Classification morphologique de galaxies

- Caractérisation de la floculence et des Hotspots -

Rapport de stage

du 15/03/2007 au 15/09/2007

Jean DUMONCEL

Sous la responsabilité de :

M. Henri MAÎTRE : Professeur à l'ENST.
Mme. Marine CAMPEDEL : Enseignant chercheur à l'ENST.

Remerciements

Je remercie avant tout Mr Henri Maître qui m'a donné l'opportunité de faire mon stage dans un domaine particulièrement motivant. Je le remercie également pour les orientations de recherche qu'il m'a suggérées.

Un grand remerciement à Mme Marine Campedel pour ses conseils critiques qui m'ont permis d'avancer tout au long de ce stage.

J'adresse mes remerciements à Mr Emmanuel Bertin, astronome-adjoint à l'Institut d'Astrophysique de Paris, pour m'avoir guidé dans mes recherches. Merci également à Mr Anthony Baillard, en thèse à l'Institut d'Astrophysique de Paris , pour l'aide qu'il m'a apportée.

Je remercie également le personnel de l'ENST et particulièrement celui du laboratoire TSI pour les diverses aides qu'il a pu m'apporter tout au long du stage.

Résumé

Dans un souci de classification des images issues des grands relevés astronomiques, plusieurs méthodes de description des galaxies sont implémentées dans des algorithmes devant servir à la communauté des astronomes et des astrophysiciens. Ce stage a eu pour but de chercher des méthodes d'analyse des regroupements d'étoiles et des spots de formation stellaire. Cette étude est motivée par le fait que ces critères sont liés à l'histoire de la formation des galaxies. Les images étudiées sont issues du catalogue PGC 1.3 fourni par le SDSS, et la longueur d'onde étudiée est la bande verte. Afin de fournir une réponse au problème de la floculence, le système doit être capable d'avoir un modèle satisfaisant de la floculence, puis de la détecter et de la quantifier. Pour cela, on effectue une analyse des textures locales par extraction de caractéristiques dans les images de galaxies. Deux systémes ont alors été développés : le premier est basé sur la classification manuelle des textures afin d'effectuer un apprentissage supervisé de la texture *floculence*, le deuxième permet de classer les textures selon un nombre de classes optimal obtenu de façon automatique, puis de quantifier l'importance de chaque type de texture au sein d'une galaxie afin d'attribuer un taux de floculence à un type de distributions des classes issu de la quantification. Les formations stellaires, ou *hotspots* sont caractérisées par une intense émission de lumière en comparaison avec le reste de la galaxie. Leur détection passe par la recherche de maxima en essayant de s'affranchir des objets lumineux qui se superposent à la galaxie.

Sommaire

Introduction

La classification des galaxies est un problème important dans l'étude des astres observables dans le ciel. En effet, les outils informatiques doivent pouvoir fournir aux astronomes une exploitation rapide et efficace des images de galaxies disponibles en très grande quantité. C'est dans cette optique qu'est né le projet EFIGI (Extraction de Formes Idéalisées de Galaxie en Imagerie) qui est un projet ACI masse de données de trois ans. Il a pour but de fournir des outils aux astronomes leur permettant d'effectuer une classification morphologique suffisamment complète des images de galaxies provenant des caméras électroniques à grand champ. A terme, le classificateur doit pouvoir reproduire le jugement d'un astronome. La première véritable classification des galaxies a été établie par Hubble, décrivant les galaxies en trois catégories principales : elliptiques, spirales et irrégulières, mais elle ne tenait pas compte de toutes les caractéristiques que l'on peut observer dans une galaxie. C'est pourquoi cette classification a été enrichie par la suite afin de pouvoir décrire toutes les galaxies, notamment les galaxies lointaines qui ont subi de nombreuses perturbations pendant leur évolution (zone de formation stellaire, fusion de galaxies, ...).

De nombreuses méthodes ont été developpées ces dernières années pour obtenir une description paramétrique satisfaisante des galaxies, et la méthode la plus largement utilisée est l'extraction du profil de lumière des galaxies. Les principaux problèmes rencontrés sont la définition précise de ce profil ainsi que l'optimisation des temps de calcul, car ces profils dépendent de nombreux paramètres. La problématique est alors de fournir des estimateurs de ces critères, qui soient suffisamment fiables pour se rapprocher avec une faible erreur des estimations faites par les astronomes.

Il reste alors de nombreux autres paramètres[1] (2) qui sont exploités par les astronomes, mais qui ne sont pas, ou peu, implémentés dans les algorithmes. Certaines galaxies contiennent des zones actives de formation stellaire, qui se caractérisent dans les images de la galaxie par des régions plus lumineuses localement : les *hotspots*. Ils sont très utiles pour les astronomes car ils permettent d'étudier la corrélation avec l'âge d'une galaxie mais aussi avec le taux de fusion dans la galaxie. Un deuxième paramètre est le taux de floculence à l'intérieur d'une galaxie. Le phénomene de floculence doit son nom à l'aspect "floconneux" que prend la distribution des poussières et des étoiles dans certains disques de galaxies. Son origine est encore mal comprise, mais pourrait traduire un régime stochastique de propagation de la formation stellaire dans des conditions dynamiques favorables, par opposition à un motif cohérent de bras spiraux dominé par un phénomène d'ondes de densité.

[1] Centre de données astronomiques de Strasbourg

Chapitre 1

Contexte

L'ENST (Ecole Nationale Supérieure des Télécommunications) est un centre de formation d'ingénieurs et qui est également membre de deux groupes scientifiques :

- le GET créé en 1996 à la suite de la privatisation de France Télécom afin de gérer les trois écoles de télécommunication : l'ENST Paris, l'ENST Bretagne et l'INT.
- Paritech qui regroupe 10 grandes écoles d'ingénieurs (ENPC, ENSAE, ENSAM, ENSCP, ENSMP, ENSTA, EP, AgroParisTech et ENST) couvrant ainsi l'ensemble des sciences et des techniques de l'ingénieur à travers plus de 130 laboratoires (3000 enseignants chercheurs et 2000 doctorants).

Le LTCI (Laboratoire de Traitement et Communication de l'Information) regroupe les activités de recherche de l'école des télécommunications et des chercheurs associés. En collaboration avec le département STIC (Sciences et Techniques de l'Information et de la Communication) du CNRS, le laboratoire effectue ses recherches autour des thématiques du signal, des communications numériques et de la reconnaissance des formes. Acteur important de la recherche française, il participe notamment aux ACI du ministère de la recherche dans son domaine de compétence.

1.1 Présentation du département TSI

A l'intérieur du LTCI, le département TSI (Traitement du signal et des images) a pour mission l'enseignement (initial et continu), la recherche (académique et contractuelle), la formation par la recherche dans les domaines du traitement du signal et des images et l'application du traitement du signal et des images dans divers contextes de la société de l'information dont les télécommunications. Il est organisé en cinq groupes qui participent à l'ensemble des missions du département, certains orientés vers la recherche académique, d'autres vers la recherche appliquée, ou l'enseignement :

- le groupe Traitement et Interprétation des Images qui met en oeuvre des schémas complets de traitement, d'analyse et d'interprétation d'images, en particulier de scènes complexes.
- le groupe Traitements Statistiques et Applications aux Communications dont les domaines d'activités sont le signal pour les communications, la séparation de sources, la modélisation statistique pour le signal et l'image.
- le groupe Perception, Apprentissage et Modélisation qui étudie le rôle des facteurs humains dans l'accès aux divers types d'information (parole, image, écrit).
- le groupe Codage qui travaille sur des techniques éprouvées de compression de sources ainsi que sur leur adaptation aux applications de l'audiovisuel et du multimédia.

– le groupe Audio, Acoustique et Ondes qui étudie la physique des ondes dans les deux domaines de l'optique et de l'acoustique.

J'effectue mon stage sous la direction de Henri Maitre et Marine Campedel au sein du groupe Traitement et Interprétation des Images qui participe au projet EFIGI.

1.2 Présentation de TERAPIX

Créé en 1999, TERAPIX[1] (Traitement Elémentaire, Réduction et Analyse des PIXels de megacam) (1) est un centre de traitement des images astronomiques issues des grands détecteurs panoramiques visibles et infra-rouge. Ce centre travaille particulièrement sur le grand relevé Canada-France-Hawaï-Telescope Legacy Survey (CFHTLS). Composée d'une dizaine de personnes, les objectifs de l'équipe de TERAPIX sont :

i). Le développement de logiciels rapides et optimisés pour le traitement des grandes images astronomiques.

ii). La production de données scientifiques calibrées et prêtes à l'exploitation scientifique pour les atronomes et les physiciens, des données telles que des séries d'images astronomiques calibrées, ré-échantillonnées, co-additionnées, des catalogues d'objets, des ensembles de méta-données,... Ces données sont en partie disponibles pour d'autres instituts de recherche.

iii). L'assistance technique aux utilisateurs qui souhaitent bénéficier de l'expérience et des ressources pour le traitement de leurs images.

Outre l'assistance et la production de données, TERAPIX fournit ses logiciels à l'ensemble de la communauté astronomique en essayant de les rendre suffisament génériques afin d'étendre leur utilisation à d'autres types de capteurs et d'images numériques. Terapix est financé par l'INSU (Institut National des Sciences de l'Univers), le PNC (Programme National de Cosmologie), des contrats européens, l'ACI EFIGI et l'IAP.

1.3 Présentation du projet EFIGI

EFIGI (Extraction de Formes Idéalisées de Galaxies en Imagerie) est un projet de 3 ans coordonné par TERAPIX et financé par le Fond National pour la Science. Il a pour but de fournir des outils aux astronomes leur permettant d'effectuer une classification morphologique suffisamment complète des images de galaxies provenant des caméras électroniques à grand champ. A terme, le classificateur doit pouvoir reproduire le jugement d'un astronome. Sept instituts sont partenaires de ce projet : l'IAP (Institut d'Astrophysique de Paris), le LTCI, le LRDE : Laboratoire de Recherche de l'EPITA, le LAM : Laboratoire d'Astronomie de Marseille, le LAOMP : Laboratoire d'Astrophysique de l'Observatoire Midi-Pyrénées, le CRAL : Centre de Recherche Astronomique de Lyon, et le SAp : Service d'Astrophysique du Commissariat à l'Energie Atomique.

Le stage est réalisé au laboratoire TSI en collaboration avec les personnes de l'IAP travaillant pour le projet EFIGI dont Emmanuel Bertin, astronome-adjoint à l'IAP. A l'intérieur du laboratoire TSI, je suis aussi rattaché au COC (Competence Center on Information Extraction and Image Understanding for Earth Observation) qui est une coopération entre le CNES (Centre Nationnal d'Etude Spatiale, France), DLR(Centre Nationnal de Recherche pour l'Aéronautique

[1] http ://terapix.iap.fr

et l'Espace, Allemagne), et l'ENST. Ce groupe travaille notamment sur l'indexation d'images satellitaires, ce que fait aussi l'équipe du projet EFIGI avec des images dont les acquisitions sont non plus dirigées vers la Terre mais vers le ciel.

Chapitre 2

Etude de la morphologie des galaxies

2.1 Les images du catalogue PGC 1.3

Les images sur lesquelles nous travaillons sont centrées sur des galaxies bien résolues du catalogue PGC/RC3 (Principal Galaxy Catalog of the Third Reference Catalog) observées par le Sloan Digital Sky Survey[1], projet visant à établir une carte 3D pour environ un million de galaxies et de quasars en couvrant plus d'un quart du ciel. Pour chaque galaxie du catalogue, les images sont disponibles dans cinq bandes spectrales : u, g, r, i, et z (de l'ultraviolet à l'infrarouge), dont le fond de ciel est directement soustrait et les images ré-échantillonnées, centrées et redimensionnées en 255×255 pixels (cf Fig2.1). Le catalogue contient actuellement plus de 4000 galaxies qui ont été vérifiées par les astronomes. Les images sont au format fits (Flexible Image Transport System), qui est le format standard des données utilisé en astronomie. Ce format, en plus de contenir l'information de l'image, permet de stocker d'autres éléments tel que le spectre.

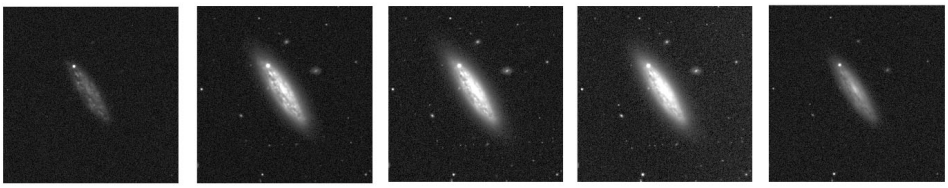


FIG. 2.1 – Galaxie observée dans les bandes u, g, r, i, et z

La base actuellement étiquetée par une dizaine d'astronomes et homogénéisée par Emmanuel Bertin sur des critères de taux de floculence et de *hotspots* (de 0 à 4 en valeur entière avec un intervalle de confiance (cf Tab2.4), le taux 0 correspondant à l'absence du critère, le taux 4 correspondant aux galaxies où le critère est le mieux représenté) comprend 3753 galaxies et a été mise à notre disposition par l'IAP. Chaque galaxie est représentée par une séquence de 5 images (cf Fig2.2) issues de la bande g qui est la plus pertinente pour les critères recherchés :

- l'image ré-échantillonnée en bande g,
- la même image où le fond de ciel a été soustrait,
- le profil de lumière de la galaxie,

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

– la réponse impulsionnelle de l'image extraite à partir d'objets ponctuels isolés tels que les étoiles,
– l'image après soustraction du profil de lumière.

