coefficient élevé pour cette composante serait suffisant pour compléter le type révisé par ajout du symbole correspondant. Etudier en profondeur de telles correspondances nous mènerait probablement à une grande base. Le problème est qu'une base comme celle-ci n'est pas orthogonale et nous serions obligés de calculer un fit au lieu d'une projection qui demande moins de temps.

La transformée de KL s'avère insuffisante pour représenter des structures spiralées. Pour cette raison, on pourrait envisager d'utiliser une base de fonctions de shapelettes en ajoutant une dépendance radiale de la phase de la composante harmonique circulaire comme l'illustre la figure 23. Il serait possible de calculer préalablement un invariant sur la position de départ des bras spiraux et un autre sur le rapport entre les deux axes principaux de la galaxie. Enfin, une Analyse en Composantes Indépendantes reste une piste à explorer.

En ce qui concerne le type de Hubble, on pourrait imaginer une méthode basée sur trois systèmes consécutifs. Le premier système nettoie les images, calcule les invariants et la projection sur une base à deux composantes, comme proposé précédemment. Un réseau de neurones classifierait les galaxies en deux classes : précoces et avancées. Ensuite, en fonction des scores de classification, nous utiliserions l'un ou les deux autres systèmes. Si la galaxie est classée en précoce, le "système précoce" spécifierait le type de Hubble (entre -6 et 2). Si la galaxie est classée en avancée, le "système avancé" spécifierait le type de Hubble (entre 2 et 10). Si le réseau de neurones n'est pas catégorique, nous pourrions utiliser les deux systèmes. Les systèmes "précoce" et "avancé" pourraient être bien plus complexes et utiliser les types morphologiques révisés (Table 1). Dans ce cas, il pourrait s'avérer nécessaire de compléter la table et l'interface de classification manuelle. Les types morpho-

logiques révisés faisant appel à des notions de certitude, il serait possible d'utiliser des techniques de logique floue. Il est aussi envisageable d'utiliser d'autres systèmes d'apprentissage comme les machines à support de vecteurs (SVM) ou une classification par prototypes grâce aux k-moyennes.

Dans le cadre de la collaboration EFIGI, il est également envisagé d'utiliser la méthodologie de sélection de caractéristiques proposée par Campedel et Moulines [5] pour les images satellitaires. Un stage a déjà été effectué sur ce sujet par Marie Lienou à l'ENST montrant ainsi la faible efficacité des shapelettes.

La comparaison des performances du système avec celles d'êtres humains doit être nuancée car les astronomes ont été "bridés". S'ils avaient agi de façon autonome, les astronomes auraient utilisé des images en couleur et des outils de visualisation permettant de jouer sur le contraste des images afin de faire ressortir plus facilement les composantes caractéristiques des différents types de galaxies. Il est donc évident que l'utilisateur humain peut obtenir des résultats de classification bien plus satisfaisants, tout comme notre système, en utilisant toutes les informations disponibles.

Un point important pour les prochains travaux est que nous utiliserons un autre jeu de données plus grand d'environ 10 000 galaxies en 5 bandes issues du SDSS. Actuellement, cet ensemble d'images ne peut pas être utilisé pour de l'apprentissage supervisé car la classification disponible contient des erreurs grossières et nous voulons que des experts classent correctement la plupart des images. Nous devrons appliquer les deux étapes de prétraitement sur ces galaxies avant de calculer une nouvelle base de Karhunen-Loeve et d'entrainer des réseaux. De plus, nous ne devons pas oublier que dans cette première étape du projet EFIGI, nous utilisons uniquement des galaxies bien résolues tandis que nous traiterons par la suite des images dégradées de façon plus significative par la réponse impulsionnelle (PSF). Dans ce contexte, les problèmes de dégénérescence des coefficients joueront un rôle essentiel dans le choix de la base. Enfin, grâce à un sous-échantillon d'images de galaxies PGC dans l'ultraviolet moyen provenant du satellite Galex, et d'images infrarouges du relevé 2MASS, nous serons en mesure d'explorer plus d'une décade en longueur d'onde et d'offrir à terme un moyen de classification véritablement pan-chromatique.

Remerciements

Je tiens à remercier monsieur Emmanuel Bertin, mon maître de stage au sein de l'Institut d'Astrophysique de Paris, ainsi que monsieur Yannick Mellier, responsable de l'équipe TERAPIX, pour leur aide précieuse. Dans le cadre du projet EFIGI, je remercie vivement pour leur travail et leur collaboration active, madame Roser Pelló pour les classifications manuelles, madame Marine Campedel pour son travail sur les shapelettes et monsieur Thierry Géraud pour son apport concernant la morphologie mathématique.

Je remercie également tous les autres membres de l'équipe TERAPIX, mesdames Delphine Charbonneau, Mireille Dantel-Fort et Chiara Marmo ; messieurs Laurent Domisse, Jean-Christophe Malapert, Henry Joy McCracken, et Frédéric Magnard, pour leur accueil et leur collaboration.

Mes remerciements s'adressent egalement à monsieur Marcin Detyniecki, mon encadrant universitaire au sein du LIP6 pour son aide et ses suggestions et à monsieur Henry Maitre pour son soutien pour le master IAD.

