République Algérienne Démocratique et Populaire

N° d'ordre :
N° de série :

Ministre de l'Enseignement Supérieur et de la Recherche Scientifique

UNIVERSITE ECHAHID HAMMA LAKHDAR D'EL-OUED

FACULTE DES SCIENCES EXACTES

Mémoire de fin d'étude

Présenté pour l'obtention du diplôme de

MASTER ACADEMIQUE

Domaine : Maths et informatique

Filière : informatique

Spécialité : systèmes distribué et intelligence artificielle

Présenté par :

HARIZ BEKKAR OUACILA

BELBEY MEDJDA

Thème

Analyse de la texture des images mammaires par une fusion des lois de Zipf et des Ondelettes pour la classification des tumeurs mammaires via l'analyse en composantes principale

Encadré par :

Dr. Hamoud Meriem

2017-2018

Dédicace

Nous dédions cette mémoire à A ma très chère mère Affable, honorable, aimable: Tu représentes pour moi le symbole de la bonté par excellence, la source de tendresse et l'exemple du dévouement qui n'a pas cessé de m'encourager et de prier pour moi. Ta prière et ta bénédiction m'ont été d'un grand secours pour mener à bien mes études. Aucune dédicace ne saurait être assez éloquente pour exprimer ce que tu mérites pour tous les sacrifices que tu n'as cessé de me donner depuis ma naissance, durant mon enfance et même à l'âge adulte. Tu as fait plus qu'une mère puisse faire pour que ses enfants suivent le bon chemin dans leur vie et leurs études. Je te dédie ce travail en témoignage de mon profond amour. Puisse Dieu, le tout puissant, te préserver et t'accorder santé, longue vie et bonheur.

Remerciement

Avant tout nous remercions bien dieu qui nous amené à ces résultats, et après nous remercions la Docteur Hamoud Meriem qui a beaucoup encouragé avec patience et a nous dirigé au bien et surtouts ces conseils et les remarques

illimité.

Qu'il trouve ici le témoignage de ma profonde gratitude.

Et nous remercions également les membres de jury qui nous a compagne pour avoir accepté d'évaluer ce travail et pour toutes leurs remarques et critique.

Et aussi nous remercions beaucoup mes parents qui a nous idées qui dieu les bénisse et tous qui m'ont donnés son confiance.

Avec beaucoup des remerciements à mes enseignants qui m'ont initié aux valeurs authentiques, en signe d'un profond respect et d'un profond amour !!

Merci à vous tous.

الملخص

ان اختيار العلاقات عدم الخطية له أهمية بالغة في تطوير أدوات قوية لتحليل الصور و الرؤية عبر الكمبيوتر.
مشكلتنا البحثية هي تطبيق قوانين القوة: زيب ا ف وزيب ا ف معكوس لتحليل صور الثدي, تميز قوانين القوة التعقيد الهيكلي لبنية الصورة عبر نمذجة التوزيع الإحصائي لوتيرة ظهورالأنماط حسب التوزيع على قانون القوة. ثم دمجنا الأوصاف التي تم الحصول عليها مع تلك التي تم إنشاؤها بعد تحليل الصور الثديية من موجات هآر.

تم اقتراح نظام في مجال المساعدة للتشخيص الطبي لسرطان الثدي عن طريق مساعدة الكمبيوتر. في الواقع، اقترحنا نظاما للكشف عن الأورام استنادا إلى قوانين دمج زيب أف و زيب أف معكوس ونظام تصنيف الأورام عن طريق حساب واصفات جديدة للبنية، لعملية التوصيف .بعد ذلك، اقترحنا نظام لفهرسة التصوير بالأشعة السينية ونظام البحث الذي يقوي أداء التشخيص بمساعدة الكمبيوتر في مرحلة تقديم التشخيص إلى أطباء الأشعة. وبالفعل ، فهم أكثر ثقة في حكم التشخيص المعتمد على الحالة الذي يتم إرجاعه من خلال تصنيف قائم على مطابقة القالب على غرار الحالة التي يتم تحليلها بدلا من النتيجة المجردة, التي تم تشخيصها مسبقا الناتجة عن المصنف. كما أعطى تقييم العمل المقترح اداء جيد و نتائج مشجعة.

كلمات البحث: تحليل الصور ,الرؤية عبر الكمبيوتر ، تصنيف , الفهرسة ، قانون زيب اف معكوس ، قانون زيب اف ، مويجات هآر , مطابقة القالب .

Résumé

Le choix de la non-linéarité est d'un intérêt crucial dans le développement de puissants outils d'analyse d'image et de vision par ordinateur. Dans ce sens, notre problématique de recherche consiste à appliquer les lois puissance : Zipf et Zipf inverse à l'analyse des images mammaires. En effet, les lois de Zipf caractérisent la complexité structurelle de la texture d'image par la modélisation de la répartition statistique de la fréquence d'apparition des motifs selon une distribution en lois puissance. En outre, nous fusions les descripteurs obtenus avec ceux générés suite à l'analyse des images mammaires par les ondelettes de Haar. Nous appliquerons une analyse en composantes principales (ACP) pour réduire le nombre de descripteurs à mesure.

Nous avons proposé un système d'indexation et de recherche des mammographies par le contenu (CBMIIR) qui renforce la performance du diagnostic assisté par ordinateur au niveau de l'étape de la présentation du diagnostic aux radiologues. En effet, ces derniers sont plus confiants d'un jugement de diagnostic basé sur des cas renvoyé par une classification basée template-matching, diagnostiqués préalablement, similaires au cas en cours d'analyse plutôt que le résultat abstraits généré par un classifieur. L'évaluation des travaux proposés a donnée des performances encourageantes.

Mots clés : Analyse d'image, Vision par ordinateur, Loi de Zipf, Loi de Zipf inverse, Classification, Indexation, Recherche par le contenu, ondelette de Haar,ACP, template-matching,

Abstract

The choice of the non-linearity is of crucial interest in the development of powerful tools for image analysis and computer vision. In this sense, our research problematic is to apply the power laws: Zipf and inverse Zipf for mammogram images analysis. Indeed, the laws of Zipf characterize the structural complexity of the image texture by modeling the statistical distribution of patterns frequency of appearance as power law distribution. In addition, we have performed a fusion of the obtained texture features with those generated once applying Haar wavelet transform for mammogram images analysis .We will apply a PCA principal component analysis to reduce the number of descriptors.

Subsequently, we have proposed a content based mammogram image indexing and retrieval system (CBMIIR) that boosts the performance of a computeraided diagnosis (CADx) at the stage of providing the diagnostic to radiologists. Indeed, radiologists feel more confident in their diagnosis decision based upon case-adaptive classification via the template-matching technique, where similar known cases, to the one under analysis, are retrieved and displayed from indexed databases; rather than the abstract result generated by a classifier. The evaluation of the proposed systems has given encouraging performance.

Keywords

Image analysis, computer vision, Zipf's law, inverse Zipf's law, Classification, Indexing, Retrieval by content, Haar wavelet transform, PCA ,template-matching.

Table de matière

Liste des figures

Introduction générale

Le cancer du sein demeure un problème de santé publique, et le plus fréquent des cancers touchant la femme. La mammographie est le pivot de l'exploration de la pathologie mammaire, cependant, l'incidence de cette dangereuse maladie ne cesse de croitre et les radiologues se trouvent face à une quantité surprenante de mammographie à analyser et interpréter. Evidemment plusieurs facteurs entrent en jeux comme la fatigue visuelle touchant les radiologues et générant des difficultés à l'interprétation avec exactitude d'une mammographie.

A cet effet, le diagnostic assisté par ordinateur a pris part par le développement des systèmes d'aide au diagnostic médical du cancer du sein pour fournir un deuxième avis aux radiologues durant la réalisation de leur diagnostic.

La problématique tirée et discutée pour l'élaboration de ce travail consiste en l'analyse des images mammaires par des approches non linéaires et les proposer comme alternatives aux approche purement linéaires qui montrent vite leurs limites, spécialement, dans le cas des images vu leur structure complexe.

Nous nous intéressons aux lois puissance : Zipf et Zipf inverse pour l'analyse et la caractérisation de la texture des images mammaires. Voir plus, nous proposons une fusion des lois de Zipf et de Zipf inverse avec l'approche des ondelettes de Haar via l'analyse en composante principale pour bénéficier de l'apport complémentaire qui peut être obtenu par la fusion de ces deux approches. En effet, notre objectif est d'essayer d'améliorer les résultats obtenus par l'application des lois de Zipf et de Zipf inverse individuellement.

Dans ce contexte, nous développons un système d'indexation et de recherche des mammographies par le contenu (CBMIIR) pour l'aide au diagnostic médical du cancer du sein. En effet, les radiologues jugent les systèmes d'aide au diagnostic médical du cancer du sein assisté par ordinateur et se basant sur la sortie numérique d'un classifieur (CADx) comme ambigus et il leur est difficile d'interpréter un résultat abstraits. Ils se sentent à l'aise dans l'interprétation d'un aide généré à base de cas déjà diagnostiqués et stockés dans une base de données comme est le cas des systèmes CBMIIR.

Nous pouvons affirmer que notre système aboutit à un taux de classification encourageant de l'ordre de 82.5%.

Ce mémoire se compose de quatre chapitres comme suit :

Chapitre I : Analyse d'image et vision par ordinateur

Dans ce chapitre, nous avons présenté la formalisation générale d'analyse d'image et de la vision par ordinateur, par la suite, nous avons évoqué les descripteurs de la texture du fait de leur pertinence en vision par ordinateur.

Chapitre II : Les lois puissance

Ce chapitre est consacré aux lois puissance, en effet, nous nous intéressons aux lois puissance de type Zipf et Zipf inverse. Spécialement, l'application de ces lois dans l'analyse d'image et la vision par ordinateur.

Chapitre III : Aide au diagnostic médical du cancer du sein assisté par ordinateur

Ce chapitre détaille l'aide au diagnostic médical du cancer du sein assisté par ordinateur.

Chapitre IV : Conception et implémentation

Durant ce chapitre, nous avons présenté les diverses phases de réalisation de notre système d'indexation et de recherche des mammographies par le contenu ainsi que ses fonctionnalités.

Ce mémoire sera achevé par une conclusion générale et des perspectives.

--_

Chapitre I : Analyse d'image et vision par ordinateur

I.1 Introduction

Dans le domaine vaste du traitement d'image, la vision par ordinateur a pris une part de plus en plus importante engendrant le fait que les opérations de base telles que la segmentation et la classification d'image se perfectionnent, spécialement, dans le domaine de l'imagerie médicale. La vision par ordinateur englobe deux parties distinctes : une de "bas niveau" et une autre de "haut niveau". La vision de bas niveau nécessite très peu d'informations sur le contenu des images et concerne des problèmes liés à la numérisation et filtrage d'images, au codage et la compression de données pour le stockage ou la transmission… etc. D'autre part, la vision de haut niveau qui fonctionne en aval de celle de bas-niveaux et encapsule la reconnaissance d'objets ainsi que l'interprétation des scènes et les processus cognitifs de façon générale.

Nous exposons dans le premier chapitre les notions de base de l'analyse d'image et de la vision par ordinateur. Par la suite, nous évoquons les descripteurs de la texture du fait de leur pertinence en vision par ordinateur. En effet, notre travail porte sur la caractérisation de la texture par une fusion des lois puissance : Zipf et Zipf inverse et les ondelettes de Haar pour l'aide au diagnostic médical du cancer du sein assisté par ordinateur.

I.2 Lien entre l'analyse d'image et la vision par ordinateur

La vision par ordinateur a pour but de reproduire quelques fonctionnalités de la vision humaine au travers de l'analyse d'images. C'est un problème difficile en raison du fait que l'information disponible : des images 2D fournies par des capteurs (CCD, ...), correspond à une projection du monde 3D. La projection 3D-2D entraine une perte d'informations importante, de plus l'information disponible n'est pas parfaite (numérisation des capteurs, déformation des objectifs, bruitages) (Edmond Boyer.2016).

La perception visuelle par ordinateur vise à trouver une relation reliant une image d'entrée aux modèles du monde réel (Hamoud M.2015).

I.3. Analyse d'image

Une image numérique est avant tout un signal 2D (x, y), souvent, cette image représente une réalité 3D (x ,y, z) constituée d'une grille rectangulaire d'échantillonnage dont les constituants sont des pixels portant des informations sur l'intensité lumineuse des différents lieux au sein de l'image. En effet, une image numérique consiste en une matrice bidimensionnelle, dont les éléments sont des nombres naturels correspondant à des niveaux de quantification dans l'échelle de l'intensité lumineuse. Par contre pour l'humain une image désigne plusieurs informations sémantiques dont faut interpréter le contenu au-delà de la valeur des nombres (TSAL1.INSA.Lyon) (Hamoud M.2015).

L'analyse d'images a pour objectif de décomposer ("décortiquer") une image afin d'en extraire des informations pertinentes tout en fournissant une description quantitative de l'image. De multiples éléments s'impliquent:

- Des primitives images : pixels, points d'intérêts, segments, contours.
- Des caractéristiques photo-métriques : niveaux de gris, de couleurs.
- Des caractéristiques géométriques : caméras, mouvements.
- Des caractéristiques statistiques (Philippe B et al.2017).

Nous allons évoquer les niveaux d'analyse d'une image comme suit :

I.3.1. Analyse de bas niveau d'image

D'après (Sonka et al.2008), le traitement de bas niveau d'images consiste en les opérations effectuées sur les images au plus bas niveau d'abstraction citons l'exemple de l'élimination du

bruit au sein de l'image. De plus, le traitement de bas niveau a tendance à diminuer généralement les informations contenues au sein de l'image par la suppression des distorsions. La segmentation d'image est une étape clé du processus d'analyse d'image, en effet, l'ordinateur sépare les objets de l'arrière-plan de l'image désignant la segmentation partielle. Par la suite, seuls les indices qui aideront l'analyse de haut niveau sont extraits (Hamoud M.2015).