La réponse impulsionnelle est très importante dans la recherche des différents critères dans les images de galaxies. En effet les données bruitées issues des relevés astronomiques nécessitent d'extraire les galaxies et de les redimensionner afin de pouvoir les comparer. Cette étape est effectuée à l'aide du logiciel SWarp[2] qui ré-échantillonne et additionne plusieurs images fits ensemble.

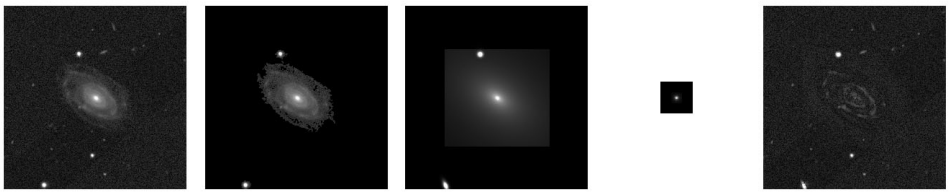


FIG. 2.2 – Banque d'images fournies par l'IAP pour une galaxie

Sur les résidus issus des soustractions des profils de lumière, les bras de spirales, les barres ou les anneaux se retrouvent dans ces images. Les derniers ajustements de profils permettent de soustraire ces éléments qui sont modélisés à l'aide de la dernière version de SExtractor[3] développée à l'IAP (cf Fig2.3). Les galaxies sont modélisées comme étant la somme d'un bulbe, d'un disque, d'une barre et de bras spiraux (19 paramètres au total), l'ensemble étant convolué avec la réponse impulsionnelle locale. Quelques difficultés sont encore présentes au niveau de la modélisation de la partie interne dans les structures spirales, et seuls les objets les plus larges sont modélisés.

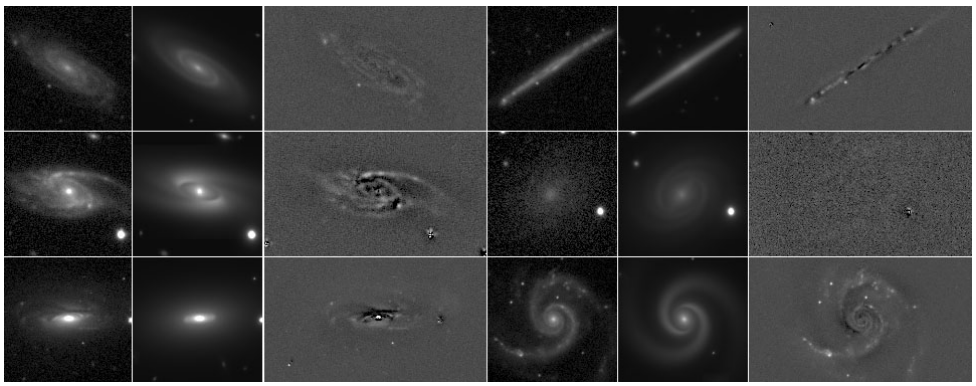


FIG. 2.3 – Exemple de modélisation des galaxies obtenues en faisant la somme du bulbe, du disque, d'une barre et des bras de spirale, convoluée avec la réponse impulsionnelle

La difficulté principale lorsqu'on travaille sur de telles images est la présence d'objets très lumineux qui n'appartiennent pas à la galaxie mais qui peuvent lui être superposés. En

[2] http ://terapix.iap.fr/soft/swarp
[3] http ://terapix.iap.fr/soft/nfigi

effet un observateur humain est capable de faire abstraction de ces objets dans son jugement, à condition que ces objets ne soient pas trop larges. Par contre un traitement automatique de l'image entraînera la prise en compte de ces parasites. Il est donc important d'essayer de prétraiter ce type d'images afin de prévenir une erreur de classification ultérieure. Un logiciel (nfigi) est en cours de développement à l'IAP, qui permet de nettoyer les images. Pour cela, les images en niveau de gris sont symbolisées par des arbres où chaque feuille représente un niveau de gris, donc l'image est assimilée à un relief où la hauteur de chaque pixel représente son niveau de gris. Les maximas apparaissent alors clairement dans ce relief. Différents opérateurs morphologiques sont appliqués à cet arbre :

- une ouverture d'aire avec une taille donnée afin de localiser les aires lumineuses.
- un seuillage pour séparer les endroits les plus lumineux.
- un étiquetage en 4-connexités pour traiter l'ensemble de la zone lumineuse.
- une suppression de la composante centrale qui correspond au bulbe de la galaxie
- une dilatation en 8-connexités pour chercher la valeur de remplacement des zones lumineuses.
- un nettoyage en fonction de la position de la composante

Le logiciel nfigi permet de nettoyer les images efficacement mais altère aussi parfois lourdement les galaxies. Notamment la distinction étoile superposée / *hotspots* étant difficile à modéliser, c'est le type d'information qui peut être perdu lors de ces pré-traitements. C'est la raison pour laquelle les résidus ne peuvent pas encore être calculés sur des images qui ont subi ces pré-traitements.

2.2 Etat de l'art de la floculence et des *hotspots*

Afin de comprendre l'histoire de l'évolution des galaxies, les astronomes ont été amenés à classer les images de galaxies suivant les différentes formes que l'on peut distinguer dans les images . La première classification établie par Hubble (cf Fig2.4) comprenait trois types de galaxies : les galaxies elliptiques constituées d'un sphéroïde dont la luminosité décroît de façon radiale, et composées majoritairement de vieilles étoiles ; les galaxies spirales comprenant un bulbe et un disque où différentes formes telles que des bras, une barre, ou un anneau apparaissent et sont constituées d'étoiles jeunes et de gaz ce qui favorise la formation des étoiles ; et les galaxies lenticulaires dont la forme se rapproche des galaxies spirales mais elles ne possèdent pas de bras et la proportion du bulbe est beaucoup plus importante. Quelques années après sa première classification, Hubble a rajouté une quatrième catégorie : les galaxies irrégulières qui n'ont pas de formes définies et qui comprend par conséquent toutes les galaxies qui ne peuvent pas être classées dans l'une des trois premières catégories. La majorité des galaxies sont des galaxies spirales, c'est une statistique qui se retrouve dans le catalogue de nos galaxies (cf Tab2.1).

Types de galaxies	Nombre de galaxies
Elliptiques	225
Lenticulaires	451
Spirales	2668
Irrégulières	309

TAB. 2.1 – Nombre de galaxies suivant leur type dans notre base

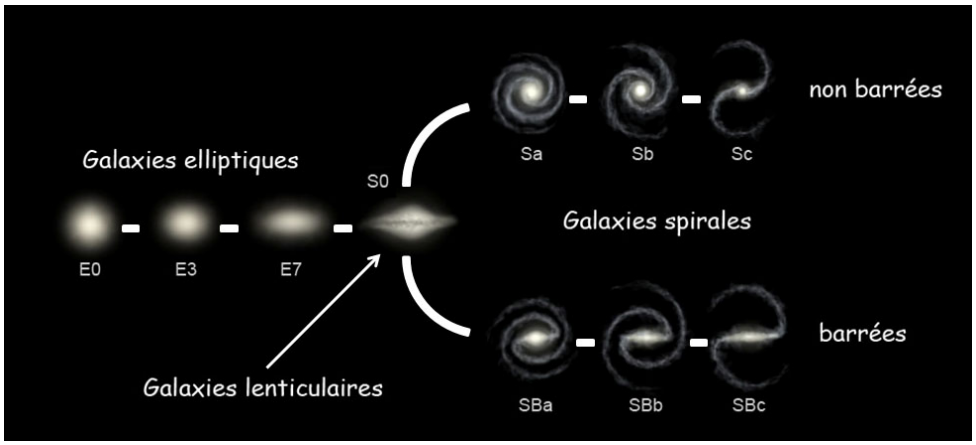


FIG. 2.4 – Classification des galaxies établie par Hubble

La classification de Hubble, et celles qui ont suivi, tiennent compte de critères morphologiques tels que la présence de spirales, de bras, ou de zones d'activité à l'intérieur des galaxies. Certaines galaxies sont constituées de zones de formation stellaire, ce qui se caractérise dans les images de la galaxie par des régions plus lumineuses localement : les *hotspots* (cf Fig2.5). Ils sont très utiles pour les astronomes car ils permettent d'étudier la corrélation avec l'âge d'une galaxie mais aussi avec le taux de fusion dans la galaxie. Un deuxième paramètre est le taux de floculence (cf Fig2.5) à l'intérieur d'une galaxie. La floculence est un effet floconneux qui s'observe dans certaines galaxies autour de regroupement d'étoiles, particulièrement autour de certains bras de galaxie.

Agrandir les images (Annexes)

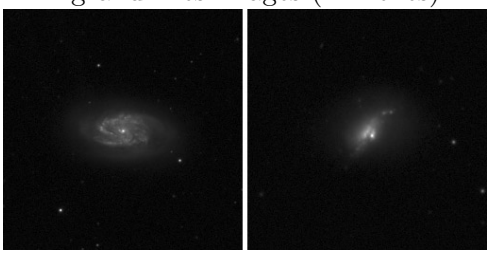


FIG. 2.5 – galaxies présentant un taux élevé de floculence (à gauche) et un taux élevé de hotspots (à droite)

Certains travaux ont tenté d'évaluer des paramètres d'irrégularité afin de décrire toutes les galaxies. Une méthode statistique a été développée dans (3) pour déterminer la distribution de la luminosité à l'intérieur de la galaxie. Ils ont utilisé les coefficients de Gini qui permettent de savoir avec quelle inégalité la lumière d'une galaxie est répartie parmi ses pixels. Cet outil est indépendant de la géométrie des galaxies et fournit un complément à la concentration pour les galaxies proches. Il fournit aussi une nouvelle approche pour l'étude des galaxies lointaines qui n'ont pas de formes prédéfinies dans la classification de Hubble. Dans (4), les auteurs présentent un ensemble de paramètres (C A S : concentration, asymétrie, clumpiness) qui devraient permettre de décrire l'évolution des galaxies à travers leur formation d'étoiles et les différentes fusions qu'elles ont subies. Le type de concentration utilisé est le rapport entre les rayons effectifs comprenant 80 % et 20 % de la luminosité totale de la galaxie. L'asymétrie est évaluée en

tournant l'image de 180 degrés et en effectuant la différence avec l'image originale. Enfin la texture (clumpiness) recherchée dans l'image est celle comprenant les hautes fréquences spatiales. Ce système a été utilisé dans le but de montrer sa corrélation avec des modes d'évolution des galaxies : la formation des étoiles et l'activité de fusion. Dans (5) un nouveau paramètre est utilisé : le moment du second ordre des 20 % les plus brillants du flux de la galaxie. Ce paramètre a été testé avec les coefficients de Gini et le système CAS. Cet article montre l'importance d'utiliser différents paramètres suivant les types de galaxies et leur éloignement. Dans (6), les auteurs introduisent un nouveau paramètre de texture. Pour cela ils découpent la galaxie suivant plusieurs anneaux et ils déterminent l'intensité lumineuse moyenne à l'intérieur de ces anneaux. Ce paramètre, utilisé avec la concentration, a permis de différencier les galaxies elliptiques des galaxies spirales avec un taux de réussite de 88%.

La plupart des études ont été réalisées sur des bases dont les galaxies sont triées selon des critères de formes ou de distances. Les distributions des galaxies selon leur taux de floculence et de *hotspots* sont représentées Fig2.6. On remarque que les distributions présentent une forte dissymétrie gauche, dûe au fait que les galaxies très floculentes et présentant beaucoup de *hotspots* sont largement minoritaires dans les relevés astronomiques. La connnaissance à priori sur la galaxie à traiter permet d'éviter des erreurs de classifications. En effet pour le critère de la floculence, les galaxies elliptiques et lenticulaires doivent présenter des taux de floculence relativement faibles en raison de leur déficience en gaz moléculaire susceptible d'alimenter la formation stellaire et de créer des motifs "floconneux". C'est ce que l'on retrouve dans notre catalogue de galaxies (cf Tab2.2 et Tab2.3), les taux de floculence élevés se retrouvent principalement dans les galaxies spirales (entre 80 % et 96 % pour les taux 1 a 4) et sont quasi inexistantes dans les galaxies elliptiques et lenticulaires qui représentent respectivement 24 %et 48.5 % des galaxies de taux 0. Il faut noter que les trois-quarts des galaxies spirales et irrégulières ont un taux de 1 ou de 2. Les distributions pour les *hotspots* suivent à peu près les mêmes chiffres, avec une différence un peu moins marquée pour les galaxies très floculentes, même si les galaxies spirales représentent tout de même toujours 70 % des exemples de chaque taux. De plus pour tous les types de galaxies, la majorité a un taux 0 et très peu de galaxies ont un taux supérieur à 2.