Classes	Families	Varieties	Stages	T	Type	Code
Ellipticals		Compact		−6	cE	cE...
			Ellipt. (0–6)	−5	E0	.E.0.
			Intermediate	−5	E0–1	.E.0+
		"cD"		−4	E+	.E$^+$..
Lenticulars					S0	.L
	Non-barred				SA0	.LA
	Barred				SB0	.LB
	Mixed				SAB0	.LX
		Inner ring			S(r)0	.L.R
		S-shaped			S(s)0	.L.S
		Mixed			S(rs)0	.L.T
			Early	−3	S0$^-$	.L..−
			Intermediate	−2	S0°	.L..0
			Late	−1	S0$^+$	.L..+
Spirals	Non-barred				SA	.SA
	Barred				SB	.SB
	Mixed				SAB	.SX
		Inner ring			S(r)	.S.R
		S-shaped			S(s)	.S.S
		Mixed			S(rs)	.S.T
			0/a	0	S0/a	.S..0
			a	1	Sa	.S..1
			ab	2	Sab	.S..2
			b	3	Sb	.S..3
			bc	4	Sbc	.S..4
			c	5	Sc	.S..5
			cd	6	Scd	.S..6
			d	7	Sd	.S..7
			dm	8	Sdm	.S..8
			m	9	Sm	.S..9
Irregulars	Non-barred				IA	.IA
	Barred				IB	.IB
	Mixed				IAB	.IX
		S-shaped			I(s)	.I.S
			Non-Magellanic	90	I0	.I.0
			Magellanic	10	Im	.I..9
		Compact		11	cI	cI
Peculiars				99	Pec	.P
Peculiarities			Peculiarity		pec	.....P
(All types)			Uncertain		:	....*
			Doubtful		?	.... ?
			Spindle		sp	...../
			Outer ring		(R)	R.....
			Pseudo-outer R		(R′)	P.....

TAB. 1 – Codage des types morphologiques révisés

Références

[1] R. G. Abraham, S. van den Bergh, and P. Nair. A new approach to galaxy morphology : I. analysis of the sloan digital sky survey early data release. 2003.

[2] G. M. Bernstein and M. Jarvis. Shapes and Shears, Stars and Smears : Optimal Measurements for Weak Lensing. 123 :583–618, February 2002.

[3] E.J. Breen and R. Jones. Attribute openings, thinnings and granulometries. *Computer Vision and Image Understanding*, 64 :377–389, 1996.

[4] R. Buta and F. Combes. Galactic Rings. *Fundamentals of Cosmic Physics*, 17 :95–281, 1996.

[5] Marine Campedel and Eric Moulines. Méthodologie de sélection de caractéristiques pour la classification d'images satellitaires. In *CAP*, pages 107–108, 2005.

[6] D. L. Donoho, I. M. Johnstone, G. Kerkyacharian, and D. Picard. Wavelet shrinkage : Asymptopia ? *J. R. Statist. Soc. B.*, 57(2) :301–337, 1995.

[7] A. Elisseeff and E. Paugam-Moisy. Size of multilayer networks for exact learning : Analytic approach. In *NIPS*, pages 162–168, 1996.

[8] S. Goderya, D. Andreasen, and N. S. Philip. Advances in automated algorithms for morphological classification of galaxies based on shape features. *ASP Conference Series*, 314(1) :617–620, 2004.

[9] J. Hampshire and B. Perlmutter. Equivalence proofs for multilayer perceptron classifiers and the bayesian discriminant function, 1990.

[10] Henk J. A. M. Heijmans. Connected morphological operators for binary images. *Computer Vision and Image Understanding : CVIU*, 73(1) :99–120, 1999.

[11] I. T. Jolliffe. *Principal Component Analysis*. Springer, 1986.

[12] B. C. Kelly and T. A. McKay. Morphological Classification of Galaxies by Shapelet Decomposition in the Sloan Digital Sky Survey. 127 :625–645, February 2004.

[13] Robert Lupton. Sdss image processing i : The deblender. *AJ, submitted*, 2001.

[14] Robert Lupton, James E Gunn, Zeljko Ivezic, Gillian R Knapp, Stephen Kent, and Naoki Yasuda. The sdss imaging pipelines. *ASP CONF.SER.*, 10 :269, 2001.

[15] A. Naim, K. U. Ratnatunga, and R. E. Griffiths. Quantitative morphology of moderate redshift galaxies : How many peculiars are there ? 1996.

[16] L. Najman and M. Couprie. Quasilinear algorithm for the component tree. In *Vision Geometry XII. Edited by Latecki, Longin Jan ; Mount, David M. ; Wu, Angela Y. Proceedings of the SPIE, Volume 5300, pp. 98-107 (2004).*, pages 98–107, April 2004.

[17] S. C. Odewahn, S. H. Cohen, R. A. Windhorst, and N. S. Philip. Automated galaxy morphology : A fourier approach. *apj*, 568 :539–557, apr 2002.

[18] Alexandre Refregier. Shapelets : I. a method for image analysis. *Mon. Not. Roy. Astron. Soc.*, 338 :35, 2003.

[19] M. Riedmiller and H. Braun. A direct adaptive method for faster backpropagation learning : The rprop algorithm, 1993.

[20] P. Salembier and L. Garrido. Binary partition tree as an efficient representation for image processing, segmentation, and information retrieval. 2000.

[21] W. Sarle. Stopped training and other remedies for overfitting, 1995.

[22] J. Serra. *Image Analysis and Mathematical Morphology*. Academic Press, 1982.

[23] F. Simien and G. de Vaucouleurs. Systematics of bulge-to-disk ratios. 302 :564–578, March 1986.

[24] Robert Endre Tarjan. Efficiency of a good but not linear set union algorithm. *J. ACM*, 22(2) :215–225, 1975.

[25] L. Vincent. Morphological area openings and closings for greyscale images. In *Proc. Shape in Picture '92, NATO Workshop*, Driebergen, The Netherlands, September 1992. Springer-Verlag.

[26] George Wolberg. *Digital Image Warping*. IEEE Computer Society Press, Los Alamitos, CA, USA, 1994.