I.3.2. Analyse de haut niveau d'image

Selon (Sonka et al. 2008) , l'analyse de haut niveau d'image s'appuie sur des connaissances relatives au contenu de l'image. En effet, elle fait intervenir les méthodes issues d'intelligence artificielle du fait que la vision par ordinateur de haut niveau s'inspire de la cognition humaine et la capacité à prendre des décisions depuis l'information contenue dans l'image (Hamoud M.2015).

I.4. L'image médicale

L'imagerie médicale est une image en niveaux de gris représentant différents organes de l'organisme, permettant de les examiner sans la nécessité d'opérer le patient. L'utilité de l'image médicale est d'aider le médecin lors du diagnostic, ou bien le chirurgien lors de la réalisation d'un geste opératoire. Suite à l'amélioration de l'image médicale par le rehaussement du contraste, l'élimination du bruit ou la mise en évidence des détails; elle pourra détecter et localiser le positionnement des tumeurs ainsi que la mesure de leurs dimensions et volumes (Mekki et al 2017).

I.5. Analyse de la texture d'une image

L'analyse de la texture consiste en un ensemble de techniques quantifiant les différents niveaux de gris présents dans une image en termes d'intensité et de distribution dans le but de calculer un certain nombre de descripteurs caractérisant la texture en vue d'analyse. Notons que le but de l'analyse de texture est de formaliser une description attributs qui serviraient à l'identifier (Chekkaf et al 2012).

I.5.1. Définition de la texture

La texture (Sheni et al. 1996) (Aptoulai et al. 2011) présente une caractéristique qui réfère à des propriétés représentant la surface ou la structure d'un objet en la voyant comme une répétition d'un motif. En fait, il n'existe pas de définition précise de la texture malgré son omniprésence dans les images (Hamoud M.2015).
Structurée et aléatoire, une texture se décrit spatialement ou statistiquement, en général, nous qualifierons une texture de fine ou grossière, plus ou moins homogène, rugueuse, plus ou moins régulière, dense, linéaire, isotrope ou directionnelle, tachetée, marbrée…etc.(Hamoud M.2015).

I.5.2. Les méthodes d'analyse de la texture

Nous distinguons 4 classes (modèles) de textures (Tuceryan et al.1998) :

- **Les méthodes statistiques**

Les méthodes statistiques étudient les relations entre un pixel et ses voisins et définissent des paramètres discriminants de la texture en se basant sur des outils statistiques. Ces modèles statistiques sont efficaces pour de nombreuses textures naturelles ayant des primitives discernables ainsi que pour la caractérisation des structures fines, sans régularité apparente. Notons que plus l'ordre de la statistique est élevé et plus le nombre de pixels mis en jeu est important. Parmi ces méthodes, citons la méthode basée sur les matrices de cooccurrences, celle des matrices de longueurs de plage (Bagadi A.2016). La méthode basée sur les matrices de cooccurrence, proposée par Haralick (Haralick. 1979), constitue à explorer les dépendances spatiales des pixels en construisant d'abord une matrice de cooccurrence basée sur l'orientation et la distance entre les pixels de l'image. A partir de ces matrices, nous pouvons extraire des caractéristiques de la texture, comme le contraste, l'entropie ou la différence inverse des moments. Ces modèles sont efficaces pour de nombreuses textures naturelles qui ont des primitives discernables ainsi que pour la caractérisation des structures fines. Plus l'ordre de la statistique est élevé et plus le nombre de pixels mis en jeu est important (Hamoud M.2015).

- **Les méthodes structurelles**

Les méthodes structurelles permettent de décrire la texture en définissant les primitives et les "règles" d'arrangement qui les relient. Conduisant leur application sur les textures aléatoires.

Ces modèles purs structurels de texture sont les meilleurs pour la caractérisation des textures où il y a une grande résolution de l'image (Bagadi A.2016).
D'après Haralick (Haralick, 1982), ces modèles présument que les textures se forment de primitives engendrant des motifs répétitifs dans les directions de l'espace de manière quasi-régulière et tentent de décrire les règles qui régissent leur organisation spatiale (Hamoud M.2015).

- **Les méthodes basées sur des modèles**

Ces méthodes d'analyse de la texture à base de modèles se basent sur le principe de la construction d'un modèle d'image permettant de décrire une texture ainsi que d'en générer (Hamoud M.2015). Nous distinguons deux méthodes couramment utilisées dans cette catégorie qui sont les fractales qui furent utilisées pour mesurer la rugosité d'une texture et la répétitivité (spatiale ou à différentes résolutions) d'un motif (Journet et al. 2007) et les champs de Markov (Hamoud M.2015).

- **Les méthodes spatio-fréquentielles**

Ces méthodes décrivent une texture à partir de ses caractéristiques fréquentielles. En effet, certaines sont basées sur le filtrage spatial (domaine spatial), tandis que d'autres extraient les descripteurs de texture à partir de la représentation de l'image dans le domaine de Fourier. Notons qu'il existe des méthodes qui exploitent à la fois le domaine spatial et le domaine fréquentiel pour le calcul des descripteurs (Bagadi A.2016).
D'autres méthodes préservent à la fois les informations globales et locales existent, qui s'articulent autour des transformations spatio-fréquentielles dans le but de caractériser la texture à différentes échelles. Citons la transformée en ondelettes, la transformée de Gabor, la transformée de Hermite (Bagadi A.2016).

I.5.3. Les descripteurs de la texture

La description de la texture par un observateur humain est qualitatives et sera comme suit: fine, grossière, foncée, grainée, lisse… etc. Une quantification de ces propriétés a été réalisée pour la génération de descripteurs numériques mathématiques servant à identifier la texture précisément dans le but de rendre la reconnaissance par machine possible (Hamoud M.2015). Les chercheurs en vision par ordinateur se sont basés sur des descripteurs de la texture tels que la régularité, la répétitivité et le contraste des textures, citons l'exemple de la matrice de co-

occurrence des niveaux de gris de l'image (Haralick) qui indique, dans une image le nombre d'apparitions de couples de pixels ayant des niveaux de gris (i, j) selon une direction et un déplacement donné (d = (dx,dy)). Des descripteurs calculés sur la matrice de co-occurrence permettent de caractériser la régularité, la répétitivité et le contraste des textures (Hamoud M.2015).

Nous distinguons une autre famille de descripteurs basée sur les filtres de Gabor où les descripteurs sont extraits pour certaines bandes de fréquence et dans des directions choisies. Les performances des descripteurs de Gabor sont dépendantes du contexte dans lequel apparaît la zone à extraire. Cependant, en analysant les descripteurs de Gabor, on observe qu'ils permettent la distinction de différents types de graphiques (Mehri et al. 2013).

I.6. La vision par ordinateur

I.6.1. Qu'est-ce que la vision ?

L'espace du monde qui nous entoure présente une structure tri dimensionnelle. Si nous demandons à un humain de décrire ce qu'il aperçoit, il n'éprouvera aucune difficulté à nommer les objets qui l'entourent. L'information qui est réellement disponible sur la rétine des yeux est une collection de points environ un million en chaque point ou pixel il y a tout simplement une information qui donne une indication quant à la quantité de lumière et la couleur qui proviennent de l'espace environnant et qui ont été projetées à cet endroit de la rétine (Radu H.2011).

La rétine et par ses propres connaissances reconnait les objets via un processus d'interprétation qui fait partie du système de vision. Le système de vision doit fournir les connaissances nécessaires pour permettre une interprétation non ambigüe (Radu H.2011).

I.6.2. Comprendre la vision par ordinateur

La vision par ordinateur est une discipline dont les premières bases théoriques ont vu le jour dans les années 60. Depuis, un spectre très large d'applications industrielles, militaires aérospatiales et médicales envisageaient d'incorporer la vision par ordinateur et donc de dépasser le cadre relativement restreint des laboratoires de recherche.

La vision permet aux humains de percevoir et de comprendre le monde qui les entoure, de l'autre part la vision par ordinateur reproduit l'effet de la vision humaine en percevant électroniquement une image tout en simulant ou imitant les systèmes humains. Cependant, donner à l'ordinateur la possibilité de voir n'est pas une tâche facile, en effet, le passage du monde réel (3D), au monde (2D) est en ordre du fait que lorsque les ordinateurs tentent

d'analyser des objets dans l'espace 3D, les capteurs visuels disponibles comme les caméras de télévision, génèrent des images bidimensionnelles (2D), et suite à cette projection en un nombre inférieur de dimensions, une énorme perte d'informations est encaissée (Hamoud M.2015).

Les techniques de vision par ordinateur se basent sur les résultats issus des méthodes de mathématiques, reconnaissance de formes, intelligence artificielle, psycho-physiologie, informatique, électronique.

L'utilisation des algorithmes ou des dispositifs d'analyse d'image numérique, est conditionnée par la prise en compte des principes de la perception humaine de l'image. Les descripteurs psychophysiques utilisés dans la vision humaine sont : la couleur, le contraste, les contours, la forme, la texture...etc. Nous allons nous intéresser à l'analyse de la texture vue son importance dans la vision par ordinateur. En effet, l'analyse de la texture est d'un grand intérêt en vision humaine ainsi qu'en psychophysique, en revanche, les caractéristiques de texture ne sont pas interprétées de façon évidente par la vision humaine. A cet effet, les chercheurs en analyse d'image et vision par ordinateur cherchent à caractériser cette texture à travers des descripteurs numériques pertinents et discriminants qui seront incontournables pour la détection des objets et des régions saillantes (Hamoud M.2015).

Dans ce mémoire, nous allons nous baser sur la caractérisation de la texture par une fusion des lois puissance : Zipf et Zipf inverse et les ondelettes de Haar afin de bénéficier de leur apport complémentaire dans un processus d'aide au diagnostic médical du cancer du sein assisté par ordinateur (Hamoud M.2015).

I.7. Conclusion

L’analyse d’image et la vision par ordinateur deviennent de plus en plus un domaine de recherche actif. En effet, l’analyse numérique de l'image en vue de son interprétation désigne l'outil principal de la vision par ordinateur.

A cet effet, nous nous basons sur cet enjeu pour la réalisation de ce mémoire. Effectivement, nous nous intéressons aux modèles de loi puissance de type: Zipf et Zipf inverse ainsi qu’à l’ondelette de Haar pour la caractérisation de la texture des zones d’intérêt extraites à partir d’images mammaires afin d’offrir un aide au diagnostic médical du cancer du sein assisté par ordinateur.

Chapitre II : Les lois puissance

II.1 Introduction

Dans divers domaines, des lois puissance ont été appliquées pour la modélisation de différents phénomènes.

A travers ce chapitre, nous évoquons les lois de puissance, en effet, nous nous intéressons aux lois puissance de type Zipf et Zipf inverse. À cet effet, nous exposons le processus d'application de ces lois dans l'analyse d'image et la vision par ordinateur qui fut notre domaine de recherche.

II.2. Qu'est-ce qu'une loi de puissance ?

Des répartitions statistiques en lois puissance ont été constatées dans de nombreux domaines. Les modèles de loi puissance les plus connus sont la loi de Pareto et les lois de Zipf et de Zipf inverse. En effet, elles sont omniprésentes dans l'économie, la démographie, la musique, l'internet ainsi que la génétique (Caron.2004).

La loi de puissance est une relation mathématique entre deux quantités. Si une quantité est la fréquence d'un événement et l'autre est sa taille, alors la relation est une distribution selon une loi puissance si les fréquences diminuent lentement lorsque la taille de l'événement augmente (Hamoud M.2015).

En science, une loi de puissance est une relation entre deux quantité x et y qui peut s'écrire comme suit :

$$Y = ax^{-b} \qquad \text{(II.1)}$$

Où a et b sont des constantes. Sa représentation dans un repère linéaire est mentionnée sur la figure II.1:

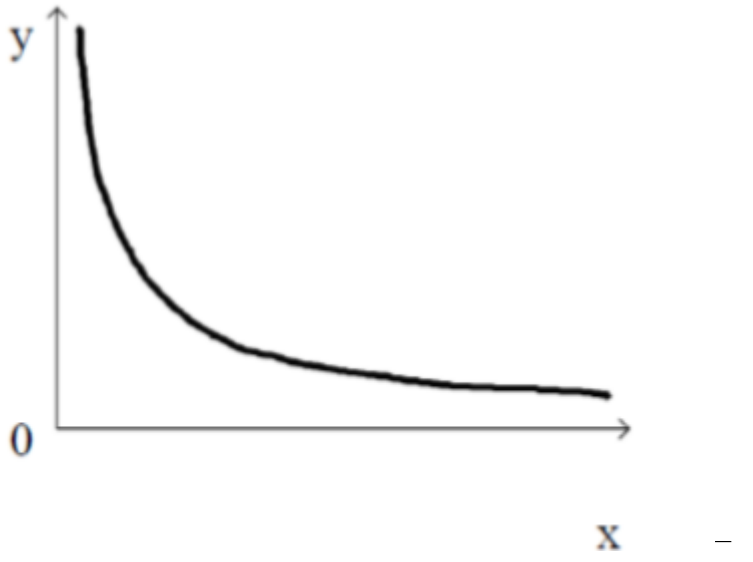


Figure II.1 Représentation d'une loi puissance dans un repère linéaire.