Taux de floculence ou de *hotspots* / Types de galaxie	0	1	2	3	4
Elliptiques	24 - 6.3	0.5 - 5.4	0.3 - 6.2	0 - 7.7	0 - 5
Lenticulaires	48.5 - 13.8	7.4 - 15.4	2.6 - 17.2	0.8 - 14.3	0.7 - 15
Spirales	25.2 - 71.8	79.9 - 71	87.3 - 67.3	91.2 - 70.3	96.3 75
Irrégulières	2.3 - 8.1	12.2 - 8.2	9.8 - 9.3	8 - 7.7	3 - 5
	100 - 100	100 - 100	100 - 100	100 - 100	100 - 100

TAB. 2.2 – Répartition des taux de floculence (rouge) et de hotspots (bleu) suivant les types de galaxies (en pourcentage)

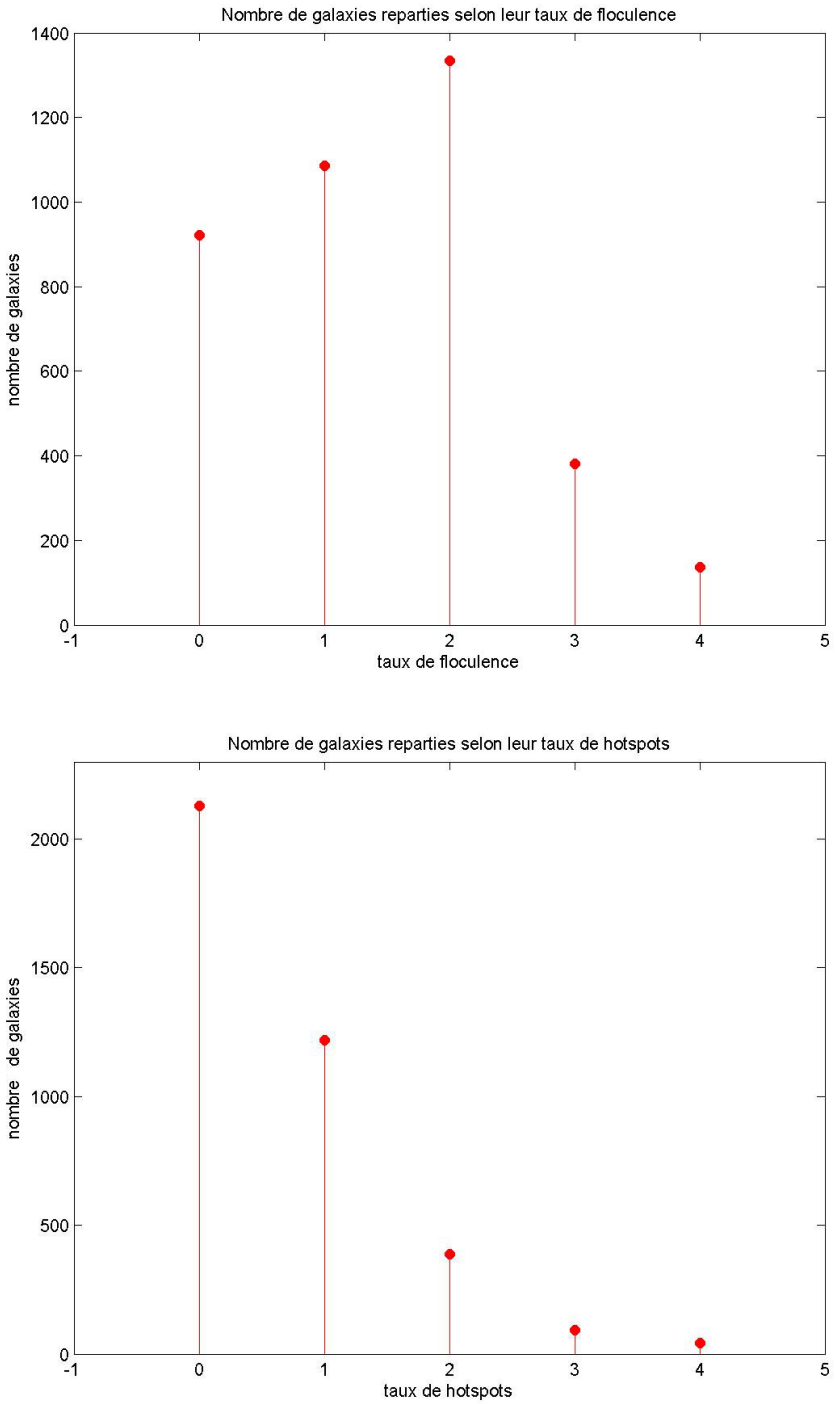


FIG. 2.6 – Distribution des galaxies suivant les taux de floculence et de hotspots

Types de galaxie \ Taux de floculence ou de hotspots	0	1	2	3	4	
Elliptiques	96	2.2	1.8	0	0	100
	57.3	28.4	10.2	3.1	1	100
Lenticulaires	79	14.1	6.2	0.5	0.2	100
	51.5	33.4	11.6	2.4	1.1	100
Spirales	8.5	31.4	42.5	12.8	4.8	100
	55.3	31.7	9.4	2.4	1.1	100
Irrégulières	6.5	41.4	41.1	9.7	1.3	100
	53.7	32	11.3	2.3	0.7	100

TAB. 2.3 – Répartition des types de galaxies suivant leur taux de floculence (rouge) et de hotspots (bleu) (en pourcentage)

2.3 Problématique

Le problème de la détection de la floculence peut se rapprocher de l'étude des nuages dans les images satellitaires. En effet la floculence se présente sous la forme de perturbations dans certains bras de galaxie, mais même si elle est souvent localisable pour les galaxies les plus floculentes, c'est parfois une impression de floculence globale sur la galaxie qui permet de déterminer son taux. Les études de classification réalisées au LTCI montrent que sur des images terrestres prises par des satellites, on peut obtenir de très bons résultats en classant des images selon des critères de texture (9). Notamment, les nuages peuvent être détectés par rapport au sol. Sinon, les autres recherches sur les détections ou les reconnaissances de nuages ont montré que les principales solutions étaient des analyses spectrales sur les images de nuages ou plus directement des analyses temporelles, pour détecter la présence ou non des nuages. Certaines études météorologiques ont utilisé un apprentissage à l'aide de réseaux de neurones entrainés à discriminer certains types de nuages. Le problème de toute classification est la définition de la base d'apprentissage, qui va permettre aux classificateurs de discriminer au mieux les différentes classes.

La principale difficulté est de pouvoir caractériser ces deux indices, floculence et *hotspots*, en s'affranchissant de la réponse impulsionnelle. Pour déterminer la floculence au sein des galaxies, la méthode employée est une approche par texture. Les images peuvent être considérées comme des mosaïques de régions de différentes textures qui présentent des caractéristiques différentes. On veut donc pouvoir extraire des données statistiques de ces régions afin de pouvoir séparer les régions floculentes des autres régions. Pour cela, deux méthodes ont été réalisées dont l'implémentation est représentée Fig2.7 :

– La première méthode pour réaliser cette séparation est la création d'une base d'apprentissage à partir de fenêtres des images de galaxies qui ont été étiquetées à l'aide des données fournies par les astronomes. Ensuite, nous utilisons un classificateur de plus proches voisins afin d'entraîner cette base et de la tester sur les images de galaxies. L'indice représentant le taux de floculence est un indice global de la galaxie. Il faut donc pouvoir déterminer de quelle manière va être calculé ce taux. Si l'on parvient à déterminer les zones de floculence au sein de la galaxie, le taux peut être exprimé comme un rapport entre l'aire de la zone

floculente et l'aire de la galaxie (cela pourrait aussi être le rapport entre la luminosité de la zone floculente et la luminosité totale de la galaxie).

– La deuxième méthode s'affranchit de l'étiquetage des fenêtres et permet de caractériser une galaxie à l'aide d'un unique vecteur. Pour cela le nombre de classes permettant de décrire au moins les fenêtres des galaxies est déterminé à l'aide d'une étude comprenant l'algorithme des kmeans et un algorithme de minimisation d'erreur. Chaque galaxie est représentée suivant son histogramme de classes. La classification s'effectue alors avec des machines à support de vecteurs sur les histogrammes étiquetés suivant les taux de floculence.

Il a été choisi de travailler à la fois sur la base d'images des résidus fournies par l'IAP et sur les images originales.

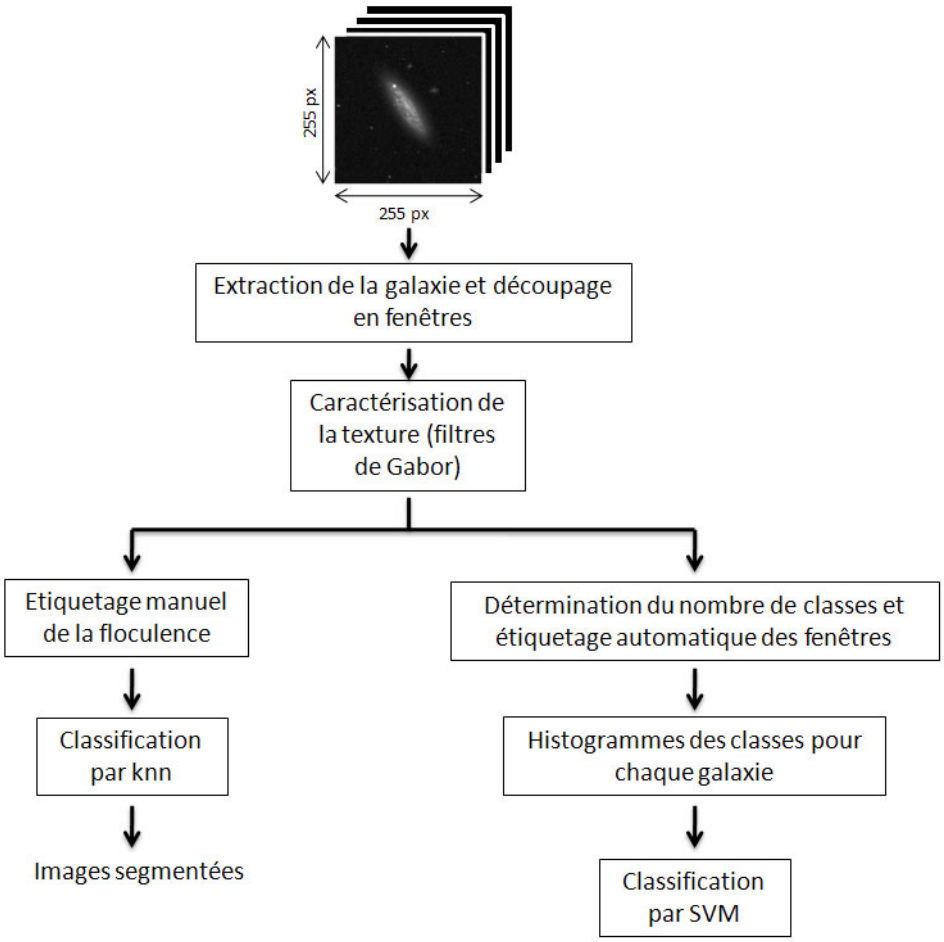
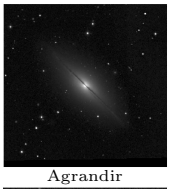
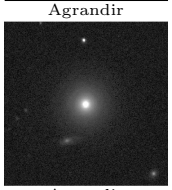
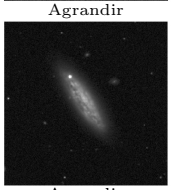
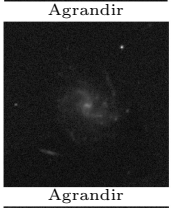
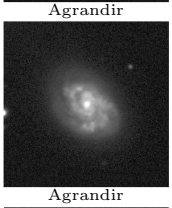
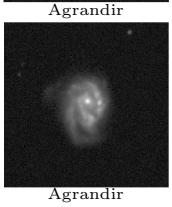
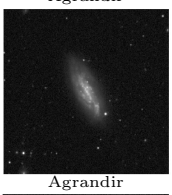
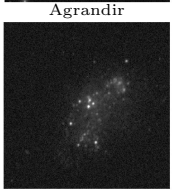


FIG. 2.7 – Schéma représentant les différentes classifications

Galaxies	Taux de floculence	Taux de hotspots	
PGC0000218	0 0 1	0 0 0	Agrandir 
PGC0008718	0 0 0	0 0 0	Agrandir 
PGC0010065	4 4 4	1 1 2	Agrandir 
PGC0027993	1 2 2	0 1 1	Agrandir 
PGC0029209	3 4 4	1 2 2	Agrandir 
PGC0030136	1 1 2	1 2 3	Agrandir 
PGC0039925	2 3 4	2 3 3	Agrandir 
PGC0044491	1 2 2	3 4 4	Agrandir 

Tab. 2.4 – Exemples de galaxies de la base PGC 1.3 avec leur taux de floculence et de hotspots associés. Les valeurs entourant la valeur centrale correspondent à un intervalle de confiance

Chapitre 3

Analyse de la texture des galaxies

3.1 Découpage des fenêtres

Afin de limiter les effets de textures qui se situent en dehors des galaxies, il a été choisi de réduire la zone d'intérêt à un rectangle englobant la galaxie. Pour cela, la méthode employée sur les images des galaxies où le fond de ciel a été soustrait, mais sur lesquelles il reste les objets lumineux, a été de labeliser ces objets (dont la galaxie) à l'aide d'une analyse en composantes connexes (8 connexités). Parmi les labels obtenus, le label contenant la galaxie est conservé en utilisant le fait que les galaxies sont centrées donc le label du pixel central appartient à la galaxie. Ainsi nous pouvons conserver les limites des rectangles englobant la galaxie qui serviront de limite pour les fenêtres d'étude. Cette méthode permet de diminuer l'influence du fond de ciel dans notre étude.

Les caractéristiques sont extraites localement sur les images de galaxie. L'image est parcourue par une fenêtre dont la taille a été fixée dans un premier temps à 15 pixels sur la première base de 400 galaxies, puis à 25 pixels sur la base la plus récente. Pour cela seuls les centres des fenêtres sont sauvegardés. Les premiers tests ont été effectués sur des fenêtres extraites dans un repère cartésien. Le décalage entre chaque fenêtre est de 12 pixels, soit un recouvrement de 50 %.

3.2 Filtrage de Gabor

Les filtres de Gabor permettent de caratériser les textures à l'aide de filtrages passe-bande. L'usage des filtres de Gabor en analyse d'images est souvent motivé par leur similitude avec les fonctions de réponse des champs réceptifs visuels humains. Une telle approche est très intéressante car même si des études commencent à expliquer le système visuel humain, son efficacité d'analyse est encore loin d'être modélisée. Les filtres de Gabor permettent de séparer les textures et sont très largement utilisés dans les méthodes de segmentation et de classification (7) (8). Un filtre de Gabor est une gaussienne modulée par une sinusoïde (cf Fig3.1). La fonction de transfert d'un filtre de Gabor s'écrit dans le domaine spatial :

$$h(x, y, u, \theta) = \frac{1}{2\pi\sigma_x\sigma_y} e^{(-\frac{1}{2})(\frac{x_\theta^2}{\sigma_x^2}+\frac{y_\theta^2}{\sigma_y^2})} cos(2\pi u x_\theta) \tag{3.1}$$

où $x_\theta = xcos(\theta) + ysin(\theta)$ et $y_\theta = -xsin(\theta) + ycos(\theta)$ et u et θ représentent la fréquence et

la phase selon l'axe des x. Les variances σ_x et σ_y permettent de caractériser l'enveloppe de la gaussienne et sont liées à la largeur de bande du filtre de Gabor.

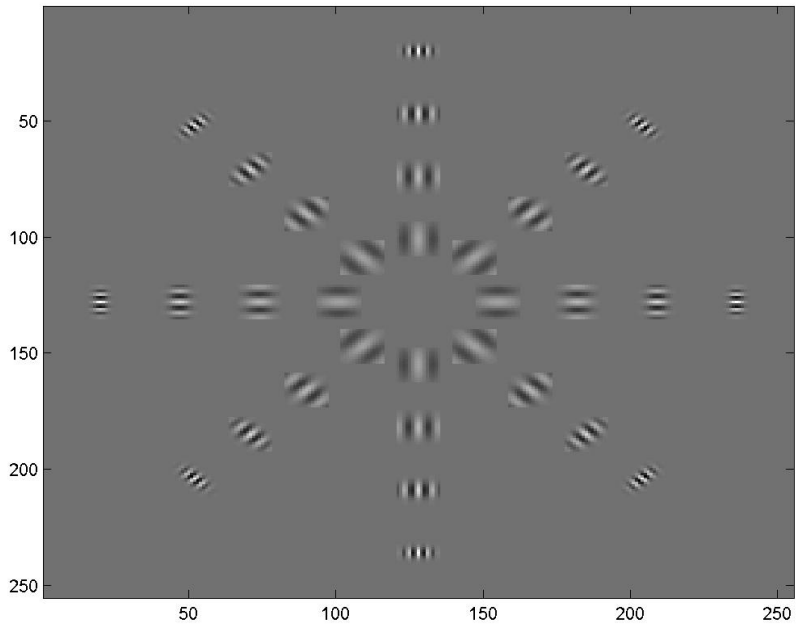


FIG. 3.1 – Affichage dans le domaine spatial des filtres de Gabor pour une taille de 20 pixels, 4 échelles variant de 0.1 à 0.3, 4 orientations variant de 0 à $\frac{3\pi}{4}$

Ce type de filtre permet d'analyser l'image sous différentes orientations et sous différentes échelles. Ainsi en appliquant un banc de filtres correspondant à ces différentes échelles et orientations sur des fenêtres carrées de l'image, on peut extraire un vecteur de paramètres en effectuant une analyse statistique des résultats des filtrages. Le but d'une telle approche est de pouvoir caractériser une texture *floculente* afin qu'un classificateur puisse séparer les fenêtres de galaxies en trois classes : le fond de la galaxie, les parties de la galaxie où il n'y a pas de floculence, et les parties de la galaxie où il y a de la floculence. Pour pouvoir établir un tel classificateur, il faut au préalable créer une base d'apprentissage, c'est-à-dire découper la galaxie en fenêtres et les étiqueter. Cet étiquetage est très important car il détermine la base du classificateur. Il a été réalisé en tenant compte des indices fournis par l'IAP pour chaque galaxie. Ces indices représentent un taux global de la floculence au sein des galaxies, les astronomes l'évaluent en prenant en compte à la fois la surface occupée par la floculence et sa luminosité par rapport à la galaxie. Le choix des paramètres est plus significatif dans le domaine des fréquences spatiales que dans le domaine spatial, c'est pourquoi on effectue la transformée de Fourier de l'équation 3.1, on obtient :

$$H(U,V) = A(e^{(-\frac{1}{2}[\frac{(U-u)^2}{\sigma_u^2}+\frac{v^2}{\sigma_v^2}])} + e^{(-\frac{1}{2}[\frac{(U+u)^2}{\sigma_u^2}+\frac{v^2}{\sigma_v^2}])}) \tag{3.2}$$

avec $\sigma_u = \frac{1}{2\pi\sigma_x}, \sigma_v = \frac{1}{2\pi\sigma_y} et A = 2\pi\sigma_x\sigma_y$. Soient U_h et U_b les fréquences hautes et basses délimitant le domaine fréquentiel, nE le nombre d'échelles et nO le nombre d'orientations. Les

équations suivantes de σ_u et σ_v assurent un recouvrement de 50 % des pics de réponse, ainsi que l'uniformité des fréquences dans l'intervalle $[U_b\ U_h]$ défini par l'utilisateur :

$$a = (U_h/U_b)^{(1/(nE-1))}, \qquad \sigma_u = \frac{(a-1)U_h}{(a+1)\sqrt{2ln(2)}},$$

$$\sigma_v = tan(\frac{\pi}{2k})[U_h - 2ln\frac{\sigma_u^2}{U_h}][2ln2 - \frac{(2ln2)^2\sigma_u^2}{U_h^2}]^{-\frac{1}{2}} \tag{3.3}$$

Chaque fenêtre est ensuite convoluée par chaque filtre. Puis, on extrait les caractéristiques issues de ce filtrage qui sont la moyenne et la variance. Ces deux paramètres sont utilisés pour représenter les régions afin de pouvoir établir la classification, et ils sont normalisé en ramenant la moyenne à 0 et la variance à 1 pour la base de vecteurs obtenus. Ainsi pour une fenêtre, le vecteur de caractéristiques C contient 32 composantes issues des 16 filtres utilisés :

$$C = (moy\_filtre1, var\_filtre1, ..., moy\_filtre32, var\_filtre32) \tag{3.4}$$

Afin de prendre en compte le fait que les galaxies sont composées d'objets en rotation, et donc que les régions assimilables à une texture de floculence doivent contenir cette information de rotation, il semble judicieux de placer les centres des fenêtres dans un repère polaire et de tourner autour du centre de la galaxie.

Pour pouvoir utiliser une telle méthode, il faut calculer les centres des fenêtres dans ce repère à l'aide des équations :

$$x = \rho cos(\theta) y = \rho sin(\theta) \tag{3.5}$$

Afin de fournir une description homogène de la galaxie, les fenêtres doivent avoir un recouvrement presque constant sur toutes les zones de la galaxie. On place le centre du repère au centre de la galaxie. Le pas dans la direction radiale ρ reste constant et égal à 12 pixels. Le pas de l'angle de rotation θ dépend de la valeur de ρ donc il est variable suivant la distance par rapport au centre. Il est défini de la manière suivante en fonction du pas P_ρ de ρ et du pas initial $P_{\theta 0}$ de θ

$$P_\theta = \frac{P_{\theta 0}}{P_\rho} \tag{3.6}$$

Le pas initial a été choisi à $P_{\theta 0} = \frac{12\pi}{4}$ pour pouvoir assurer le recouvrement des fenêtres sur la totalité de la galaxie.

On peut imaginer que les réponses des filtres seront proches dans des bras spiraux lorsque ceux-ci ont la même orientation. Or les orientations des filtres appliqués sur les fenêtres d'une galaxie sont définies dans le repère cartésien. Il est intéressant de décaler les filtres en fonction de l'angle θ de la position du centre de la galaxie, mais cette méthode n'est pas envisageable car il faudrait calculer tous les filtres de Gabor pour chaque orientation, et les temps de calcul seraient beaucoup trop longs. Une solution moins coûteuse en temps de calcul est d'effectuer des permutations circulaires sur le vecteur de paramètres en fonction de la position

de la fenêtre dont est issu le vecteur. Les fenêtres sont filtrées suivant quatre orientations (de 0 à $\frac{3\pi}{4}$), le vecteur de paramètres va donc subir des rotations en fonction de ces orientations. Les règles de rotation en fonction de la position de la fenêtre sont , si le vecteur initial est $C = (MC_0, MC_{\frac{pi}{4}}, MC_{\frac{pi}{2}}, MC_{\frac{3pi}{4}})$:

- si $\theta \in \left[0\ \frac{\pi}{8}\right[$ ou $\left[\frac{7\pi}{8}\ \pi\right[$ modulo(π), le vecteur reste le même.
- si $\theta \in \left[\frac{pi}{8}\ \frac{3pi}{8}\right[$ modulo(π), le vecteur devient : $C = (MC_{\frac{pi}{4}}, MC_{\frac{pi}{2}}, MC_{\frac{3pi}{4}}, MC_0)$
- si $\theta \in \left[\frac{3pi}{8}\ \frac{5pi}{8}\right[$ modulo(π), le vecteur devient : $C = (MC_{\frac{pi}{2}}, MC_{\frac{3pi}{4}}, MC_0, MC_{\frac{pi}{4}})$
- si $\theta \in \left[\frac{5pi}{8}\ \frac{7pi}{8}\right[$ modulo(π), le vecteur devient : $C = (MC_{\frac{3pi}{4}}, MC_0, MC_{\frac{pi}{4}}, MC_{\frac{pi}{2}})$

3.3 Paramètres statistiques des distributions

Lors des traveaux précédents pour tenter de classer automatiquement les galaxies, des études ont été menées pour pouvoir caractériser la distribution de lumière d'une galaxie. Pour cela, de nombreux indices dérivés de la concentration ont été mis en place. C'est le cas d'un indice permettant de déterminer la répartition de la lumière dans une galaxie : le coefficient de Gini (3) (développé par le statisticien italien Corrado Gini). Ces coefficients sont issus d'une courbe très utilisée en statistique : la courbe de Lorentz qui peut permettre par exemple d'étudier les répartitions de la richesse dans une population où la courbe représenterait la proportion cumulée des revenus en fonction du nombre d'individus ayant ces revenus. Ainsi si les richesses sont réparties de façon égale, alors la courbe de Lorenz est une droite, et si par exemple, les plus riches deviennent de moins en moins nombreux et de plus en plus riches, la courbe sera proche de zéro avec uniquement un pic à la fin de la courbe. Le coefficient de Gini, qui se déduit de cette courbe, est défini comme étant le rapport entre l'aire comprise entre la ligne d'égalité et la courbe de Lorentz obtenue (A) et l'aire totale du triangle sous la ligne d'égalité (A+B) (cf Fig3.2). Donc un coefficient de Gini proche de 0 traduit l'égalité entre les individus (ce qui pourrait être un système communiste idéal) et proche de 1 si tout est concentré autour d'un seul individu (une monarchie absolue par exemple). Pour les galaxies, ces coefficients vont permettre de calculer la répartition de l'intensité lumineuse sur les pixels. Il faut noter que ces coefficients ne renseignent pas sur l'emplacement des pixels lumineux mais uniquement sur leur répartition. L'expression mathématique de la méthode donnée ci-dessus est pour la courbe de Lorentz :

$$L(p) = \frac{1}{\bar{X}} \int_0^p F^{-1}(u)du \tag{3.7}$$

Où X est une variable aléatoire positive comprenant n éléments de valeur moyenne $\bar{X}$, dont la fonction de distribution cumulée est F(x). Le coefficient de Gini s'obtient alors par le calcul suivant :

$$G = \frac{1}{2\bar{X}n(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} |Xi - Xj| \tag{3.8}$$

Une façon plus rapide de calculer ce coefficient est de préalablement ranger les Xi dans l'ordre croissant, puis de sommer les éléments à l'aide de la formule :

$$G = \frac{1}{2\bar{X}n(n-1)} \sum_{i=1}^{n} (2i - n - 1)Xi \tag{3.9}$$

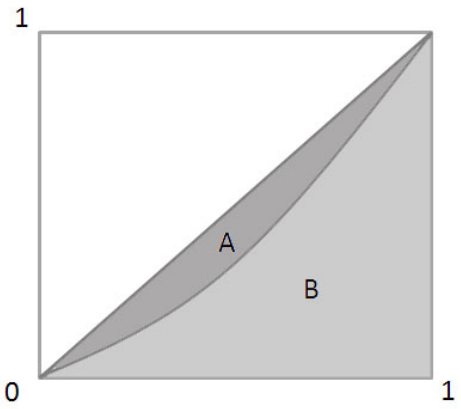


FIG. 3.2 – Interprétation géométrique des coefficients de Gini basée sur la courbe de Lorentz. Sur le graphique sont représentées la droite d'égalité et une courbe qui représente une distribution moins égale. Plus l'écart avec la droite d'égalité est grande, plus l'inégalité est grande. Le coefficient de Gini est le rapport entre l'aire A et l'aire sous la diagonale A+B.

D'autres éléments peuvent compléter cette analyse statistique : les coefficients de Skewness et de Kurtosis (11). Le coefficient de Skewness mesure le degré d'asymétrie de la distribution et se calcule comme étant le moment d'ordre 3 centré et l'écart-type au cube :

$$S = \frac{E(X-\mu)^3}{\sqrt{E(X-\mu)^2}^3} \tag{3.10}$$

Ce qui se calcule dans la pratique pour n éléments par :

$$S = \frac{n}{(n-1)(n-2)} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s} \right)^3 \tag{3.11}$$

Si S est en-dessous de 0, alors la distribution est asymétrique vers la gauche, si S est au-dessus de 0, alors la distribution est asymétrique vers la droite, et si S = 0, la distribution est symétrique.