Lois puissance se représentent dans un repère bi-logarithme. En choisissant un tel repère, nous obtenons une distribution en loi puissance sous la forme d'une droite. La figure II.2 le démontre :

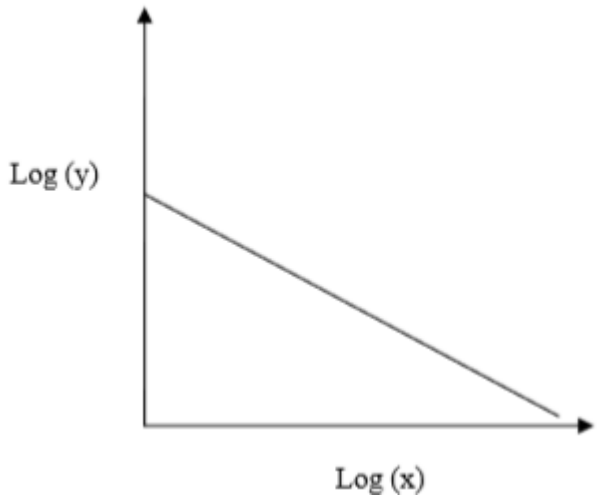


Figure II.2 Représentation d'une loi puissance dans un repère bi-logarithmique.

II.3. Les principales lois puissance

II.3.1. La loi de Pareto

L'économiste italien Pareto est le premier à appliquer la loi de puissance sur le phénomène économique, ceci en 1897. En 1906 il étudie la répartition des revenus des habitants de divers pays industrialisés et constate que 80% des richesses appartiennent à 20% de la population et ceci, quel que soit le pays observé (Hamoud M.2015).

La loi de Pareto ou loi des 80/20, est une loi empirique inspirée par les observations suivantes: 80% des effets est le produit de 20% des causes, 80% des résultats (positifs ou négatifs) ne sont obtenus que par 20% du travail, 20% des produits fabriqués par une entreprise représentent 80% du volume de production de celle-ci, 20% du réseau routier supporte 80% du trafic, 20% des individus sont responsables de 80% des problèmes(Hamoud M.2015).

II.3.2. La loi de Zipf

En 1949 fut la mise en évidence de la loi de Zipf par le linguiste américain George k. Zipf, il a constaté que le mot le plus fréquemment utilisé revenait dans le texte étudié en moyenne tous les dix mots, le second mot le plus fréquemment utilisé dans l'ordre des fréquences, revenait

tous les vingt mots et le troisième mot, tous les trente mots. Ceci n'est observé que si nous classons les mots par ordre décroissant de leur fréquence d'apparition (Hamoud M.2015).En effet, selon la loi de Zipf, les fréquences d'apparition des mots classés par ordre décroissant de leur fréquence d'apparition s'organisent suivant une loi puissance (Caron. 2004) . En effet, si nous désignons par F la fréquence d'apparition du mot de rang r dans la suite, la relation est exprimée par la formule suivante :

$$F = a\, r^{-\alpha} \qquad \text{(II.1)}$$

Dans cette formule, a et α sont deux constantes positives. Selon cette approximation, le produit rang-fréquence est constant.

II.3.3. La loi de Zipf inverse

Elle présente la seconde loi puissance qui a été mise en évidence par Zipf à partir de ses travaux sur la langue anglaise (Caron. 2004). Au contraire de la loi de Zipf précédemment évoquée et décrivant la répartition statistique des mots, cette seconde loi décrit la répartition statistique des fréquences des mots.

Selon la loi de Zipf inverse, le nombre I(f) de mots apparaissant f fois dans le texte est donné par la formule suivante :

$$I(f) = A\, f^{-\gamma} \qquad \text{(II.2)}$$

Dans cette formule, A et γ sont des constantes.

II.4 Application des lois puissances en analyse d'image

Il est connue que ces lois ne se vérifient pas uniquement aux messages linguistiques, mais également pour les messages iconiques y compris les images.

II.4.1 Les lois de Zipf et de Zipf inverse pour l'analyse des images numériques

- **La loi de Zipf**

Rappelons dans sa définition que la loi de Zipf décrit la répartition des n-uplets de symboles d'un ensemble suivant leur fréquence d'apparition. Nous considérons que l'ensemble de symboles ou de n-uplets dans le cas des images sera des motifs d'image (Caron. 2004).

Un motif est une matrice carrée de pixels adjacents de l'image centrée sur un pixel. Le motif M2k+1(i,j) de taille 2k+1 centré sur le pixel v(i,j) est défini ainsi pour k > 1 :

$$M_{2k+1} = \begin{cases} v(i-k, j-k) \dots \dots \dots v(i-k, j+k) \\ v(i,j) \\ v(i+k, j-k) \dots \dots \dots v(i+k, j+k) \end{cases}$$

Les fréquences N1, N2 ... Nn d'apparition des motifs M1, M2 ... Mn de l'image, en les triant selon l'ordre décroissant tout en leur attribuant un rang répondent à la formule suivante:

$$N_{\sigma(i)} = k.i^{-a} \qquad \text{(II.3)}$$

Où $N_{\sigma(i)}$ représente le nombre d'apparitions d'un motif de rang i, K et a sont des constantes.

- **La loi de Zipf inverse**

D'après la loi de Zipf inverse, le nombre I de motifs différents ayant une fréquence d'apparition f s'exprime par la formule suivante :

$$I(f) = af^{-b} \qquad \text{(II.4)}$$

Dans cette formule, a et b sont des constantes positives (Caron 2004).

II.4.2 Codage de l'image

Pour l'application des modelés de lois puissance en analyse d'image, nous nous trouvons face à la nécessité de définir un codage aux images pour réduire le nombre des motifs distincts et faire en sorte qu'un motif apparaisse plusieurs fois dans l'image.

- **Utilisation du codage des 9 classes**

Pour l'utilisation de ce codage, le choix du nombre de classe est nécessaire, en effet, il a été démontré (Caron. 2004) que n= 9 fut le choix optimal pour le nombre des classes : à cet effet, nous partitionnons la dynamique [0, 255] des niveaux de gris en un nombre de 9 classes de largeur égale, et nous numérotons ces classes en ordre croissant de 0 à 8.

Pour réaliser le codage de l'image nous procédons comme suit : chaque pixel du motif, se verra attribué une valeur de classe c(x, y) en fonction de la valeur g(x, y) de son niveau de gris par la formule suivante :

$$C(x, y) = \text{int}\left[\frac{n*g(x,y)}{255}\right] \quad \text{(II.5)}$$

- **Utilisation du codage des rangs généraux**

Le principe du codage des rangs généraux consiste à remplacer les niveaux de gris des pixels par leur rang dans un voisinage.

Le codage d'un motif s'effectue selon les étapes suivantes: nous ordonnons les niveaux de gris des pixels du motif selon un ordre croissant et nous affectons la valeur 0 au niveau de gris le plus bas, tout en incrémente la valeur d'une unité jusqu'à atteindre le niveau de gris le plus élevé. Les pixels ayant la même valeur de niveau de gris recevront le même rang.

La figure II.4 présente un exemple d'un motif avant et après avoir effectué son codage par la méthode des rangs généraux. Nous pouvons remarquer que ce codage permet de mettre en évidence une information plus précise concernant la texture de l'image [2-13].

II.4.3 Traçage des courbes de Zipf et de Zipf inverse

- **L'algorithme de traçage de la courbe de Zipf**

Nous procédons par un balayage séquentiel de l'image avec un masque de capture 3x3 en codant les motifs par un codage qui convient aux propriétés de l'image que nous cherchons à mettre en évidence.

Par la suite, nous devons calculer le nombre d'occurrences de chacun des motifs distincts dans l'image, en effet, nous rangeons dans un tableau tous les motifs rencontrés tout en associant la fréquence d'apparition relative à chaque motif comme suit : pour chaque motif courant trouvé, nous le comparons aux autres du tableau. S'il existe déjà alors nous incrémentons sa fréquence d'apparition d'une unité, contrairement, nous l'ajoutons au tableau en initialisant sa fréquence d'apparition à 1.

L'étape à venir consiste à trier la fréquence d'apparition des motifs selon un ordre décroissant. Dernièrement, la fréquence de chaque motif est tracée en fonction de son rang dans un repère bi-logarithmique, où en abscisse le rang R des motifs et en ordonnée leur fréquence d'apparition. Nous exposons sur la figure II.3 la courbe de Zipf d'une image.

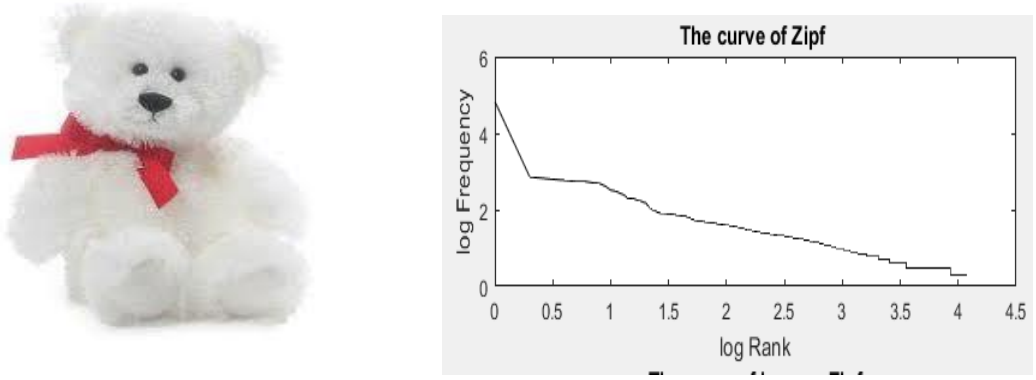


Figure II.3 Courbe de Zipf d'une image

- **L'algorithme de traçage de la courbe de Zipf inverse**

La première phase de l'algorithme est la même utilisée pour la construction de la courbe de Zipf. En effet, nous balayons l'image en comparons le motif courant aux motifs rencontrés dans l'image tout en rajoutant le motif dans le tableau s'il n'y est pas.

L'étape prochaine diffère de celle de Zipf, effectivement, nous comptons les motifs ayant la même fréquence d'apparition que la fréquence courante. A cet effet, nous initialisons la fréquence cherchée à 1 et nous parcourons séquentiellement le tableau des motifs pour compter les motifs de la même fréquence que la fréquence courante.

L'algorithme est réitéré en incrémentant à chaque fois la fréquence d'une unité jusqu'à l'atteinte de la fréquence maximale. Finalement, le nombre de motif est tracé en fonction de leur fréquence d'apparition dans un repère bi-logarithmique constituant la courbe de Zipf inverse.

Nous exposons sur la figure II.4 la courbe de Zipf inverse d'une image.

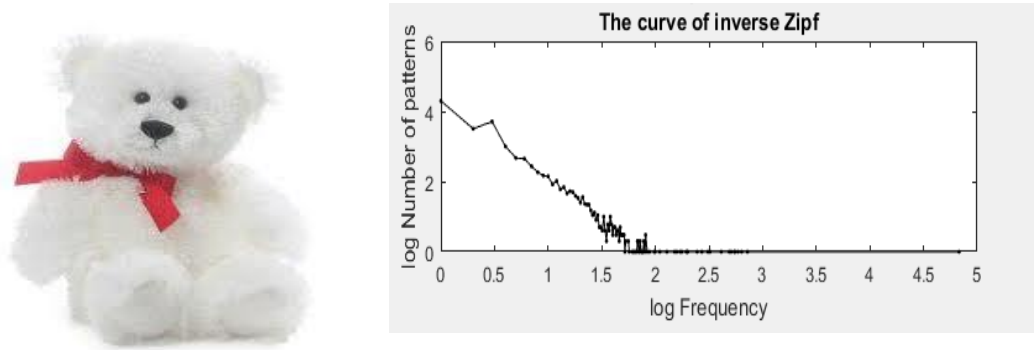


Figure II.4 Courbe de Zipf inverse d'une image

II.5 Conclusion

Nous avons démontré comment les lois de Zipf s'appliquent en imagerie pour la modélisation du contenu structurel de l'image. Pour cela, nous définissons des motifs qui sont les équivalents des mots dans le cas de l'analyse de texte.

La pertinence des lois de Zipf et de Zipf inverses pour analyser la structure complexe de l'image numérique nous pousse à réaliser une extension de l'application de ces lois au domaine de l'imagerie médicale, où ces dernières présenteront un outil efficace d'aide au diagnostic médical du cancer du sein assisté par ordinateur. En effet, les prochains chapitres de ce mémoire évoqueront davantage de détails sur ce point.

Chapitre III : Aide au diagnostic médical du cancer du sein assisté par ordinateur

III.1. Introduction

Jusqu'à ce jour l'atteinte par le cancer du sein ne cesse de s'augmente aux femmes, en réalité, il n'y a aucun remède efficace à cette dangereuse maladie, donc la détection précoce est la clé unique de diagnostic de cette maladie pour la diminution de la moyenne de mortalité associée.
La mammographie est utilisée dans ce sens, en effet, elle détecte jusqu'à 75% (Mitnick. 2005) des cancers du sein, par conséquent, l'évaluation de nouvelles techniques, assistées par ordinateur, de dépistage du cancer du sein était en ordre pour offrir un deuxième avis aux radiologues pour la prise de décision.
Les approches assistées par ordinateur ont tendance à optimiser le contraste au sein des mammographies pour faciliter la détection des tumeurs et leur classification en terme de bénignes ou malignes.
Dans ce chapitre nous proposons une approche de caractérisation de la texture des images mammaires basée fusion des lois de puissance : Zipf, Zipf inverse et l'ondelette de Haar pour la classification des zones d'intérêt extraites à partir des mammographies.

III.2. Le cancer du sein

Le taux d'atteinte par le cancer du sein augmente de plus en plus à travers le monde, et jusqu'à ce jour il n'existe pas une procédure efficace pour attaquer cette maladie en raison de sa cause inconnue. Il présente la deuxième cause de mortalité après le cancer des poumons (Hamoud M.2015) .Le cancer du sein est aussi l'une des principales causes de décès par cancer dans les pays les moins développés du fait de manque des moyens d'exploration de cette maladie (Article. 2013). Selon une analyse de la Société américaine du cancer (ACS) cette maladie peut causer le décès de 5,5 millions de femmes chaque année dans le monde jusqu'à 2030. En Algérie le cancer du sein désigne la première cause de mortalité chez la femme et le taux d'incidence est de 14.5 cas pour 100.000 habitants par ans. Selon ce taux ; il y'a 2.000 nouveaux cas chaque année les deux tiers survenant après l'âge de 45ans. L'incidence a augmenté de façon importante et constaté pendant 25 ans qu'elle elle est passé de 65.8 nouveaux cas / 100000 personne en 1980, à 101.5 nouveaux cas / 100000 personnes en 2005. (Belet et all, 2008). Le cancer du sein est une tumeur maligne qui se développe au niveau du sein, en effet, il correspond à la multiplication anarchique de cellules anormales incriminant des mécanismes très nombreux. Ces cellules envahissent progressivement les tissus voisins, atteignent les ganglions et se propagent par la circulation sanguine et lymphatique. Lorsque ces cellules anormales migrent et se fixent dans d'autres organes, elles donnent naissance à des métastases (Terki H et al.2014).

Nous ne connaissons pas les causes du développement du cancer du sein qui peut être dû à différents Facteurs :

- **L'âge**

Le risque d'avoir un cancer du sein augmente avec l'âge. Il reste rare avant 40 ans mais demeure beaucoup mais demeure beaucoup plus fréquent entre 60 et 65 ans. Le facteur âge représente le risque le plus important car on constate qu'après 40 ans le risque de développer un cancer du sein se multiplie par une fois et demie tous les dix ans. (Florian SCOTTE, ET all, 2002, P179) (Maouchich S et al .2017).

- **Les caractéristiques individuelles**

Le risque est plus important chez les femmes qui ont eu des règles précoces et une ménopause tardive. Il est élevé également chez celles qui n'ont pas eu de grossesse ou une première

grossesse après 40 ans, ou encore qui ont pris un traitement substitutif à la ménopause pendant plus ménopause pendant plus de 10 Ce risque augmente aussi avec la consommation excessive de sucres, de graisses animales et d'alcool. (Florian SCOTTE, ET all, p179) (Maouchich S et al .2017).

- **Les antécédents familiaux**

Quelques cancers du sein sont liés au fait que plusieurs femmes de la même famille ont été ou sont atteintes par cette même maladie surtout si elles avaient à l'époque moins de 40 ans. Un cancer du sein est également plus susceptible de se développer chez une femme dont plusieurs membres d'une même famille ont souffert d'un cancer du côlon ou des ovaires. Le facteur génétique n'est lié qu'à 5% des cancers du sein (Henrik THORLACIUS, 1995, P346) (Maouchich S et al .2017).

III.3. Le dépistage du cancer du sein

Le dépistage (Zivian. 2008) est le test systématique de personnes asymptomatiques de maladie préclinique. L'ultime but est la détection du symptôme du cancer du sein avant l'existence d'une preuve sur maladie citons l'exemple d'une masse palpable ainsi que l'interdiction ou le retard du développement de la maladie à travers la détection précoce et le traitement. En fait le dépistage systématique aux femmes à risque normal est effectué annuellement à l'âge de 40 ans. L'efficacité de la mammographie est reconnue depuis le milieu des années 1960 (Griff et al. 2002) (Hamoud M.2015). Donc le dépistage consiste en une démarche pour la détection, au plus tôt, en l'absence de symptômes, des lésions pouvant être cancéreuses ou pouvant converger vers un cancer. Il est aujourd'hui possible de dépister ou de détecter précocement certains cancers. Le dépistage peut être réalisé soit dans le cadre d'un programme organisé par les autorités de santé publique, ou bien de façon individuelle à l'initiative du professionnel de la santé ou de la patiente (Touami R et al.2011).

III.3.1. La mammographie

La mammographie est un examen radiologique consacré à l'étude du sein. En effet, cet examen s'effectue avec un appareil à rayons X dédié uniquement à cet usage : le mammographe. Nous distinguons deux types d'examen : la mammographie diagnostique et la mammographie de dépistage. La mammographie diagnostique est effectuée chez des personnes qui ont des

symptômes : masse palpable, écoulement (Article.2014). La mammographie de dépistage recherche la présence d'une lésion chez des personnes qui n'ont aucun symptôme. La détection radiologique des lésions du sein exige des images de grande précision, qui peuvent être obtenues seulement avec le mammographe. Notons que les mammographies numériques peuvent être interprétées à partir d'un poste de travail informatique plutôt qu'un film et qu'elles sont fructueuses pour l'imagerie des femmes ayant des seins denses pour l'évaluation des microcalcifications subtiles (Article.2014). La figure III.1 présente un appareil de génération des mammographies.

Figure III.1 Appareil de génération des mammographies. Source [Grower-Thomas. 2009].

Si nous évoquons maintenant l'utilité d'une mammographie, nous dirions (Touami R et al.2011) :

- Qu'elle permet d'étudier la glande mammaire et permet de dépister à un stade précoce le cancer du sein.
- Qu'elle détecte des anomalies telles que des opacités et des micros calcifications.
- Qu'elle repère des lésions durant l'intervention chirurgicale : un fil métallique est placé sous contrôle de la mammographie et guide le chirurgien pour enlever la zone anormale.

III.4. Détection et classification assistées par ordinateur (CADe/CADx) des tumeurs dans la mammographie

Dernièrement, les recherches s'évoluent d'une façon rapide dans le domaine de diagnostic des tumeurs à l'aide de l'ordinateur. En effet, dû à sa propagation rapide, Il apparait indispensable de développer de nouvelles méthodes ou de nouveaux protocoles pour le dépistage du cancer du sein. De plus, l'imagerie médicale est certainement l'un des domaines de la médecine qui a connu une remarquable révolution pendant ces vingt dernières années (Hamoud M.2015), donc ; la mise en œuvre de système de détection et de diagnostic assistés par ordinateur du cancer du sein semblait indispensable. Ces récentes découvertes permettent non seulement un meilleur diagnostic mais évitent aussi les biopsies effectuées aux tumeurs bénignes, en effet, ces systèmes d'aide à la prise de décision dédiés à l'imagerie médicale assistent les radiologues dans la tâche délicate d'interprétation des images médicales (Cheikhrouhou I.2012).

Nous distinguons deux types de systèmes pour le diagnostic assisté par ordinateur : les systèmes de détection des tumeurs assistée par ordinateur (CADe) et les systèmes de diagnostic ou de classification des tumeurs assistés par ordinateur (CADx). La figure III.2 présente un résumé schématisé éclaircissant la différence entre ces deux systèmes (Cheikhrouhou I.2012).

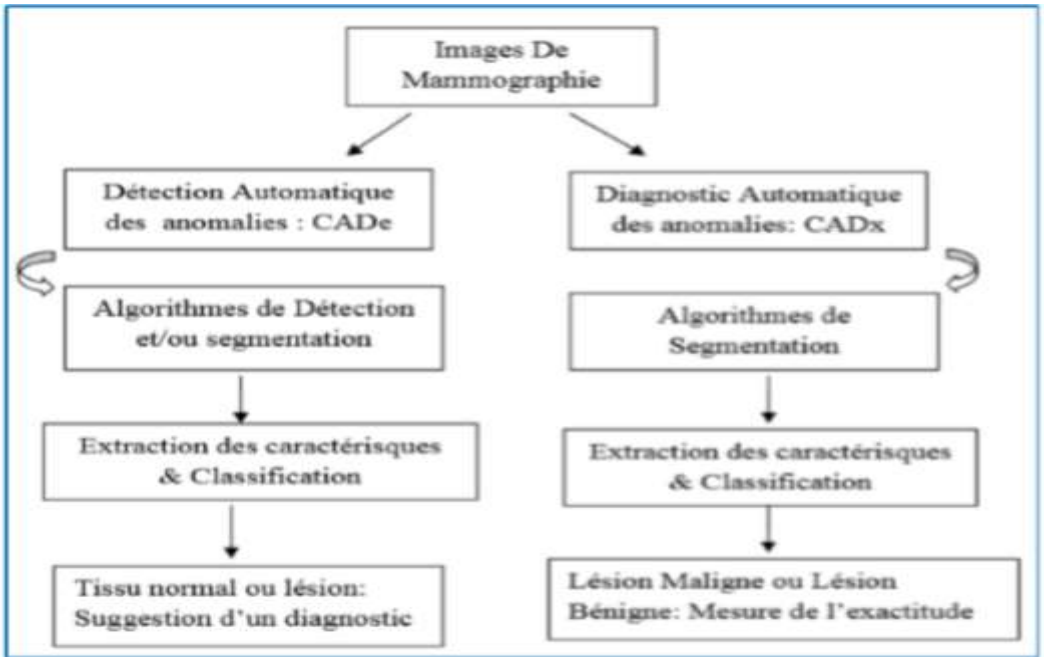


Figure III.2 Différence entre les systèmes CADe et les systèmes CADx.

III.4.1. Détection des tumeurs assistée par ordinateur (CADe)

Les systèmes de détection automatisée des anomalies dans des images de mammographie utilisent des techniques de détection basées régions en segmentant des régions d'intérêt pour en extraire les caractéristiques qui seront par la suite utilisées pour classifier les régions comme suspect ou non suspect (Tang, Rangayyan, Xu, El Naqa, & Yang, 2009). De plus, nous distinguons d'autres techniques basées analyse de pixels (Minh-Nguyen et al. 2013) qui admettent que les pixels à l'intérieur d'une tumeur présentent des descripteurs différents par rapport aux autres pixels du tissu parenchyme mammaire. Ces descripteurs peuvent consister en des valeurs de niveaux de gris, des mesures de texture ou des mesures morphologiques.
Nous pouvons attester que le but des systèmes de détection des tumeurs assistée par ordinateur (CADe) est de localiser des masses (Hamoud M.2015)**.**

III.4.2. Diagnostic des tumeurs assisté par ordinateur (CADx)

La classification des tumeurs mammaires par un radiologue est une classification humaine subjective pouvant classifier une même lésion en deux classes différentes. En revanche, un système automatique de diagnostic assisté par ordinateur (DAOx), basé sur les descripteurs de la lésion donne toujours les mêmes résultats (Cheikhrouhou I.2012). Donc l'extraction de descripteurs mathématiques discriminants est nécessaire du fait que le CADx repose sur la caractérisation des lésions détectées pour réaliser une classification ayant comme objectif de distinguer les tumeurs malignes / bénignes et le tissu parenchyme mammaire (Hamoud M.2015)**.**
Nous pouvons attester que la tâche du diagnostic est modélisée comme une tâche de classification, en effet, le CADx extrait des descripteurs à partir des zones suspectes segmentées et les donne comme entrée à un classifieur pour évaluer les tumeurs. Les méthodologies de classification des tumeurs se basent sur les sorties fournies par les classifieurs (Hamoud M.2015)**.**

III.5. Indexation et recherche des mammographies par le contenu (CBMIIR) pour l'aide au diagnostic médical du cancer du sein assisté par ordinateur

Suite à la révolution dans le dépistage du cancer du sein, une grande augmentation de la quantité des mammographies produites et stockées a été observée. Naturellement, une recherche manuelle dans ces ressources précieuses sera impossible et par conséquence il était devenu

obligatoire de développer un système performant d'indexation et de recherche des mammographies par le contenu (CBMIIR : Content Based Mammogram Image Indexing and Retrieval). Ce système offre une solution automatisée assistée par ordinateur avec un but ultime, celui de l'accès efficace aux mammographies archivées désirées, pathologiquement similaires, à la mammographie en cours de traitement. Contrairement aux travaux initiaux sur l'indexation et la recherche d'images qui se basaient sur l'annotation textuelle manuelle posant plusieurs limitations telles que la subjectivité de l'annotation textuelle ou la consommation d'un temps considérable, les systèmes de recherche des images par le contenu (CBIR : Content Based Image Retrieval) présentent la capacité à effectuer la recherche en fonction du contenu visuel des images comme mentionné sur la figure III.3. Précisons, que dans le domaine médical, nous effectuons le diagnostic à travers 'une comparaison entre les cas déjà diagnostiqués précédemment et les cas actuels en cour de diagnostic en s'appuyant sur des conditions pathologiques et donc à réaliser un diagnostic à base de cas (Weidong Cai et al. 2008).