Enfin le coefficient de Kurtosis mesure le degré d'écrasement d'une distribution et est défini par le rapport entre le moment d'ordre 4 centré sur la variance au carré :

$$K = \frac{E(X-\mu)^4}{[E(X-\mu)^2]^2} \tag{3.12}$$

Ce qui se calcule dans la pratique pour n éléments par :

$$K = \frac{n(n+1)}{(n-1)(n-2)(n-3)} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s} \right)^4 - \frac{3(n-1)^2}{(n-2)(n-3)} \tag{3.13}$$

Un coefficient de Kurtosis positif indique que la forme de la distribution est pointue tandis qu'un coefficient négatif indique une forme de distribution écrasée.

3.4 Etiquetage manuel des fenêtres

La première méthode consiste à considérer qu'un astronome étant capable de détecter et de localiser la floculence présente dans une galaxie, on peut penser pouvoir extraire les textures floculentes de l'image et quantifier l'apparition de cette texture au sein d'une galaxie. L'une des principales motivations de la mise en place d'une méthode de classification de textures dans les images à l'aide de classificateurs type knn ou SVM est la très grande facilité d'une personne lambda à pouvoir discriminer ces textures sans être un spécialiste et avec un taux d'erreurs quasi-nul. Pour les images de galaxies, le problème est légèrement biaisé car l'extraction des différents paramètres morphologiques nécessitent certaines connaissances préalables. Afin de créer une base disponible pour la communauté des astronomes, l'évaluation des différents paramètres des galaxies a été effectuée par différents astronomes afin de pouvoir recouper et comparer les différents résultats pour éviter les erreurs de jugement qui peuvent survenir étant donné la quantité importante de galaxies à traiter. La première méthode utilisée a été d'étiqueter manuellement les fenêtres en trois classes : le fond de la galaxie (C_1), les parties de la galaxie non floculentes (C_2), et les parties de la galaxie floculentes (C_3) tout en respectant le taux global de la galaxie. Le but de la classification est alors d'associer une observation $X = (x_1, x_2, ..., x_n)$ à une classe C_i. Deux exemples d'étiquetage d'une galaxie en coordonnées polaires sont donnés dans la Fig3.3

Type de fenêtres	fond de galaxie	zone de la galaxie non floculente	zone de la galaxie floculente	Total
Nombre de fenêtres	23919	29056	2793	55768

TAB. 3.1 – Nombre de fenêtres de chaque classe pour 220 galaxies

L'écart du nombre de fenêtres entre les classes (Tab3.1) impose d'effectuer une sélection de vecteurs décrivant au mieux chacune des classes. Le nombre élevé de vecteurs ne permet pas d'étudier de façon statistique cette base. Les premières classifications ont été effectuées en conservant toutes les fenêtres de la floculence sur 220 galaxies et en choisissant de façon aléatoire le même nombre de fenêtres de galaxies non floculentes et de fond du ciel. Ainsi chacune des trois classes était représentée par environ 3000 vecteurs. Chaque fenêtre est représentée par un vecteur de 32 composantes issues des filtrages par les filtres de Gabor, le classificateur utilisé est un k-ppv (k plus proches voisins) ou k = 5.

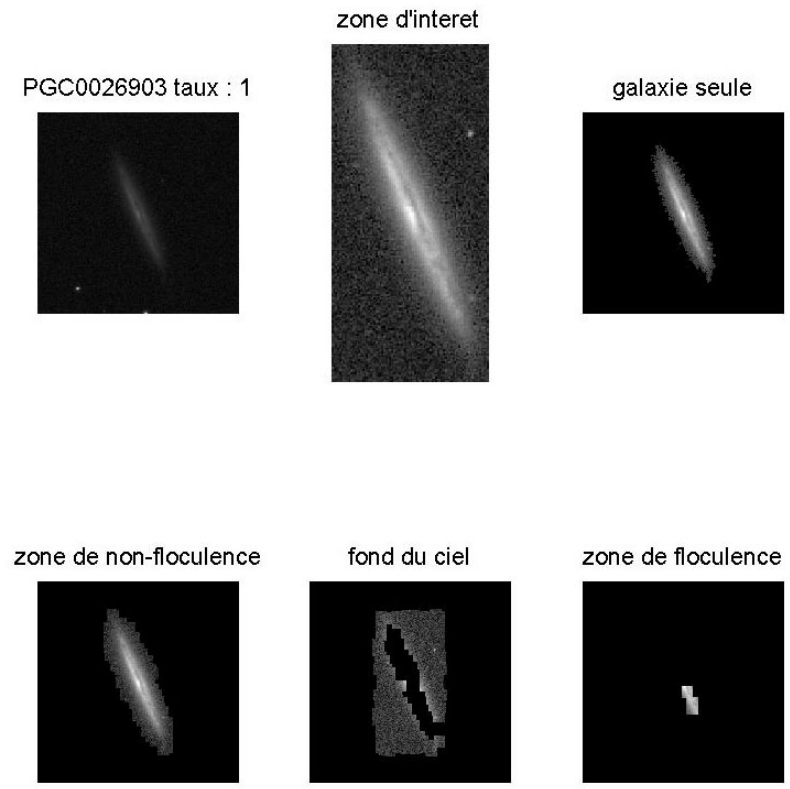


FIG. 3.3 – Exemples des différentes étapes d'étiquetage des galaxies en coordonnées polaires.

Etiquetage automatique des fenêtres

De part leur formation, les galaxies présentent de nombreux objets ou textures et le découpage en trois classes semble très insuffisant pour pouvoir extraire correctement la floculence, la classe non floculente présente trop de similitude avec la classe floculente. Le nombre de classes que l'on peut extraire dans une série d'images est toujours délicat à déterminer et dépend généralement de l'utilisateur. Certaines méthodes se sont développées afin de pouvoir déterminer de façon automatique le nombre de classes présentes dans une base d'images. Pour cela il faut pouvoir comparer entre eux les résultats obtenus lorsqu'on change le nombre de classes. La méthode qui suit permet de caractériser une galaxie par un seul vecteur. Or suivant leur inclinaison, leur forme et leur taille, la quantité de fond de ciel présent dans un rectangle englobant la galaxie n'est pas la même. C'est pourquoi, contrairement à la méthode précédente, les fenêtres d'étude appartiennent toutes à la galaxie, seules les fenêtres étudiées au bord des images contiennent des pixels se rapportant au fond de ciel.

Algorithmes des k-moyennes

Un algorithme couramment utilisé pour regrouper des données dans plusieurs classes est l'algorithme des k-means (12) ou k-moyennes, où k représente le nombre de nuages que l'on veut obtenir. Cet algorithme est basé sur le calcul des distances quadratiques entre chaque point d'un nuage et son centre.

i). Dans un premier temps, les centres z des nuages sont choisis de façon aléatoire parmi les points de la base.

ii). Ensuite tous les points appartiennent à la classe C_j dont ils sont les plus proches :

$$x \in C_j(m) si \forall i \|x - z_j(m)\| < \|x - z_i(m)\| \tag{3.14}$$

iii). Puis les nouveaux centres de chaque nuage sont calculés :

$$z_j(m+1) = \frac{1}{N_j} \sum_{x \in C_j(m)} x \tag{3.15}$$

iv). Alors, si $\forall j z_j(m+1) = z_j(m)$, la procédure s'arrête sinon on retourne en 2.

Cet algorithme a l'avantage d'être rapide lorsque le nombre de nuages n'est pas trop important, mais il a pour défaut la nécessité de la connaissance du nombre de nuages k, et le résultat de ce classement dépend de l'initialisation des centres des nuages de points.

Détermination du nombre de classes

Afin de déterminer le nombre de classes qui permet la meilleure classification, on utilise une méthode décrite par Yvan Kyrgyzov (10) basé sur un algorithme de minimisation du critère MDL (Minimum Length Description). Le but d'une telle méthode est de trouver les différentes catégories de textures présentes dans les images de galaxies, sans connaissance à priori sur le type et le nombre de ces catégiories. Le principe de cette algorithme est de comparer différents résultats obtenus en modifiant le nombre de classes dans la méthode des k-moyennes. Le critère de comparaison est le critère MDL, proposé par N. Rissanen dans les années 1970, et qui est

utilisé afin de décrire de façon optimal une distribution de données pour un certain modèle de données. Cette méthode a obtenu de très bons résultats lorsque la base possédait beaucoup de données, c'est pourquoi nous l'utilisons ici. La figure 3.4 donne les résultats du critère MDL en fonction du nombre de classes. Ce graphique devrait comporter un minimum global, mais les calculs au-dela de 70 classes créaient des erreurs de mémoire sous matlab. On peut tout de même considérer que pour un nombre de classes supérieur à 20, le critère MDL n'évolue plus de façon significative.

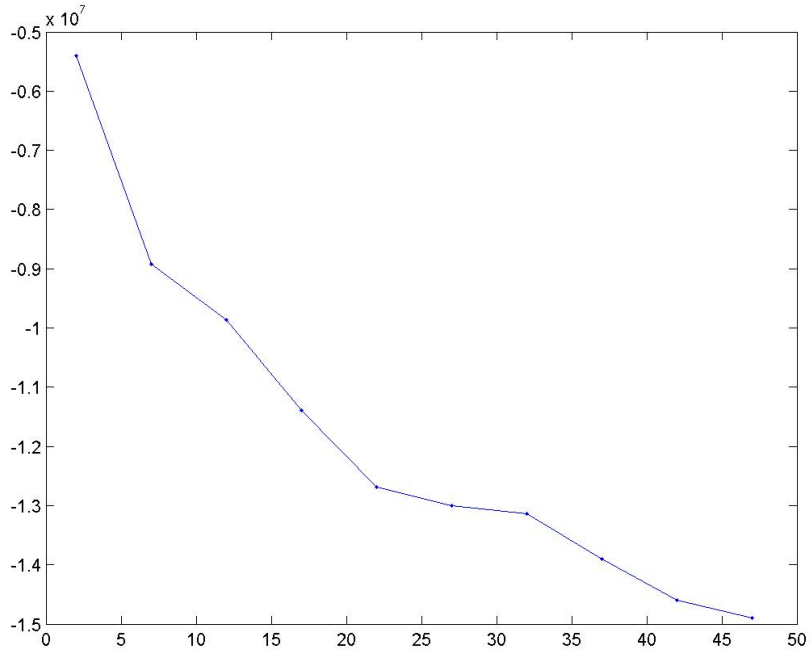


FIG. 3.4 – Détection du nombre optimal de classes. La courbe représente le critère MDL en fonction du nombre de classes.

Evaluation de l'histogramme

La méthode précédente permet de définir les centres de classes parmi une base de taille importante. La principale difficulté de la première méthode de classification est de pouvoir décrire localement la floculence afin de la quantifier sur la galaxie entière. En effet considérer un ensemble de galaxies représentées chacune par environ 350 vecteurs de dimension 32 n'est pas envisageable dans un problème de classification. La solution revient à trouver un nombre raisonnable de classes composant les images de galaxies, et d'étiqueter chaque fenêtre de chaque galaxie dans l'une de ces classes. La méthode du MDL montre qu'une vingtaine de classes semble suffisante pour décrire l'ensemble des fenêtres des 3860 galaxies (cf Fig3.4). Ainsi la base initiale est : $X = [X_1, X_2, ..., X_n]$ où chaque X_i représente un vecteur de caractéristiques issu des filtres de Gabor (ce qui représente n = 468428 vecteurs pour les 3860 galaxies). A chaque vecteur X_i, on associe une classe Cl_i à l'aide de l'algorithme des k-moyennes avec i $\in$ [1..20]. Puis on évalue le nombre d'occurences de chaque classe qui se trouve dans chaque galaxie. On obtient ainsi un histogramme (le nombre d'occurences en fonction des numéros des classes, cf Fig3.5) pour chaque galaxie. Chaque histogramme est normalisé par rapport à l'ensemble de la base

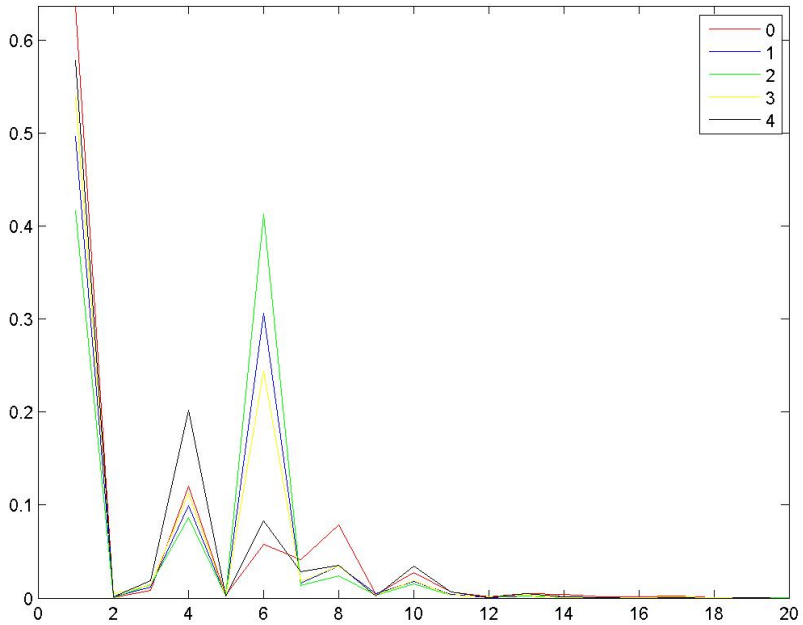


FIG. 3.5 – Moyenne des histogrammes pour chaque taux de floculence.

d'histogrammes obtenue. Or on dispose de la table de chaque galaxie en fonction du taux de floculence. Ainsi à chaque histogramme, on peut assimiler un taux de floculence. L'objectif est de pouvoir assimiler les différents histogrammes aux taux de floculence. On peut alors quantifier ces classes pour chaque galaxie. On obtient les histogrammes que l'on normalise afin de pouvoir les comparer.

SVM (Support Vector Machine)

Les SVM ou Séparateurs à Vastes Marges ont été introduits par Vapnik en 1995. C'est une méthode de classification par apprentissage supervisé dont les paramètres du modèle sont appris par un jeu de données d'apprentissage ((13) et (14)). Le modèle repose sur l'existence d'un classificateur linéaire dans un espace approprié. Le but est de séparer les données et de maximiser la distance entre deux classes. Pour un exemple à deux classes, il faut chercher à définir l'hyperplan optimal séparant les deux classes, c'est-à-dire en maximisant les distances entre les exemples des classes et cet hyperplan. La distance entre le point le plus proche et l'hyperplan est appelée la marge. La classification sera toujours ramenée à un problème à deux classes, c'est-à-dire chaque classe contre toutes les autres. Les données à classer sont clairement non linéairement séparables, donc c'est le cas de SVM avec un noyau qui va être utilisé ici afin de changer l'espace de données dans lequel les données seraient linéairement séparables. Cet espace augmente la dimension des données ce qui augmente les chances de trouver une séparation linéaire.

Le problème de la classification est de définir la classe d'appartenance C_i d'un vecteur d'entrée $x_i \in \mathbf{X}$ à partir d'un apprentissage sur une base dont on connaît les classes de chaque vecteur. On étudie donc une fonction $C_i = f(x_i)$.

Pour notre problème non linéairement séparable, l'hyperplan optimal est obtenu en résolvant le système :

$$\begin{cases} min(\frac{1}{2}||w||^2 + C\sum_{i=1}^{n}\xi_i \\ \text{sous la contrainte : } \forall i, y_i(w.x_i + b) \geq 1 - \xi_i \end{cases} \tag{3.16}$$

Les variables à estimer sont les poids w ainsi que le biais b et où le terme $C\sum_{i=1}^{n}\xi_i$ est un terme de pénalisation. La forme duale de ce système est donnée par :

$$\begin{cases} max\sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i,j}\alpha_i\alpha_j x_i x_j \\ \text{sous la contrainte : } \forall i, 0 \leq \alpha_i \leq CC\sum_{i=1}^{n}\alpha_i y_i = 0 \end{cases} \tag{3.17}$$

Dans le cas non linéaire, la dimension initiale de l'espace ne permet pas de résoudre la classification on passe donc à un espace de plus grande dimension, donc on augmente la dimension de l'espace des données par la transformation :

$$\begin{aligned} \Phi : \mathbb{R}^d &\rightarrow F \\ x &\rightarrow \Phi(x) \end{aligned} \tag{3.18}$$

Le système 3.17 devient alors :

$$\begin{cases} max\sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i,j}\alpha_i\alpha_j\Phi(x_i)\Phi(x_j) \\ \text{sous la contrainte : } \forall i, 0 \leq \alpha_i \leq CC\sum_{i=1}^{n}\alpha_i y_i = 0 \end{cases} \tag{3.19}$$

dont la solution est de la forme :

$$f(x) = \sum_i = 1^n \alpha_i^* y_i \Phi(x_i)\Phi(x) + b \tag{3.20}$$

La solution dépend du produit scalaire $\Phi(x_i)\Phi(x)$, que l'on associe à une fonction noyau : $k : XxX \rightarrow \mathbb{R}$ où k est linéaire. Le noyau utilisé pour notre classification est un noyau gaussien :

$$k(x, x') = \exp -||x - x'||^2/\sigma \tag{3.21}$$

La base d'apprentissage est donc composée de p vecteur $\{x_1, ..., x_p\}$ de classes connues $\{Cl_1, ..., Cl_p\}$, $Cl_i \in \{1, 2, 3, 4, 5\}$. Cette base permet de construire la fonction de décision qui va permettre de discriminer les différentes classes suivant sa valeur. Les données à classer sont clairement non linéairement séparables, donc c'est le cas de SVM avec un noyau gaussien qui va être utilisé ici. La classification sera toujours ramenée à un problème à deux classes, c'est-à-dire chaque classe contre toutes les autres. De plus, afin d'optimiser les résultats de classification et afin d'utiliser toutes les données de la base, la classification sera effectuée avec cinq boucles de validation croisées. La validation croisée a pour but de mieux définir les classes, notamment lorsque le nombre d'exemples de la base est réduit. Son principe est de changer la base d'apprentissage et de test sur chaque boucle. Cette méthode permet de diviser la base initiale en k parties (ici k = 5), et sur chaque boucle, l'apprentissage est réalisé sur k-1 parties de la base, la dernière partie n'ayant pas servi à l'apprentissage va être utilisée pour effectuer le test. Le taux de classification est alors la moyenne des classifications obtenues à chaque étape.

3.5 Résultats

La base utilisée pour l'apprentissage et le test contient 400 galaxies découpées en fenêtres de 15x15 pixels. A chaque fenêtre est associé un vecteur de 32 caracéristiques obtenu par le filtrage des 16 filtres de Gabor ainsi qu'un label correspondant à l'appartenance de la fenêtre : soit au fond de ciel (C1), à la galaxie non floculente (C2), ou à la galaxie floculente (C3). La classification s'effectue sur 9000 vecteurs (3000 vecteurs par classe, qui sont choisis de façon aléatoire pour les classes C1 et C2), en utilisant cinq boucles de validation croisées.

Apprentissage supervisé sur les classes fond de ciel, galaxie floculente et non floculente

La matrice de confusion permet de juger les performances du classificateur. C'est un tableau à deux entrées permettant de comparer les classes attribuées par l'étiquetage manuel et par le classificateur. Le but de la classification est de détecter la floculence, donc la troisième colonne et la troisième ligne du tableau sont très importantes. La troisième colonne permet de connaître les taux de fausses alarmes, c'est-à-dire les fenêtres classées floculentes alors qu'elles sont non floculentes. La troisième ligne indique les taux de non-détection, c'est-à-dire les fenêtres floculentes classées comme fond de ciel ou galaxies non floculentes. Les matrices de confusion obtenues (Tab3.2 etTab3.3)sont représentées ci-dessous. La première matrice représente les taux de reconnaissance sur la base de test qui ne contient que quelques vecteurs de chaque classe issus de la base initiale mais ne contenant pas les vecteurs de la base d'apprentissage. La deuxième matrice de confusion a été obtenue en testant le classificateur sur une base comprenant tous les vecteurs de plusieurs galaxies.

Classe initiale \ Classe prédite	C1	C2	C3	
C1	88	8	4	100
C2	18	71	11	100
C3	9	16	75	100

TAB. 3.2 – Matrice de confusion (en %) sur une base de test réduite
"

Classe initiale \ Classe prédite	C1	C2	C3	
C1	72	20	8	100
C2	27	33	40	100
C3	4	16	80	100

TAB. 3.3 – Matrice de confusion (en %) sur des galaxies entières

Sur la base réduite, on s'aperçoit que les taux de reconnaissances sont relativement acceptables, la classe 1 étant même très bien reconnue. Mais lorsqu'on teste le même classificateur sur tous les vecteurs, on se rend compte que la classe floculente et le fond de ciel sont très bien reconnus, mais qu'il y a un taux de fausses alarmes très élevé. Ces fausses alarmes sont très majoritairement issues de la classe 2, dont les éléments sont répartis de façon presque égale entre les trois classes. Lorsqu'on observe les résultats en plaquant les masques de la floculence (cf Fig3.6), on observe des résultats très intéressants sur certaines galaxies, mais beaucoup trop

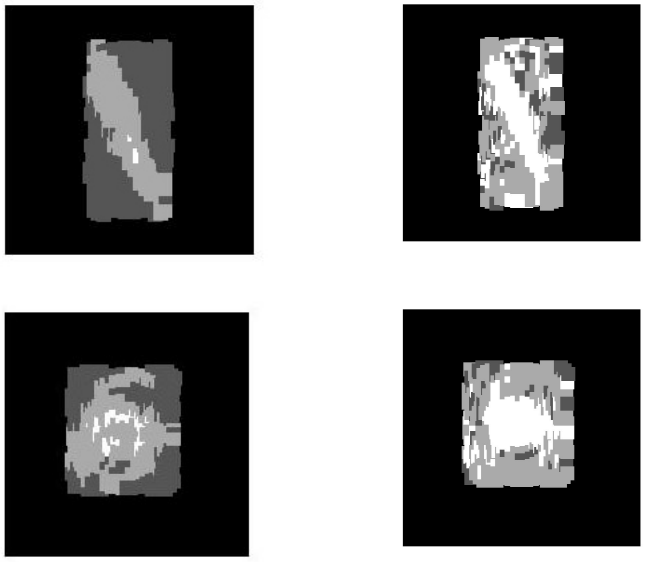


FIG. 3.6 – Résultat de la classification sur une galaxie : à gauche la galaxie avec ses étiquettes placées manuellement ; à droite, le résultat de la classification.

isolés pour être exploitables. Il faut noter aussi que ces résultats généralisés aux galaxies peuvent être améliorés en considérant que la classe floculente ne peut pas être voisine de la classe fond de ciel, mais malgré cette amélioration, le calcul de l'aire de la floculence sur l'aire de la galaxie n'est pas corrélé avec les taux de floculence. Une autre sélection a été réalisée, dans laquelle toutes les fenêtres de galaxies non-floculentes ne recouvrent aucune fenêtre floculente et toutes les fenêtres de fond de ciel ne recouvrent aucune fenêtre de galaxies (floculentes et non-floculentes) afin d'essayer de faire en sorte que les choix des vecteurs de la base d'apprentissage correspondent au mieux à chacune des classes, mais les résultats restaient à peu près identiques.

Ces premiers résultats laissent penser que la division en trois classes des images de galaxies n'est pas suffisante. En effet on observe une bonne reconnaissance du fond de ciel et de la galaxie floculente, par contre les fenêtres de galaxies non-floculentes sont réparties sur les trois classes. Ceci est principalement dû au fait qu'il existe certainement plusieurs classes de floculences, qui sont notamment déterminés par les types de galaxies et leur orientation. Le problème est de pouvoir déterminer une séparation entre des zones floculentes et des zones non floculentes, or cette séparation est difficile à déterminer. Les erreurs de classification sont aussi sans doute dûes aux nombreux objets qui se superposent aux galaxies et qui viennent fausser la classification. Enfin la grande dynamique des niveaux de gris des images fits, qui permet d'étudier des images astronomiques avec de nombreux détails, est sans doute la cause de variance intra-classes très élevée.

Apprentissage supervisé sur les histogrammes

La base est maintenant composée des histogrammes pour chaque galaxie. La classe qui leur est attribuée est celle provenant du tableau des taux de floculence établis par les astronomes. Le nombre d'exemples de chaque classe est limité par le nombre d'occurences de la classe qui est la moins représentée parmi les 3753 galaxies pour éviter les erreurs de surapprentissage. Dans notre étude, c'est le taux de floculence 4 qui est le moins représenté (133 galaxies). Afin d'augmenter

le nombre de galaxies de la base d'apprentissage et d'obtenir une meilleure représentation des résultats, différents tests ont été réalisés (Tab3.4) :

- p1 : problème à cinq classes (133 galaxies par classe)
- p2 : problème à quatre classes
- p3 : problème à trois classes : trois classes regroupées comme étant les taux <2, >2 ou =2 (849 galaxies par classes).
- p3 : problème à trois classes : soit trois classes de taux 0, 2 et 4 (133 galaxies par classe).
- p4 : problème à deux classes : le taux 0 contre les autres taux.
- p5 : problème à deux classes : le taux 0 contre le taux 5.

Les différents critères de classification sont :

- le type d'image utilisée : originale (a) ou résidu. (b)
- le type de vecteur caractéristique utilisé : les vecteurs ordonnés ou modifiés. (c)
- le changement d'échelle des paramètres des filtres de Gabor. (d)
- le nombre d'échelles et d'orientations des filtres de Gabor. (e)
- les coefficients statistiques et de Gini ajoutés au vecteur de statistiques. (f)
- La moyenne des filtres de Gabor est mise à zéro. (g)
- les filtres de Gabor sont convolués avec la réponse impulsionnelle. (h)

Méthode de calcul / Méthode de classe	a	b	c	d	e	f	g	h
p1	34.2/22.2	32.3/13.2	37.0/20.0	38.8/20.0	33,6/21.0	34.3/20.5	26.5/16.7	41.5/18.3
p2	-	-	-	-	33.0/10.2	-	-	-
p3	58.3/38.9	38.7/20.4	46.2/26.6	53.0/31.2	57.7/38.9	56.2/36.0	52.0/31.4	55.4/36.3
p4	62.7/11.2	-	-	-	-	64.7/8.5	69.2/6.5	
hline p5	81/0.6	-	-	-	-	-	-	-
p6	75.5/5	-	-	-		-	75.5/4.2	81.2/5.7

TAB. 3.4 – Taux de reconnaissance suivant les différentes méthodes employées(en moyenne des pourcentages de reconnaissance/écart-type des pourcentages de reconnaissance)

La plupart des classificateurs ne permet pas de discriminer chaque taux de floculence, pour le problème à cinq classes, les résultats sont autour de 30 % de reconnaissance pour les classes 1, 2, et 3. La méthode qui donne les meilleurs résultats est celle où on utilise la convolution. Si on considère qu'une galaxie est reconnue si le taux obtenu appartient à l'intervalle de confiance, on obtient un taux moyen de 67 % de reconnaissance avec un écart-type de 7%. Par contre la classe 0 est facilement reconnaissable (résultat aux environs de 80 % dans chaque cas). Mais le problème reste de classer les taux 1 à 4.