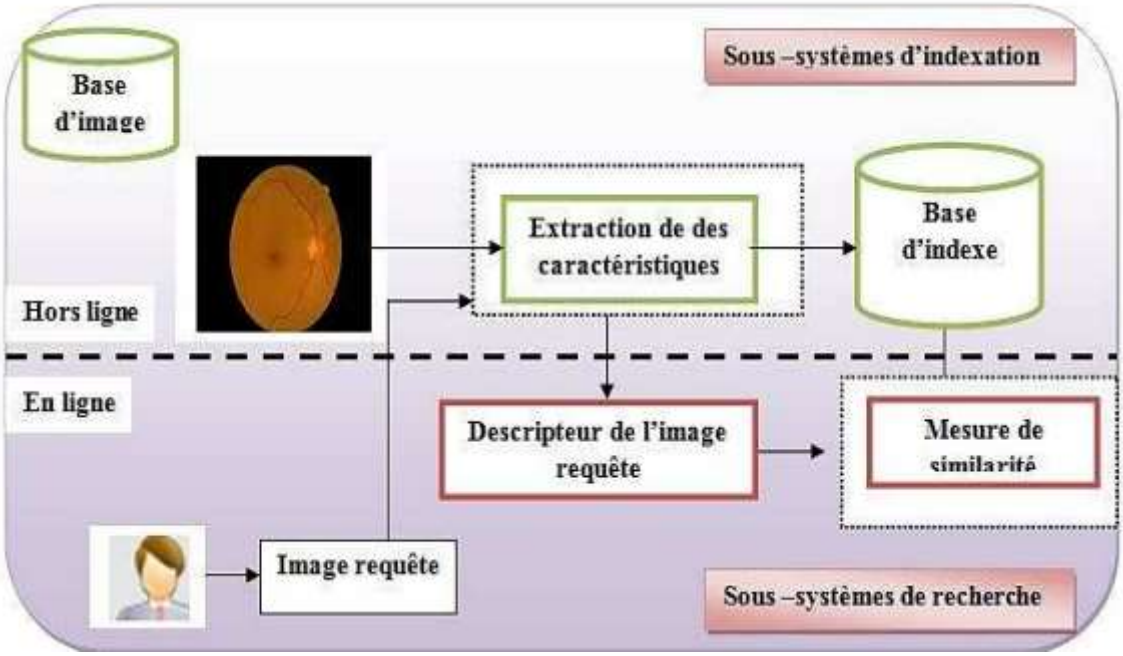


Figure III.3 Architecture d'un système d'indexation et de recherche d'image par le contenu

III.6. L'apport d'un système d'indexation et de recherche de mammographies par le contenu (CBMIIR) par rapport à un système d'aide au diagnostic assisté par ordinateur (CADx) se basant sur la sortie d'un classifieur

Des centaines d'images de mammographie sont produites dans les hôpitaux par jour, évoluant et devenant des référentiels accessibles. A cet issu, l'importance de recherche et de récupération des images mammaires de la même maladie augmente de plus en plus pour aider à la décision. Dans les systèmes d'aide au diagnostic médical (CADx), un classifieur est optimisé sur un ensemble d'apprentissage, après cela, la classification d'un nouveau cas d'entrée s'effectue. Evidemment, les radiologues obtiennent l'aide suite à des symboles numériques abstraits indiquant les classes à qui appartiennent les régions d'intérêt requêtes (Hamoud M.2015)**.**

.

Les radiologues jugent que le résultat ou la sortie d'un classifieur n'est pas évidente, et préfèrent construire leur diagnostic final à base de cas déjà diagnostiqués et similaires à celui en cours d'analyse, en les récupérant et affichant à partir de bases de données indexées de cas déjà analysés (Hamoud M.2015)**.**

.

L'utilité d'utiliser des régions d'intérêt similaires pour le diagnostic d'un nouveau cas est manifestée par l'hypothèse que les radiologues répondent favorablement aux régions d'intérêt similaires au cas en cours d'analyse, que des chiffres abstraits. En effet, suite à leur formation basée sur la lecture de diverses mammographies, ils apprennent à reconnaître l'anatomie normale, l'anomalie bénigne, et quel est le cancer (Hamoud M.2015)**.**

.

III.7. Approche proposée basée fusion des lois de puissance : Zipf, Zipf inverse et les ondelettes de Haar pour la classification des zones d'intérêt des mammographies

Le but de la classification des images mammaires consiste en la répartition d'un ensemble de régions d'intérêt, extraites à partir de ces dernières, en groupes similaires de régions d'intérêt avec un tissu parenchymateux normal, régions d'intérêt portant des tumeurs bénignes et régions d'intérêt portant des tumeurs malignes. Dans ce sens, des descripteurs discriminants et efficaces doivent être utilisés pour la discrimination. Notons qu'en général, les descripteurs de texture des motifs portant des tumeurs bénignes, malignes ainsi que ceux du tissu

parenchymateux normal sont différents. En effet, les tumeurs causent des manifestations dans la texture du tissu du sein comme suit: les tumeurs bénignes possèdent une texture plutôt homogène et les tumeurs malignes se dotent d'une texture hétérogène et complexe en cause de leur caractère invasif. Tandis que le tissu normal possède une texture totalement homogène.

III.7.1. Analyse et caractérisation de la texture des zones d'intérêt par les lois de Zipf et de Zipf inverse

Dans cette sous-section, nous allons analyser les courbes de Zipf et de Zipf inverse pour une zone d'intérêt du tissu sein, une zone d'intérêt portant une tumeur bénigne et une autre portant une tumeur maligne comme mentionné sur la figure III.4.

(a) **(b)** **(c)**

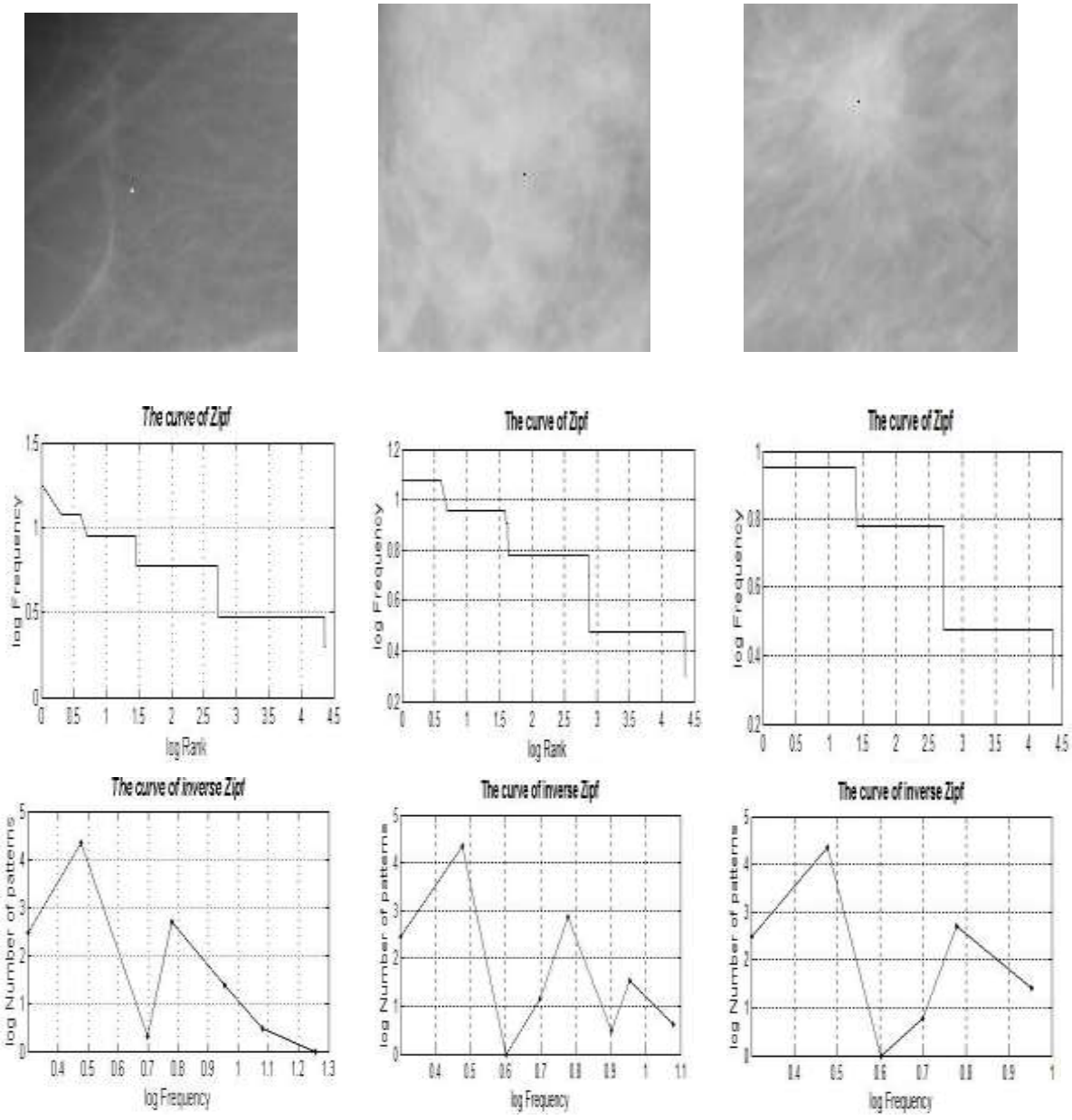


Figure III.4. Courbes de Zipf et de Zipf inverse de trois zones d'intérêts encodées par le codage des rangs généraux : (a) zone d'intérêt sans tumeur, (b) zone d'intérêt portant une tumeur bénigne et (c) zone d'intérêt portant une tumeur maligne.

Comme mentionné sur cette figure, les courbes de Zipf et de Zipf inverse changent effectivement selon le contenu structurel de la région d'intérêt. En effet, le tissu parenchymateux sans lésions, le tissu de la région d'intérêt portant une tumeur bénigne ou celui de la région d'intérêt portant une tumeur maligne ne sont pas identique.

En comparant les courbes de Zipf des régions d'intérêt qui portent une tumeur bénigne et celles portant une tumeur maligne, nous observons que la région d'intérêt de la tumeur maligne se constitue de diverses détails, donc, une texture complexe due à la nature hétérogène de la texture des tumeurs malignes. D'autre part, la texture des tumeurs bénignes est plus homogène que celles des tumeurs malignes, de ce fait, nous observons que le nombre de motifs homogènes est plus grand générant une ordonnée à l'origine de la courbe de Zipf plus élevée.

Si en a comparer les courbes de Zipf inverse des régions d'intérêt portant une tumeur bénigne et celles portant une tumeur maligne, nous attesterons également un changement des apparences des courbes selon le contenu structurel des régions d'intérêt.

Ainsi, pour la distinction entre les tumeurs malignes, bénignes et les tissus parenchymateux normaux, nous explorons les caractéristiques intéressantes de ces courbes à travers l'extraction des descripteurs suivants :

- **L'aire délimitée par la courbe de Zipf :**

Nous calculons l'aire délimitée par la courbe de Zipf à partir des courbes de Zipf obtenues suite au codage de l'image par le codage des rangs généraux. Soit *n* le nombre de motifs de la courbe (Hamoud M.2015)**.**

, f_i la fréquence et r_i le rang du motif i, l'aire de la courbe est donnée par la formule III.1 :

$$A = \sum_{i=1}^{n-1} \frac{(f_i + f_{i+1})(r_{i+1} - r_i)}{2} \qquad \text{(III.1)}$$

- **Les pentes des courbes de Zipf et de Zipf inverse :**

La pente moyenne d'une courbe est le coefficient directeur de la droite des moindres carrés. Elle est donnée par la formule III.2 :

$$p = \frac{n\sum_{i=1}^{n} y_i x_i - \sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2} \qquad \text{(III.2)}$$

- **L'entropie de la courbe de Zipf**

Il est possible de définir une mesure d'entropie à partir des fréquences d'apparition des motifs de l'image. Deux formulations de l'entropie sont possibles, l'entropie relative aux motifs et l'entropie relative aux fréquences d'apparition.

a) L'entropie relative aux motifs

L'entropie relative aux motifs est définie par la formule III.3 pour un ensemble de R motifs distincts :

$$H_w = -\sum_{r=1}^{R} \frac{f(r)}{T} \log_R \frac{f(r)}{T} \qquad \text{(III.3)}$$

Dans cette formule, *f(r)* représente la fréquence du motif de rang *r*, T représente le nombre total de motifs différents ou non, et on utilise un logarithme de base R.

b) L'entropie relative à la fréquence

L'entropie relative à la fréquence est définie par la formule:

$$H_f = -\sum_{f=1}^{F} \frac{I(f)}{R} \log_F \frac{I(f)}{R} \qquad \text{(III.4)}$$

I(f) représente le nombre de motifs distincts de fréquence f et F le nombre total d'occurrences des motifs dans l'image. On utilise un logarithme à base F afin que l'entropie soit comprise entre 0 et1.

- **La constante alpha**

Pour tout motif appartenant à l'image, la fréquence d'apparition de ce motif * son rang dans la liste ordonnée décroissante des fréquences des motifs de l'image représente une constante.

- **Les ordonnées à l'origine des courbes de Zipf et de Zipf Inverse**

Nous extrayons l'ordonnée à l'origine à partir des courbes de Zipf et de Zipf inverse.

III.7.2. Analyse et caractérisation de la texture des zones d'intérêt par les ondelettes de Haar

- **Principe des ondelettes de Haar dans le cas unidimensionnel (1D)**

Soit une séquence d'origine S0 = [2 4 8 12 14 0 2 1], correspondant aux données initiales avec la transformation par ondelettes de Haar .Si nous appliquons la transformation en ondelettes de Haar sur la séquence S0, nous obtenons alors deux nouvelles séquences, de taille moitié, dont l'une (notée S1) contient les coefficients de moyenne, et la seconde (notée D1) contient les coefficients de détail (Article –2011).

Le calcul des coefficients se fait en réunissant les coefficients de S0 2 par 2, ce qui permet d'avoir :

$$S_1 = \left[\frac{2+4}{2} \quad \frac{8+12}{2} \quad \frac{14+0}{2} \quad \frac{2+1}{2}\right], \text{ soit } S_1 = [3 \;\; 10 \;\; 7 \;\; 1.5]$$

$$D_1 = \left[\frac{2-4}{2} \quad \frac{8-12}{2} \quad \frac{14-0}{2} \quad \frac{2-1}{2}\right], \text{ soit } D_1 = [-1 \;\; -2 \;\; 7 \;\; 0.5]$$

Nous appliquons à nouveau les mêmes opérations sur la séquences S2, afin d'obtenir deux nouvelles séquences, de taille moitié de S1, dont l'une (notée S2) contenant les coefficients de moyenne, et la seconde (notée D2) contenant les coefficients de détail.
A cet effet, nous obtenons la table de décomposition en ondelettes de Haar de S0 :

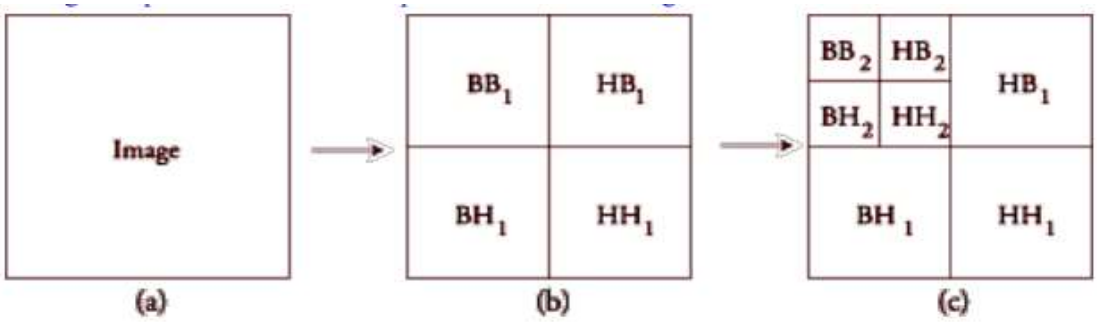
Niveau de résolution n	Coefficients de moyenne S_n	Coefficients de détail D_n
0 – Image d'origine	$S_0 = [2 \;\; 4 \;\; 8 \;\; 12 \;\; 14 \;\; 0 \;\; 2 \;\; 1]$	
1	$S_1 = [3 \;\; 10 \;\; 7 \;\; 1.5]$	$D_1 = [-1 \;\; -2 \;\; 7 \;\; 0.5]$
2	$S_2 = [6.5 \;\; 4.25]$	$D_2 = [-3.5 \;\; 2.75]$
3	$S_3 = [5.375]$	$D_3 = [1.125]$

La séquence finale obtenue après décomposition en ondelettes 1D de Haar est de même taille que la séquence initiale, et les coefficients correspondent au dernier coefficient de moyenne (soit S3) suivi des coefficients de détail des suites Dn du niveau le moins détaillé (n = 3) et niveau le plus détaillé (n = 1). Nous aurrons alors:

$$F = [5.375 \;\; 1.125 \;\; -3.5 \;\; 2.75 \;\; -1 \;\; -2 \;\; 7 \;\; 0.5]$$

- **Passage au cas bidimensionnel (2D) pour les images BITMAP**

Dans le cas d'une image BITMAP, la méthode est la même sauf que nous appliquons, à chaque niveau de détail, la transformation en ondelettes 1D de Haar dans chacune des directions « largeur » puis « hauteur ». Donc, nous obtiendrons le diagramme suivant:



Où B recueille les coefficients de moyenne, et H les coefficients de détail. Un bloc noté HB a été construit en appliquant la transformation en ondelettes 1D de Haar suivant la largeur, pour obtenir les coefficients de moyenne, sur lesquels nous avons appliqué la transformation en ondelettes 1D de Haar suivant la hauteur, pour obtenir les coefficients de détail.

Supposons l'image I suivante :

$$I = \begin{bmatrix} 0 & 250 & 25 & 50 & 200 & 0 & 0 & 0 \\ 50 & 50 & 50 & 25 & 50 & 0 & 25 & 0 \\ 25 & 50 & 0 & 250 & 0 & 50 & 0 & 0 \\ 75 & 200 & 200 & 0 & 250 & 250 & 0 & 200 \\ 250 & 25 & 250 & 200 & 0 & 75 & 25 & 25 \\ 50 & 0 & 50 & 0 & 0 & 75 & 250 & 250 \\ 250 & 250 & 25 & 250 & 50 & 25 & 50 & 50 \\ 250 & 200 & 50 & 50 & 50 & 25 & 0 & 200 \end{bmatrix}$$

1. Transformation en ondelettes 1D de Haar sur le niveau 0 :

a. Suivant la largeur :

125	37.5	100	0	−125	−12.5	100	0
50	37.5	25	12.5	0	12.5	25	12.5
37.5	125	25	0	−12.5	−125	−25	0
137.5	100	250	100	−62.5	100	0	−100
137.5	225	37.5	25	112.5	25	−37.5	0
25	25	37.5	25	25	25	−37.5	0
250	137.5	37.5	50	0	−112.5	12.5	0
225	50	37.5	100	25	0	12.5	−100

b. Suivant la hauteur:

−87.5	37.5	62.5	6.25	−62.5	0	62.5	6.25
−87.5	112.5	137.5	50	−37.5	−12.5	−12.5	−50
−81.25	125	37.5	25	68.75	25	−37.5	0
−237.5	93.75	37.5	75	12.5	−56.25	12.5	−50
37.5	0	62.5	6.25	−62.5	−12.5	37.5	−6.25
−50	12.5	−112.5	−50	25	−112.5	−12.5	50
56.25	100	0	0	43.75	0	0	0
12.5	43.75	0	−25	−12.5	−56.25	0	50

2. Transformation en ondelettes 1D de Haar sur le niveau 1 (c'est-à-dire sur le cadran haut-gauche de la matrice précédente) :

a. Suivant la largeur :

62.5	34.375	25	28.125	−62.5	0	62.5	6.25
100	93.75	−12.5	43.75	−37.5	−12.5	−12.5	−50
103.125	31.25	−21.875	6.25	68.75	25	−37.5	0
165.625	56.25	71.875	−18.75	12.5	−56.25	12.5	−50
37.5	0	62.5	6.25	−62.5	−12.5	37.5	−6.25
−50	12.5	−112.5	−50	25	−112.5	−12.5	50
56.25	100	0	0	43.75	0	0	0
12.5	43.75	0	−25	−12.5	−56.25	0	50

b. Suivant la hauteur :

81.25	64.0625	6.25	35.9375	−62.5	0	62.5	6.25
134.375	43.75	25	−6.25	−37.5	−12.5	−12.5	−50
−18.75	29.6875	18.75	−7.8125	68.75	25	−37.5	0
−31.25	−12.5	−46.875	12.5	12.5	−56.25	12.5	−50
37.5	0	62.5	6.25	−62.5	−12.5	37.5	−6.25
−50	12.5	−112.5	−50	25	−112.5	−12.5	50
56.25	100	0	0	43.75	0	0	0
12.5	43.75	0	−25	−12.5	−56.25	0	50

3. Transformation en ondelettes 1D de Haar sur le niveau 2 (c'est-à-dire sur le cadran haut-gauche de la matrice précédente) :

a. Suivant la largeur :

72.65625	8.59375	6.25	35.9375	−62.5	0	62.5	6.25
89.0625	45.3125	25	−6.25	−37.5	−12.5	−12.5	−50
−18.75	29.6875	18.75	−7.8125	68.75	25	−37.5	0
−31.25	−12.5	−46.875	12.5	12.5	−56.25	12.5	−50
37.5	0	62.5	6.25	−62.5	−12.5	37.5	−6.25
−50	12.5	−112.5	−50	25	−112.5	−12.5	50
56.25	100	0	0	43.75	0	0	0
12.5	43.75	0	−25	−12.5	−56.25	0	50

b. Suivant la hauteur :

80.959375	26.953125	6.25	35.9375	−62.5	0	62.5	6.25
−8.103125	−18.359375	25	−6.25	−37.5	−12.5	−12.5	−50
−18.75	29.6875	18.75	−7.8125	68.75	25	−37.5	0
−31.25	−12.5	−46.875	12.5	12.5	−56.25	12.5	−50
37.5	0	62.5	6.25	−62.5	−12.5	37.5	−6.25
−50	12.5	−112.5	−50	25	−112.5	−12.5	50
56.25	100	0	0	43.75	0	0	0
12.5	43.75	0	−25	−12.5	−56.25	0	50

Le résultat R après la transformée en ondelettes de Haar est comme suit :

$$R = \left[\begin{array}{cccc|cccc} 80.959375 & 26.953125 & 6.25 & 35.9375 & -62.5 & 0 & 62.5 & 6.25 \\ -8.103125 & -18.359375 & 25 & -6.25 & -37.5 & -12.5 & -12.5 & -50 \\ \hline -18.75 & 29.6875 & 18.75 & -7.8125 & 68.75 & 25 & -37.5 & 0 \\ -31.25 & -12.5 & -46.875 & 12.5 & 12.5 & -56.25 & 12.5 & -50 \\ \hline 37.5 & 0 & 62.5 & 6.25 & -62.5 & -12.5 & 37.5 & -6.25 \\ -50 & 12.5 & -112.5 & -50 & 25 & -112.5 & -12.5 & 50 \\ 56.25 & 100 & 0 & 0 & 43.75 & 0 & 0 & 0 \\ 12.5 & 43.75 & 0 & -25 & -12.5 & -56.25 & 0 & 50 \end{array}\right]$$

Nous présentons sur la figure III.5, l'analyse d'une zone d'intérêt portant une tumeur bénigne par la transformée en ondelettes de Haar.

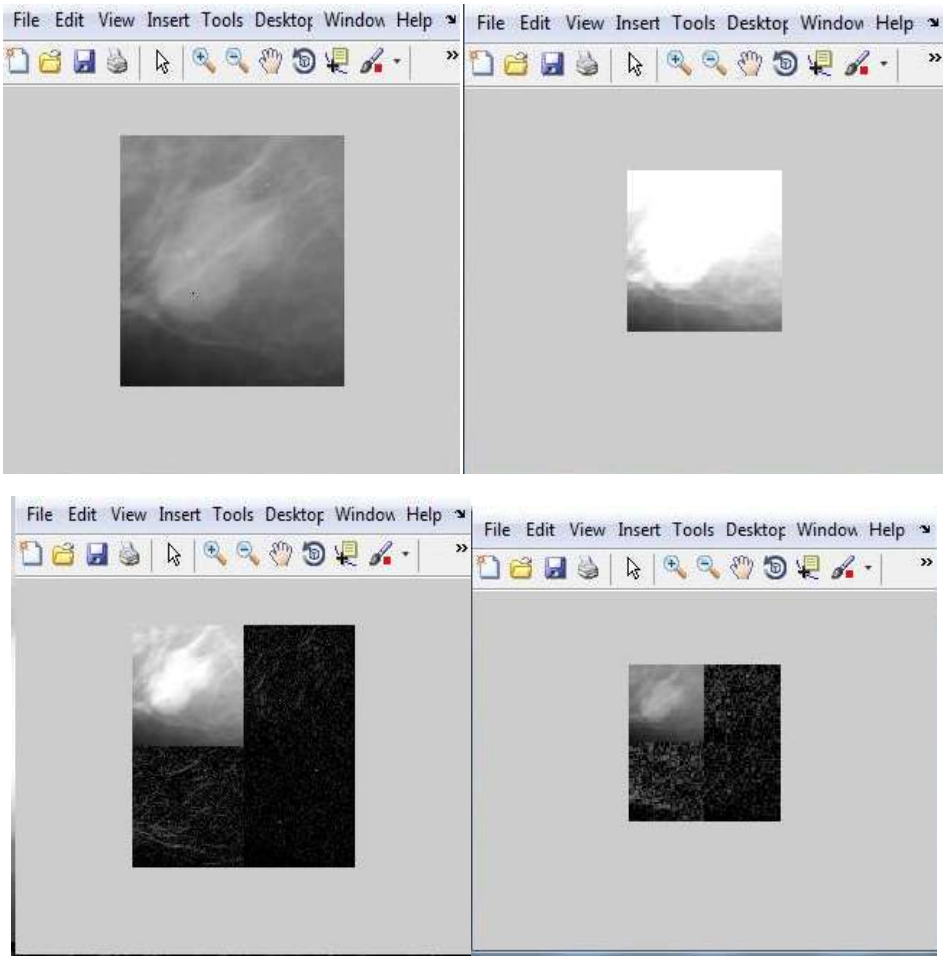


Figure III.5 L'analyse d'une zone d'intérêt portant une tumeur bénigne par la transformée en ondelettes de Haar**.**

III.8. Conclusion

De plus en plus, la détection et le diagnostic du cancer du sein assisté par ordinateur se propagent dans le monde. Dans ce cadre, réside l'origine de notre travail se basant sur la contribution d'une fusion des lois de puissance Zipf et Zipf inverse ainsi que les ondelettes de Haar pour aboutir à une caractérisation judicieuse du tissu mammaire.

Plus précisément, nous nous intéressons aux systèmes d'indexation et de recherche des mammographies par le contenu (CBMIIR) donnant des possibilités nouvelles et passionnantes pour l'analyse et le diagnostic des images mammaires. Dans ce sens, le prochain chapitre sera consacré à la présentation de notre système d'aide au diagnostic médical du cancer du sein assisté par ordinateur.

Chapitre IIII : Conception et Implémentation

IV.1. Introduction

L'indexation et la recherche d'images médicales par le contenu offrent aux radiologues la possibilité d'accéder, d'interroger ainsi que d'exploiter directement les bases d'images médicales en utilisant leur contenu.

Dans ce chapitre nous allons présenter en détails, les techniques utilisées pour le développement d'un système qui classifie les tumeurs mammaires en se basant sur la technique du template-matching ainsi que la caractérisation texturales des images mammaires par une fusion des lois de puissance : Zipf, Zipf inverse et les ondelettes de Haar à travers la technique d'indexation et de recherches des mammographies par le contenu. Bien évidemment, nous évoquerons par la suite son fonctionnement.

IV.2. Problématique

Notre problématique de recherche consiste en la suggestion d'une technique performante d'analyse de la texture d'images mammaires. En effet, les relations qui semblent les plus simples à utiliser sont les approches linéaires comme un polynôme ou les transformées intégrales linéaires. Néanmoins, ces dernières ne sont pas assez puissantes pour la modélisation d'une image avec sa structure complexe.

Dans (Hamoud M.2015) (Hamoud et al 2011), les auteurs ont proposé les lois de puissance Zipf et Zipf inverse durant la caractérisation de la texture des images mammaires pour l'aide au diagnostic médical du cancer du sein assisté par ordinateur en aboutissant à des premiers résultats très satisfaisants. En effet, les lois de Zipf et Zipf inverse modélisent parfaitement le contenu structurel des images mammaires et les descripteurs texturaux générés sont précieux pour le diagnostic, mais, difficiles à explorer par les radiologues.