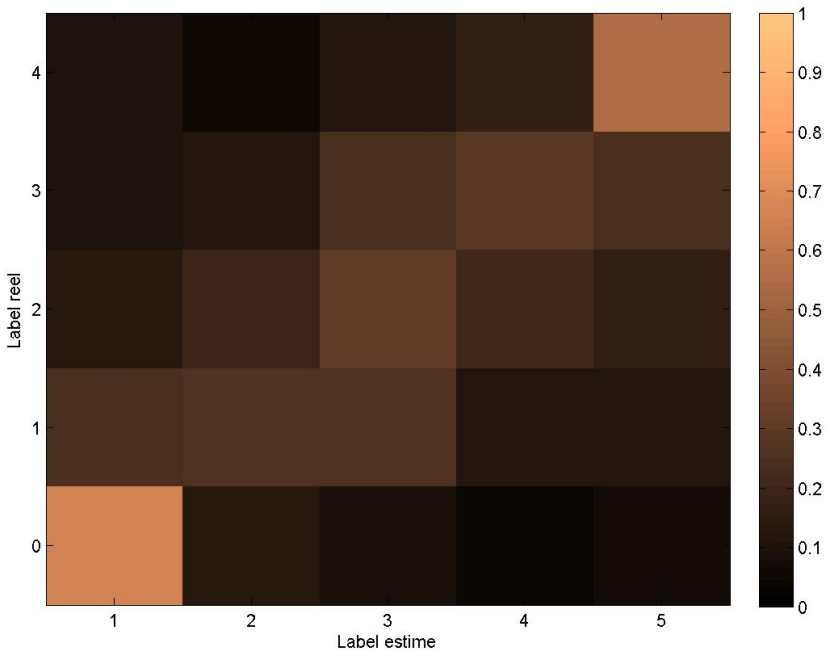


FIG. 3.7 – Affichage de la matrice de confusion pour le problème à cinq classes.

3.6 Perspective

Les classifications obtenues mettent en évidence que les classes sont mal représentées, donc les classificateurs ont des problèmes pour discriminer les différents taux de floculence. La classification avec la connaissance préalable du type de galaxie doit permettre de pouvoir mieux définir ces classes. Une autre méthode pour améliorer les résultats est d'utiliser des algorithmes de boosting. Le boosting est une méthode d'agrégation de classificateurs la plus efficace en pratique. Le principe est d'effectuer la classification n fois sur différentes distributions de probabilités des exemples des échantillons d'apprentissage, à chaque étape les poids des exemples sont mis à jour ainsi que le modèle final de classification. L'initialisation des distributions se fait généralement de manière uniforme mais étant donné qu'on possède un intervalle de confiance pour chaque galaxie on pourrait initialiser cette distribution selon la longueur de chaque intervalle.

Chapitre 4

Evaluation des taux de Hotspots

Les *hotspots* sont des régions dont le rayonnement est très important ce qui permet de les analyser sur les images de galaxies. Ils sont souvent définis comme étant des régions HII, régions dont l'émission est dominée dans le domaine visible par le rayonnement des étoiles jeunes massives et les raies d'émission des atomes de gaz excités par ce rayonnement. Ces régions contiennent généralement des étoiles chaudes et treès lumineuses qui ont une durée de vie relativement courte (par rapport au soleil par exemple).

Pour pouvoir analyser les taux de *hotspots* dans les images, la méthode mise en place doit s'affranchir de la présence des objets superposés aux galaxies. En effet les *hotspots* sont relativement facilement reconnaissables, car ce sont des points très lumineux localement, et une approche locale de la luminosité en étudiant la moyenne et la variance semblerait judicieuse afin d'extraire ces zones dans les galaxies. Mais une telle segmentation entraîne le choix de différents seuils qui sont très différents dans chaque image, surtout lorsque dans l'image se trouve une étoile qui modifie considérablement la luminosité moyenne totale de l'image. Les *hotspots* représentent aussi des perturbations à l'intérieur d'une galaxie, donc une autre approche serait d'analyser les perturbations locales dans une galaxie par rapport aux perturbations globales de cette même galaxie.

4.1 Recherche des maxima

Afin de détecter les maxima présents dans les images de galaxie, la première méthode consiste à rechercher les maxima locaux (cf Fig4.1). Pour cela on étudie les niveaux de gris à travers des fenêtres de $20x20$ pixels. Afin de pouvoir être significative, la taille d'un Hotspot doit être supérieure à la taille de la réponse impulsionnelle del'image. On étudie donc les propriétés d'une fenêtre dont la taille est définie par la réponse impulsionnelle, et dont on étudie les propriétés suivantes :

- la moyenne des niveaux de gris de cette fenêtre doit être suffisament élevée par rapport au niveau moyen de la galaxie
- la variation au niveau local est prise en compte en étudiant l'écart type entre les valeurs de la zone à étudier et les valeurs d'un anneau rectangulaire entourant cette zone.

Mais ces paramètres nécessitent de définir des seuils permettant la détection des *hotspots*, or ces seuils varient suivant les galaxies. Une autre difficulté est de définir la taille des *hotspots*. Car si les *hotspots* sont des phénomènes généralement isolés et de petite taille comparée à celle de la galaxie, certaines galaxies contiennent des hotspots qui apparaissent sur presque la totalité de sa surface.

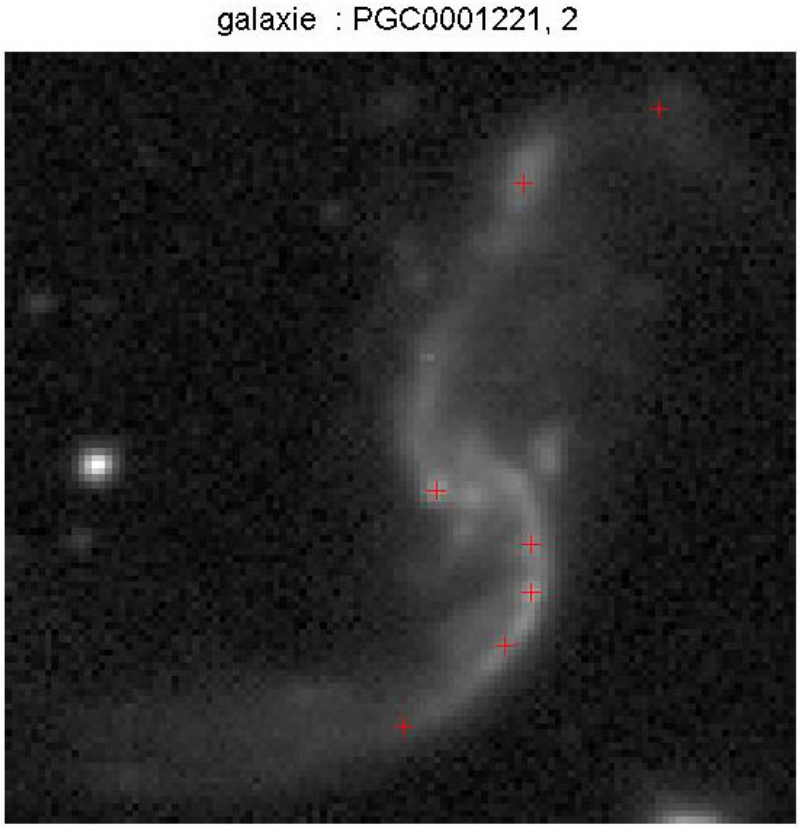


FIG. 4.1 – Exemple de la détection de hotspots par la recherche des maxima

4.2 Perspective

Comme nous l'avons vu précédemment, la détection des *hotspots* ne peut pas se faire par une simple analyse de la valeur des niveaux de gris de l'image. Une future version du programme nfigi, qui permet de nettoyer les images, permettra peut-être d'utiliser une telle méthode, mais pour le moment, le nettoyage des images entraîne trop de dégradations. Un outil d'analyse de signaux semble intéressant pour étudier les *hotspots* : l'analyse multifractale. Née dans les années 1980, elle a servi à expliquer des observations effectuées sur des signaux de turbulence. L'analyse multifractale dans le traitement des images consiste à définir des mesures à partir des niveaux de gris de l'image, à en calculer le spectre, puis à traiter l'information sur la régularité à la fois locale et globale qui en résulte. Le critère de régularité local est obtenu en calculant l'exposant local de Hölder. En calculant cet exposant sur une image, on est capable de dresser la carte des iso-hölder (point ayant la même valeur pour l'exposant de Hölder). A partir de ces iso-hölder, on peut calculer leur dimension de Hausdorff et obtenir le spectre de singularité de l'image. A partir de ce spectre, on effectue une segmentation de l'image, la dimension de Hausdorf est à valeur entière et renseigne sur l'appartenance d'un point à une région homogène, à un contour régulier, à un contour irrégulier, ... ce qui semble adapté à la détection de hotspots dans la galaxie, car cela permettrait de tester l'homogénéité de différentes parties de la galaxie.

Conclusion

Tout au long du stage, l'étude a porté sur deux types d'objets présents dans les galaxies : la floculence et les hotspots. L'hypothèse d'assimiler la floculence à une texture qu'on pourrait reconnaître semblait se justifier au vu des premiers résultats sur le découpage des galaxies en trois classes. Mais en testant sur des galaxies entières, les résultats montrent que la classe des zones de galaxies non floculentes est très mal définie et empêche de pouvoir faire une estimation réaliste de la floculence. La deuxième méthode fait intervenir moins de possibilités d'erreurs dans son processus. En effet la première méthode se base sur une classification manuelle en fonction de la classification des astronomes. Pour la deuxième méthode, l'appartenance à une classe s'effectue à l'aide de critères mathématiques et les taux de floculence sont directement reliés aux caratéristiques de chaque galaxie. Les classes les mieux reconnues sont celles de taux 0 et 4. Les classes intermédiaires semblent être plus difficiles à discriminer sans connaissance sur la galaxie. Le type de galaxies serait un moyen pour fournir un taux d'erreurs en complément des taux estimés en établissant des statistiques sur une large base de données. La définition de la base d'apprentissage peut aussi améliorer les résultats. En effet la détection des évènements rares ou *outliers* peut fournir une meilleure définition des classes et ainsi un meilleur apprentissage. Enfin l'utilisation des intervalles de confiance n'a pas été incluse dans le processus de classification mais la connaissance à priori d'un poids pour chaque exemple devrait améliorer la classification. Un autre aspect important des méthodes mises en oeuvre est le temps de calcul. En effet la quantité d'images de galaxies présentes dans les catalogues impose des temps très brefs pour le traitement de chaque galaxie. Or le filtrage de Gabor prend du temps, mais il n'a pas été optimisé. Le calcul des différents filtrages peut être optimisé si les algorithmes s'exécutent en parallèle.

Annexes : Agrandissement des images de galaxies

Revenir à l'image originale

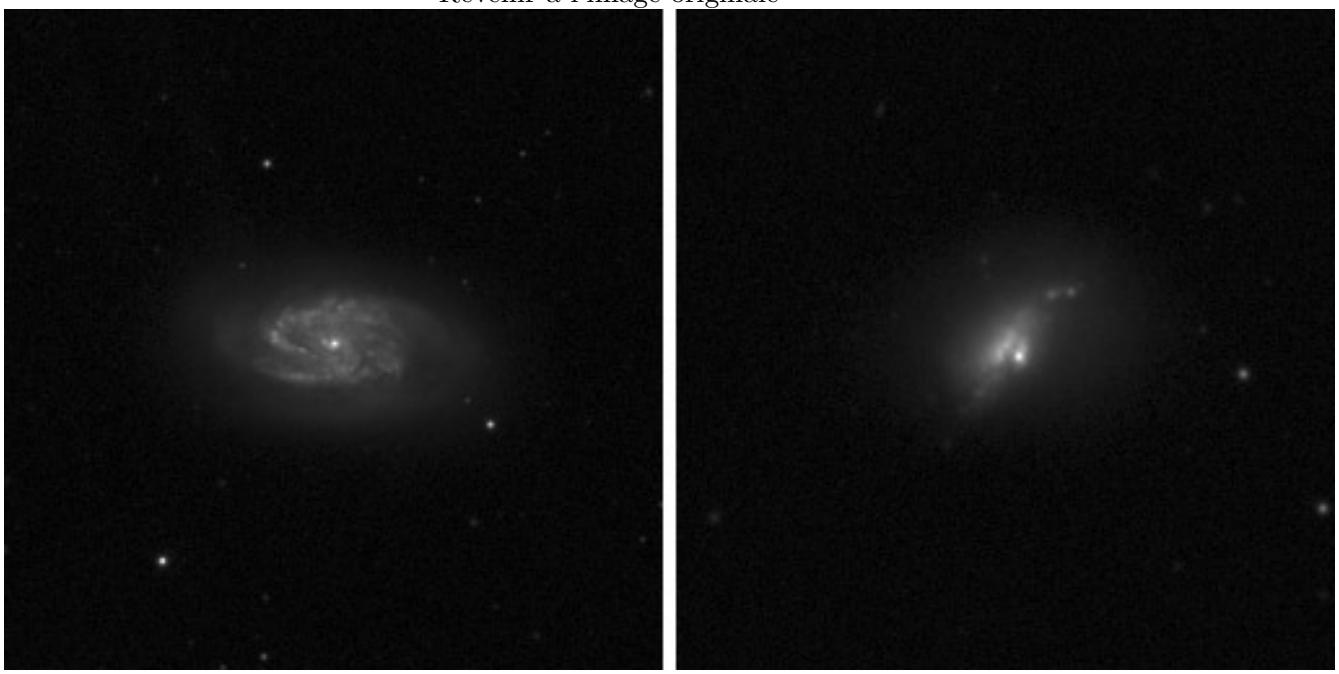


FIG. .2 – galaxies présentant un taux élevé de floculence (à gauche) et un taux élevé de hotspots (à droite)