Le travail évoqué dans ce mémoire consiste à s'inspire des perspectives de ces travaux en proposant une fusion des lois de Zipf et de Zipf inverse avec l'approche des ondelettes de Haar pour tirer profit de l'apport complémentaire de ces deux approches de caractérisation de la texture. En effet, l'ultime but est d'essayer d'améliorer les résultats aboutis par l'utilisation des lois de Zipf et de Zipf inverse uniquement.

Nous allons utiliser simultanément des descripteurs obtenus suite à l'analyse des zones d'intérêt par les lois puissance ainsi que ceux obtenus à partir des ondelettes de Haar dans le but de bénéficier de leur apport complémentaire par conséquent, nous appliquerons une analyse en composantes principales (ACP) pour réduire le nombre de descripteurs à mesure.

L'imagerie médicale a révolutionné la médecine, les systèmes d'indexation et de recherche des images médicales semblent une bonne solution pour la gestion du contenu important des images médicales archivées. Les radiologues ont exprimé leur doute du résultat du diagnostic offert par les systèmes d'aide au diagnostic médical du cancer du sein se basant sur un classifieur. Ainsi, nous renforçons les systèmes (CADx) par les systèmes d'indexation et de recherche des mammographies par le contenu (CBMIIR), motivées par l'avis favorable des radiologues pour l'utilisation du diagnostic issu de ces système du moment qu'il est obtenu à base de cas récupérés et affichés d'une bases de données indexées de cas déjà traités.

Donc nous allons proposer un système d'indexation et de recherche des images mammaires se basant sur la fusion des lois de Zipf et Zipf inverse avec les ondelettes de Haar pour la caractérisation et la quantification de la texture des zones d'intérêt extraites des images

mammaires. Tout en effectuant la classification des zones d'intérêt par la technique du template-matching.

IV.3. Principe de la classification basée template-matching et les K plus proches voisins

Cette technique est largement utilisée dans les domaines de détection d'objets tels que le suivi de l'imagerie médicale, son point crucial est la considération d'une «mesure» appropriée pour quantifier la similarité. La première phase consiste à définir une mesure pour calculer la "distance" ou la "similarité" entre les motifs d'apprentissage (connus) et le motif de test (inconnu) pour effectuer la classification durant la deuxième phase.

Notre méthodologie pour la classification des zones d'intérêts segmentées à partir des mammographies à des zones d'intérêts qui portent des tumeurs bénignes, des tumeurs malignes ou des zones d'intérêts sans la présence de lésions est basée sur l'approche template-matching. En effet, nous récupérons à partir de la base de données, les zones d'intérêts, pathologiquement similaires, à chaque zone d'intérêt requête. Ceci, par l'application de l'algorithme des K plus proches voisins (KPPV) vu qu'il est aisément paramétrable pour traiter un problème de classification avec un nombre quelconque d'étiquettes (Anne S et al.2015), de plus, il est basé sur la notion de proximité (voisinage) entre exemples et sur l'idée de raisonner à partir de cas similaires pour prendre une décision. Donc, il nous convient pour la modélisation de la classification basée template-matching. Autrement dit des entrées xi semblables devraient avoir des valeurs yi semblables (Moussati O.2016) et nous assignons à la zone d'intérêt requête sa classe pathologique, consistant en, la classe pathologique majoritairement renvoyée par cet algorithme durant la phase de recherche.

Ceci est indiqué sur la figure IV.1 présentant l'architecture de notre système d'indexation et de recherche des zones d'intérêt les plus similaires à une zone d'intérêt requête.

Le système proposé se compose de deux phases : la première phase **hors ligne**, durant cette étape, nous analysons les zones d'intérêts par les lois de Zipf et de Zipf inverse pour générer 8 descripteurs de la texture ainsi qu'une analyse par les ondelettes de Haar pour extraire 8 descripteurs texturaux également. Ensuite, l'analyse en composante principale est appliquée pour apaiser la dimension du vecteur descripteur à 10 descripteurs.

La deuxième phase **en ligne**, dans cette phase, l'utilisateur (probablement le radiologue à qui est destiné ce système d'aide au diagnostic médical du cancer du sein), introduit une zone d'intérêt requête segmentée à partir d'une mammographie, au système pour l'identification de

sa pathologie. Ensuite, le calcul du vecteur de descripteurs texturaux générés par les lois de Zipf et de Zipf inverse ainsi que les ondelettes de Haar est réalisé.

Les cas récupérés par le système suite à la recherche des K plus proches voisin relatifs à la zone d'intérêt requête seront utilisés comme prédicteurs utiles de la classe pathologique de la requête courante. Notre choix a porté sur l'utilisation de cet algorithme vu sa large utilisation dans les systèmes d'indexation et de recherche des images médicales (Eve Mathieu-Dupas.2010). En effet, la méthode des K plus proches voisins pondérés est une méthode de classification supervisée offrant des performances très intéressantes dans la recherche de nouveaux biomarqueurs pour le diagnostic (Koudri M.2011).

En partant du principe que les observations semblables sont proches l'une de l'autre et que les observations dissemblables sont éloignées l'une de l'autre, le diagnostic de la requête en cours d'analyse sera construit à partir de ses "voisines". La nouvelle observation est placée dans la catégorie ou la classe pathologique la majoritairement renvoyée par le processus de recherche.

IV.3.1 Algorithme des K-plus proches voisins

La méthode des plus proches voisins (noté parfois k-PPV ou k-NN pour (k-Nearest Neighbor) nécessite de choisir une distance, la plus classique est la distance euclidienne et le nombre de voisins à prendre en compte. Cette méthode supervisée est souvent efficace, cependant, le temps de prédiction est très long, car il nécessite le calcul de la distance avec tous les exemples, mais il existe des heuristiques pour réduire le nombre d'exemples à prendre en compte (Oumiloud H.2014).

- **Principe de l'algorithme des k plus proches voisins**

Le principe de l'algorithme des k plus proches voisins repose sur le regroupement d'individus en fonction de leur voisinage.

L'algorithme des K plus proche voisins est basé sur les éléments principaux suivants :

1. Le premier principe est d'utiliser le nombre des cas les plus proches (K) et une métrique de similarité pour mesurer le plus proche voisin.
2. Le deuxième principe est de considérer la spécification de la valeur de K à chaque application de l'algorithme qui détermine le nombre de cas considérés pour prédire la classe d'un nouveau cas.
3. La méthode des k plus proches voisins consiste à regrouper des individus en fonction de leur voisinage, en réalité, chaque individu est affecté à la classe la majoritairement renvoyée

parmi ses k plus proches voisins. Donc, un ensemble d'apprentissage est requis contenant les différentes classes afin de prédire la classe d'un nouvel individu (Koudri M.2011).

IV.3.2. L'analyse en composantes principales

L'Analyse en Composantes Principales (ACP) fait partie du groupe des méthodes descriptives multidimensionnelles appelées méthodes factorielles, qui consiste à transformer des variables liées entre elles (dites "corrélées" en statistique) en nouvelles variables décorrélées les unes des autres. Ces nouvelles variables sont nommées "composantes, principales", ou axes principaux. Elle permet au praticien de réduire le nombre de variables et de rendre l'information moins redondante (Miloudi N. 2014).

- **L'objectif de l'ACP**

Nous avons mentionné que nous allons utiliser simultanément des descripteurs obtenus suite à l'analyse des zone d'intérêt par les lois puissance ainsi que ceux obtenus à partir des ondelettes de Haar dans le but de bénéficier de leur apport complémentaire ; par conséquent, nous appliquerons une analyse en composantes principales (ACP) pour :

- Etudier et visualiser les corrélations entre les descripteurs dans le but de réduire le nombre de descripteurs à mesure.
- La génération de descripteurs non corrélés consistant en des combinaisons linéaires des descripteurs de départ, par la suite, nous utiliserons ces descripteurs dans les méthodes de modélisation, plus précisément, l'analyse discriminante afin de classifier une région d'intérêt requête en tumeur maligne ou bénigne ou bien une région seine ne présentant pas de lésion.

- **Principe de l'ACP**

Les étapes de l'analyse en composantes principales sont comme suit :

1. Ranger les individus en ligne et les variables quantitatives en colonnes

$$H=\begin{bmatrix} var^1 & var^2 & var^3 \\ x_1^1 & x_1^2 & x_1^3 \\ x_2^1 & x_2^2 & x_2^3 \\ x_3^1 & x_3^2 & x_3^3 \\ x_4^1 & x_4^2 & x_4^3 \end{bmatrix}$$

2. Calculer la moyenne de chaque variable (colonne) :

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

3. Calculer l'écart type de chaque variable (colonne) :

$$\text{Dev} = [\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2]^{\frac{1}{2}}$$

4. Centrer les variables :

$$x_c = x - \bar{x}$$

5. Réduire les variables :

$$\text{Red} = \frac{x_c}{Dev}$$

6. Calculer les corrélations

$$\text{Cor} = \text{Red}^t \times \text{Red}$$

Il faut transposer la matrice puis la multiplier par elle-même. C'est une matrice carrée dont la diagonale est l'autocorrélation de chaque variable et les autres sont les cross-corrélation.

7. Trouver les valeurs propres et les vecteurs propres correspondants de cette matrice (matrice de corrélation).
8. Trier les valeurs propres obtenues en ordre décroissant de valeur $\lambda_1 > \lambda_2 > \lambda_3$
9. Pourcentage d'inertie expliquée

$$\frac{\lambda_i}{\sum_{i=1}^{p} \lambda_i} \times 100$$

Le pourcentage de variance expliquée par les axes factoriels retenus est obtenu par somme de leurs valeurs propres divisées par la trace. Cette valeur permet d'évaluer en quelque sorte la quantité d'information initiale recueillie par chaque axe.
Le pourcentage d'inertie expliquée par un axe factoriel

10. Détermination des composantes principales

Revient à multiplier la matrice réduite avec la matrice diagonale des vecteurs propres

$$\text{Proj} = \text{Red} * \text{VP}$$

L'opérateur ' * ' est dans ce cas le produit scalaire $x .C_{1,x} + y. C_{1,y}$. Les axes choisis sont les axes principaux trouvés par l'ACP. Nous représentons donc les données sur ces axes. Les coefficients de la projection constituent les caractéristiques des données.

11. Coefficient de corrélation entre les composantes principales et les variables d'origine

Il est souvent utile d'étudier les corrélations entre les variables d'origine *x* et les nouvelles variables C.

Les corrélations entre les axes factoriels et les variables initiales indiquent la qualité de représentation de la variable sur l'axe. Les coordonnées des variables dont égales aux corrélations avec les axes. Plus une corrélation entre une variable et un axe est forte plus la variable est proche de l'axe.

Il ne faut interpréter les axes qu'à partir des variables les mieux représentées, c'est-à-dire celles dont la corrélation est proche de 1 en valeur absolue.

Le coefficient de corrélation entre une composante principale C_j et une variable x_i est égale à :

$$r_{ij} = \frac{covariance\ C_j x_i}{\sqrt{variance\ c_j}\sqrt{variance\ x_i}} \quad \text{tq } r \in [1, -1]$$

Si *r* est proche à 1 alors C et *x* sont très bien corrélés ou lié.
Si *r* est proche à 0 alors C et *x* ne sont pas corrélés.
Si *r* est proche à -1 alors C et *x* sont très bien anti-corrélé (Miloudi N. 2014).

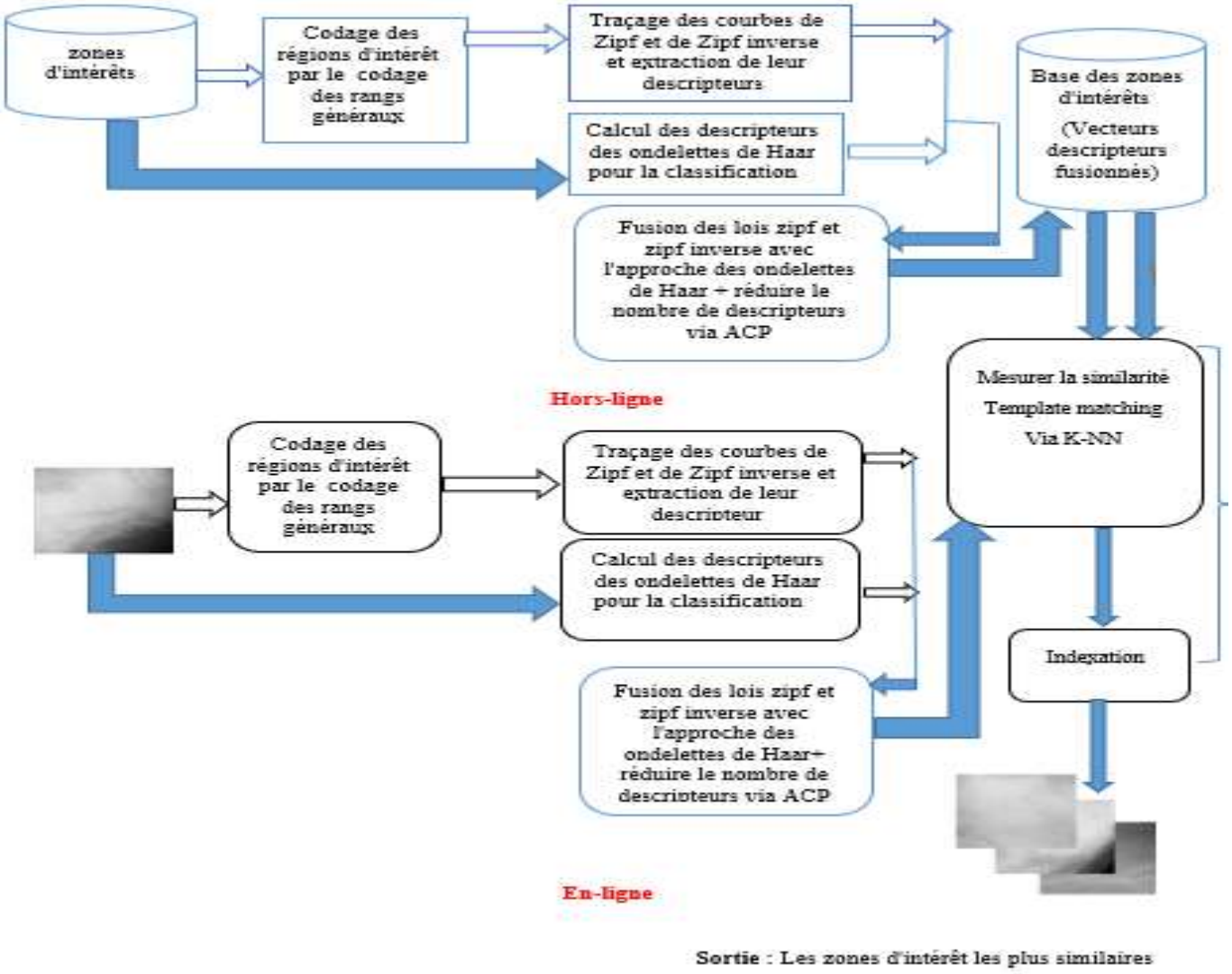


Figure IV.1 Architecture du système d'aide au diagnostic médical du cancer du sein par la méthode template-matching.