Revenir à l'image originale

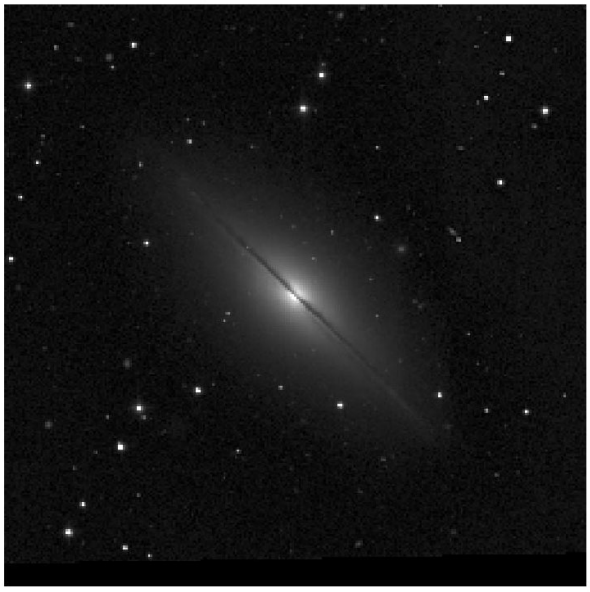


FIG. .3 – PGC00218

Revenir à l'image originale

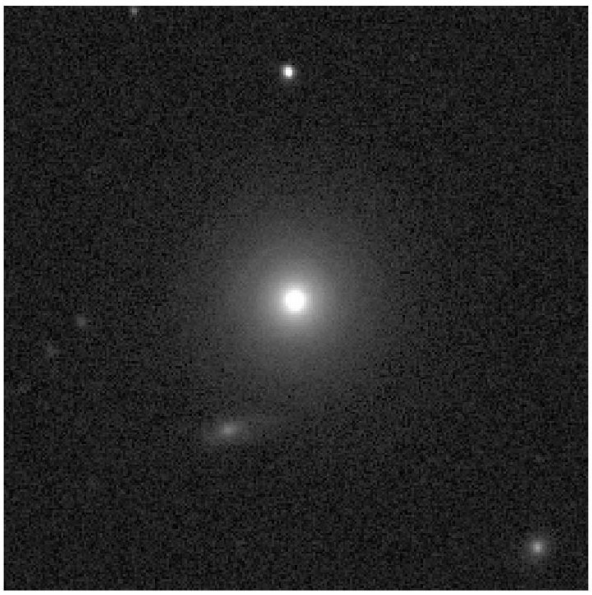


FIG. .4 – PGC08718

Revenir à l'image originale

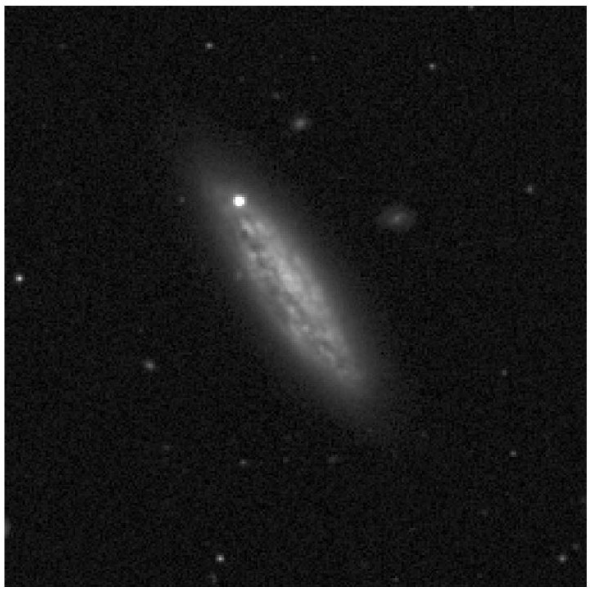


FIG. .5 – PGC10065

Revenir à l'image originale

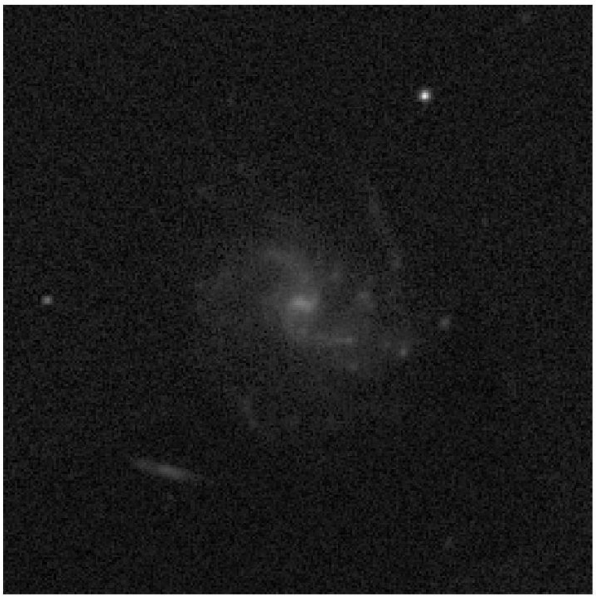


FIG. .6 – PGC27993

Revenir à l'image originale

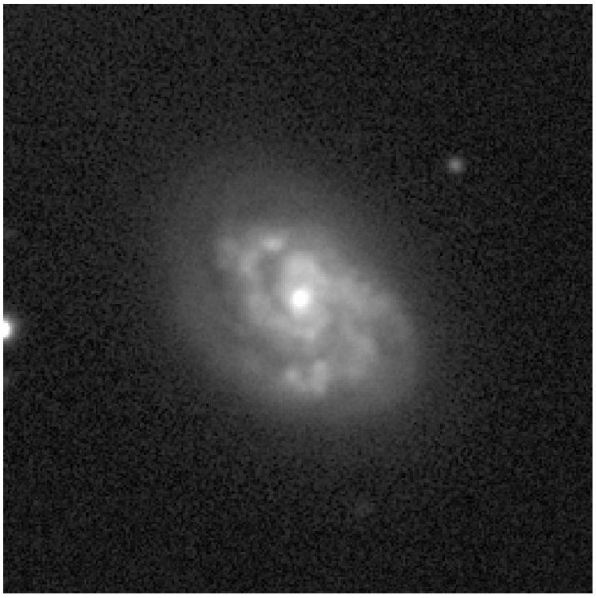


FIG. .7 – PGC29209

Revenir à l'image originale

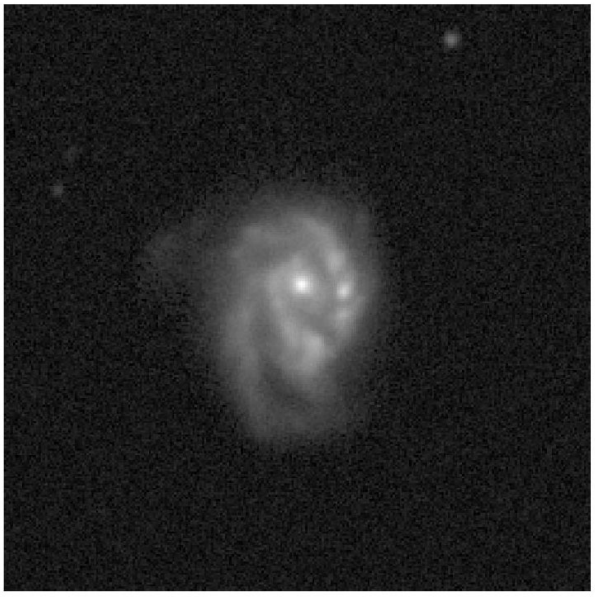


FIG. .8 – PGC30136

Revenir à l'image originale

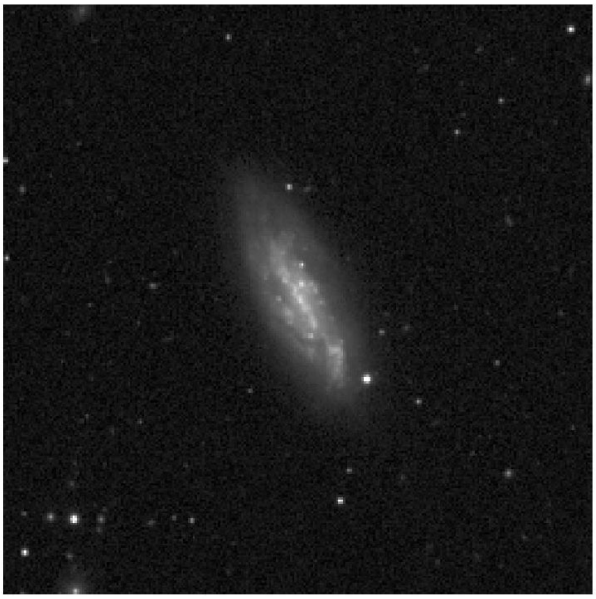


FIG. .9 – PGC39925

Revenir à l'image originale

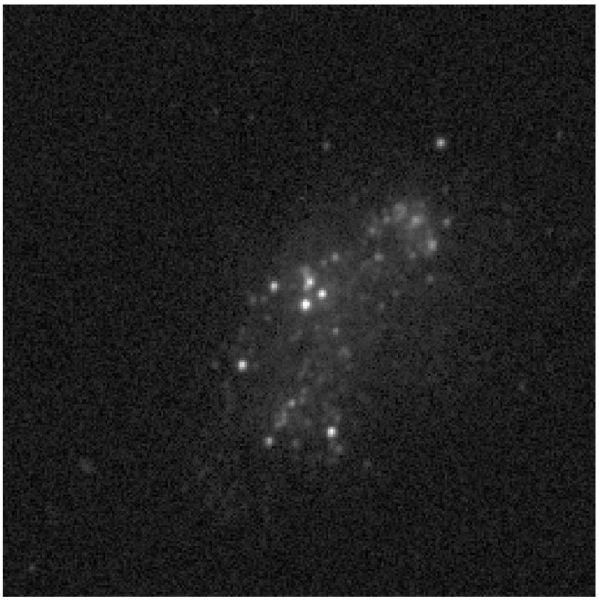


FIG. .10 – PGC44491

Bibliographie

[1] A. Baillard, E. Bertin, M. Dantel-Fort, L. Domisse, F. Magnard, J.-C. Malapert, C. Marmo, Y. Mellier, H.J. McCracken, M. Schultheis, G. SÈmah, G. Tissier. *Terapix Public Activities 2004 - 2007*, 2007. 12

[2] Nancy G.Roman,François Ochsenbein. *Centre de données astronomique de Strasbourg* [en ligne]. Disponible sur http ://cdsweb.u-strasbg.fr/cgi-bin/Cat ?VII/155(consulté le 26.03.2007) 9

[3] Roberto G. Abraham , Sidney Van Den Bergh, Preethi Nair. *A new approach to galaxy ,morphology : I. Analysis of the sloan digital sky survey early data release.* The Astrophysical Journal, 05/2003, Volume 588, Issue 1, pp. 218-229. 18, 28

[4] Christopher J. Conselice. *The relationship between stellar light distributions of galaxies and their formation histories*, The Astrophysical Journal Supplement Series, 07/2003, Volume 147, Issue 1, pp. 1-28. 18

[5] Jennifer M. Lotz,Joel Primack,Piero Madau. *A new non-parametric approach to galaxy morphological classification*,The Astronomical Journal, 2004, Volume 128, Issue 1, pp. 163-182. 19

[6] Chisato Yamauchi, Shin-ichi Ichikawa, Mamoru Doi, Naoki Yasuda, Masafumi Yagi, Masataka Fukugita, Sadanori Okamura, Osamu Nakamura, Maki Sekiguchi, Tomotsugu Goto. *Morphological Classification of Galaxies using Photometric Parameters : the Concentration index versus the Coarseness parameter*, The Astronomical Journal,2005, Volume 130, Issue 4, pp. 1545-1557. 19

[7] A. K. Jain, F. Farrokhnia. *Unsupervised texture segmentation using Gabor filters*, Pattern Recognition, 1991, vol. 24, no. 12, pp. 1167-1186. 25

[8] B.S. Manjunatah, W.Y. Ma. *Texture features for browsing and retrieval of image data*, IEEE Transactions on pattern analysis and machine intelligence, 1996, vol. 18(8), pp. 837-842. 25

[9] Marine Campedel, Bin Luo, Henri Maître, Eric Moulines, Michel Roux, Ivan Kyrgyzov. *Indexation des images satellitaires : Détection et évaluation des caractéristiques de classification*, 2004 21

[10] Ivan O. Kyrgyzov, Olexiy O. Kyrgyzov, Henri Maître, and Marine Campedel.*Kernel MDL to Determine the Number of Clusters*,2007 32

[11] D.N. Gujarati. *Basic econometrics*, 4th ed, McGraw-Hill, 2002 29

[12] Siddheswar Ray and Rose H. Turi. *Determination of number of clusters in k-means clustering and application in colour image segmentation*, Proceedings of the 4th International Conference on Advances in Pattern Recognition and Digital Techniques, pp. 137–143, 1999. 32

[13] C. Cortes and V. Vapnik. *Support-vector networks*, Machine Learning, 20(3), pp 273-297, September 1995. 34

[14] Mohamadally Hasan Fomani Boris. *SVM : Machines à Vecteurs de Support ou Spárateurs à Vastes Marges*, tutoriel, janvier 2006 34