IV.4. Matériel et outils utilisés pour le développement du système proposé

Le système que nous avons proposé a été développé et testé sous un ordinateur plutôt performant qui possède les caractéristiques suivantes :

Nom du modèle : PHP

Processeur : Intel Core i5- 4200U

Fréquence du processeur : 2,30 GHZ

Mémoire vive : 6 Go

Capacité de stockage principal : 750 Go

Système d'exploitation : Windows 10

Edition Premium : 64-bits

- **L'environnement MATLAB**

MATLAB est un environnement puissant, complet et facile à utiliser destiné au calcul scientifique. Il apporte aux ingénieurs, chercheurs et à tout scientifique un système interactif intégrant calcul numérique et visualisation. C'est un environnement performant, ouvert et programmable qui permet de remarquables gains de productivité et de créativité. Et encore le MATLAB est un environnement complet, ouvert et extensible pour le calcul et la visualisation. Il dispose de plusieurs centaines (voire milliers, selon les versions et les modules optionnels autour du noyeau Matlab) de fonctions mathématiques, scientifiques et techniques. L'approche matricielle de MATLAB permet de traiter les données sans aucune limitation de taille et de réaliser des calculs numériques et symboliques de façon fiable et rapide. Grâce aux fonctions graphiques de MATLAB, il devient très facile de modifier interactivement les différents paramètres des graphiques pour les adapter selon nos souhaits (Bentifour K et al.2016).

- **Les particularités de MATLAB ?**

MATLAB permet le travail interactif soit en mode commande, soit en mode programmation ; tout en ayant toujours la possibilité de faire des visualisations graphiques. Considéré comme un des meilleurs langages de programmations (C ou Fortran), MATLAB possède les particularités suivantes par rapport à ces langages :

• La programmation facile,
• La continuité parmi les valeurs entières, réelles et complexes,
• La gamme étendue des nombres et leurs précisions,
• La bibliothèque mathématique très compréhensive,
• L'outil graphique qui inclus les fonctions d'interface graphique et les utilitaires,
• La possibilité de liaison avec les autres langages classiques de programmations (C ou Fortran) (Bentifour K et al.2016).

Dans MATLAB, aucune déclaration n'est à effectuer sur les nombres. En effet, il n'existe pas de distinction entre les nombres entiers, les nombres réels, les nombres complexes et la simple ou double précision. Cette caractéristique rend le mode de programmation très facile et très rapide. En Fortran par exemple, une subroutine est presque nécessaire pour chaque variable simple ou double précision, entière, réelle ou complexe. Dans MATLAB, aucune nécessité n'est demandée pour la séparation de ces variables.

La bibliothèque des fonctions mathématiques dans MATLAB donne des analyses mathématiques très simples. En effet, l'utilisateur peut exécuter dans le mode commande n'importe quelle fonction mathématique se trouvant dans la bibliothèque sans avoir à recourir à la programmation.

Pour l'interface graphique, des représentations scientifiques et même artistiques des objets peuvent être créées sur l'écran en utilisant les expressions mathématiques. Les graphiques sur MATLAB sont simples et attirent l'attention des utilisateurs, vu les possibilités importantes offertes par ce logiciel (Bentifour K et al.2016).

- **Généralité sur MATLAB**

Matlab est un logiciel de calcul numérique produit par MathWorks. Il est disponible sur plusieurs plateformes. Matlab est un langage simple et très efficace, optimisé pour le traitement des matrices, d'où son nom. Matlab contient également une interface graphique puissante, ainsi qu'une grande variété d'algorithmes scientifiques. On peut enrichir Matlab en ajoutant des toolbox, profilées pour des applications particulières (traitement des images, analyses statistiques, optimisation, etc.) (Yassine A et al .2016).

La figure IV.2 présente la fenêtre principale de Matlab :

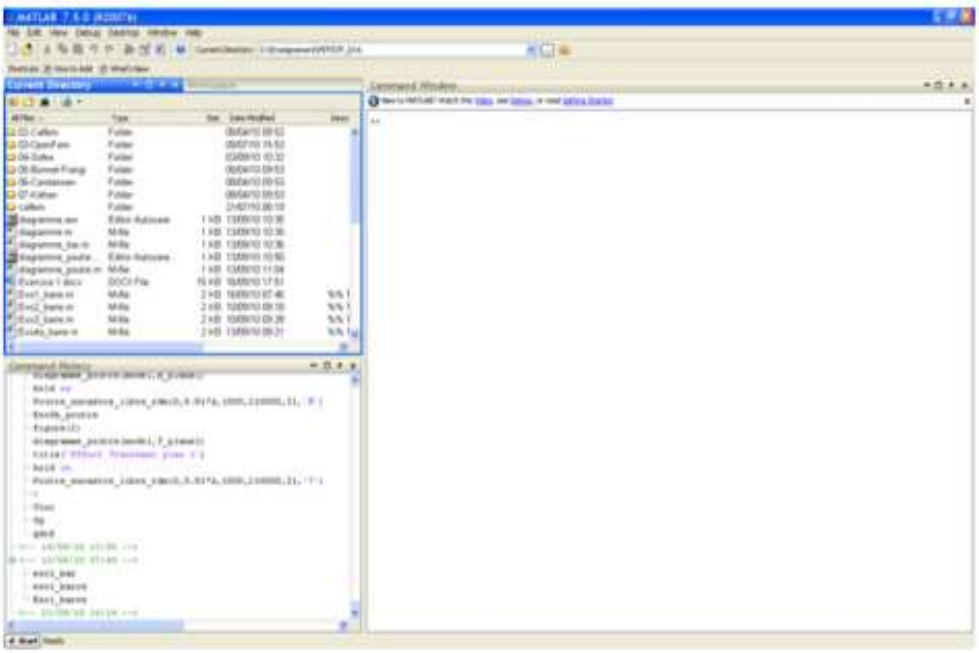


Figure IV.2 : Fenêtre principale de Matlab.

- **Représentations graphiques**

Les bibliothèques de Matlab proposent un très grand nombre de fonctions pour la manipulation d'objets graphiques. Nous ne présentons ici que quelques principes de base, utiles pour la visualisation de courbes. Si nous nous concentrons particulièrement sur la représentation graphique à 2 dimensions, il est possible d'aller bien plus loin : graphismes 3D (courbes, maillages, surfaces...), édition d'IHM (graphical user interface, GUI), animations... Quelques exemples de représentations 3D sont brièvement présentés dans (Yassine A et al .2016).

IV.5. La base des mammographies utilisée

Nous allons utiliser la base MIAS pour l'élaboration de notre travail, elle contient un ensemble de mammographies englobant 120 images mammaires numérisées et ayant une pathologie connue (24 portant une tumeur maligne, 37 portant une tumeur bénignes, 59 sans la présence de lésions). Les images mammaires sont de taille 1024 × 1024 pixels. Notons que la méthode de segmentation des zones d'intérêt a été présentée dans (Hamoud M.2015).

IV.6. Fonctionnement du système développé

Nous accédons à notre système a travers Matlab, l'interface principale du système sera obtenue en cliquant sur le bouton RUN du Matlab comme mentionné sur la figure IV.3.

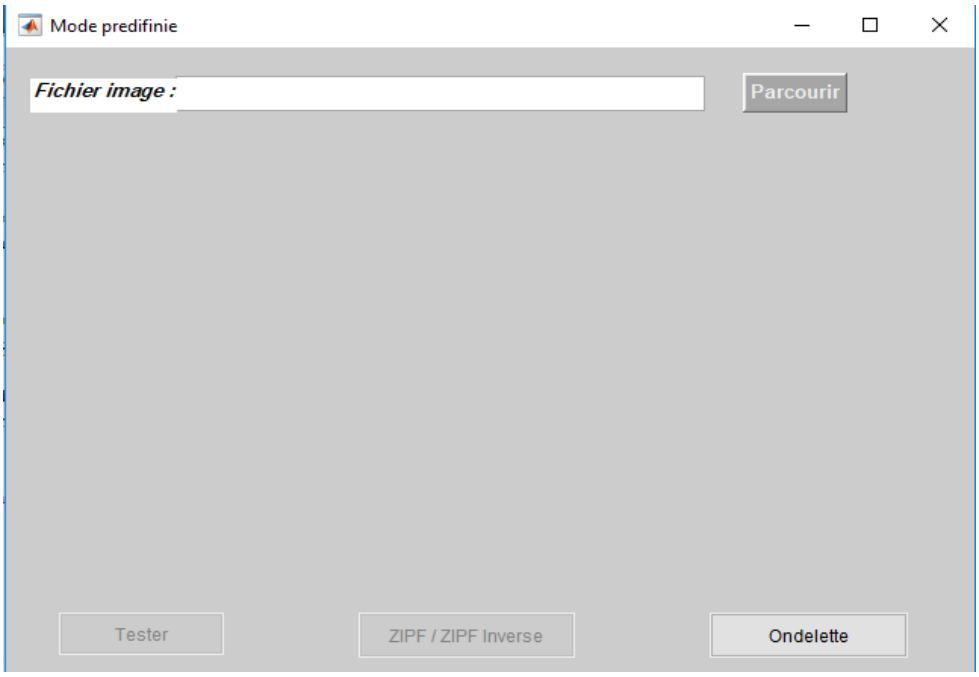


Figure IV.3 L'interface principale du système Mammo-ZipOnd.

Quant à la figure IV.4, elle reflète la sortie du système Mammo_ZipOnd. Une fois que le radiologue introduit la zone d'intérêt requête au système, ce dernier renverra et affichera les

zones d'intérêt similaires pour générer le résultat de la classification : tumeur maligne, tumeur bénigne ou tissu normal en cliquant sur le bouton « Tester ».

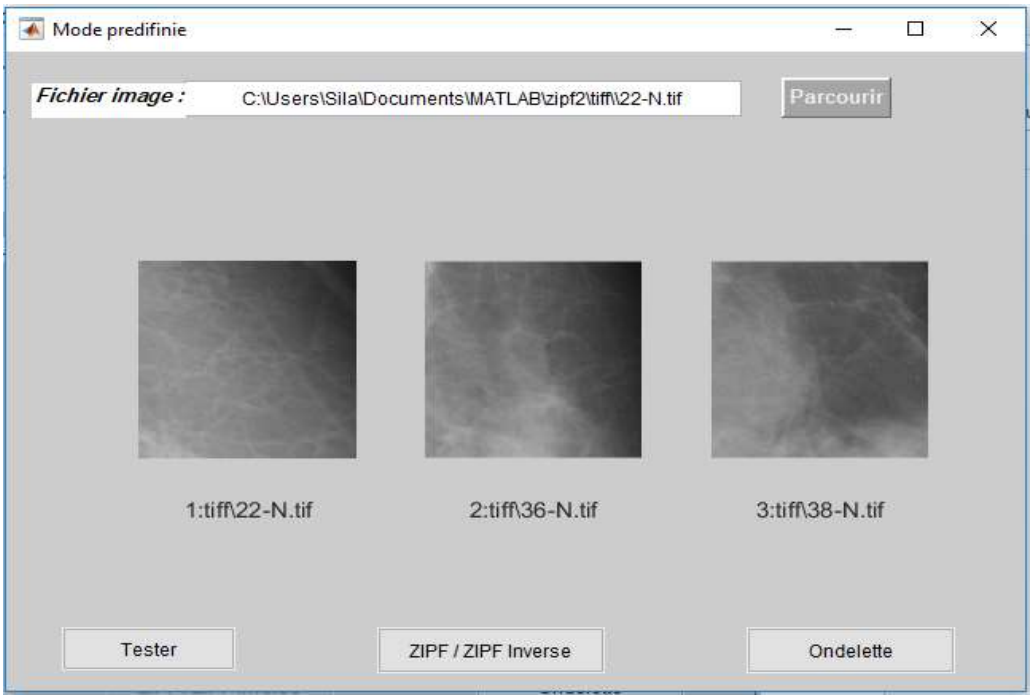


Figure IV.4 Processus de décision médicale issue de l'utilisation de la sortie de Mammo_ZipOnd par la considération de k=3.

Nous constatons que les 3 régions d'intérêt renvoyées par le processus de recherche portent un tissu normal, menant le radiologue à considérer la région d'intérêt requête comme étant d'un tissu sein, ce qui est juste. Notons également, le système Mammo_ZipOnd permet l'affichage des zones d'intérêt appartenant à la même classe pathologique de la zone d'intérêt requête.

Le système permet de calculer les ondelettes de Haar de la zone d'intérêt requête comme mentionné sur la figure IV.5.

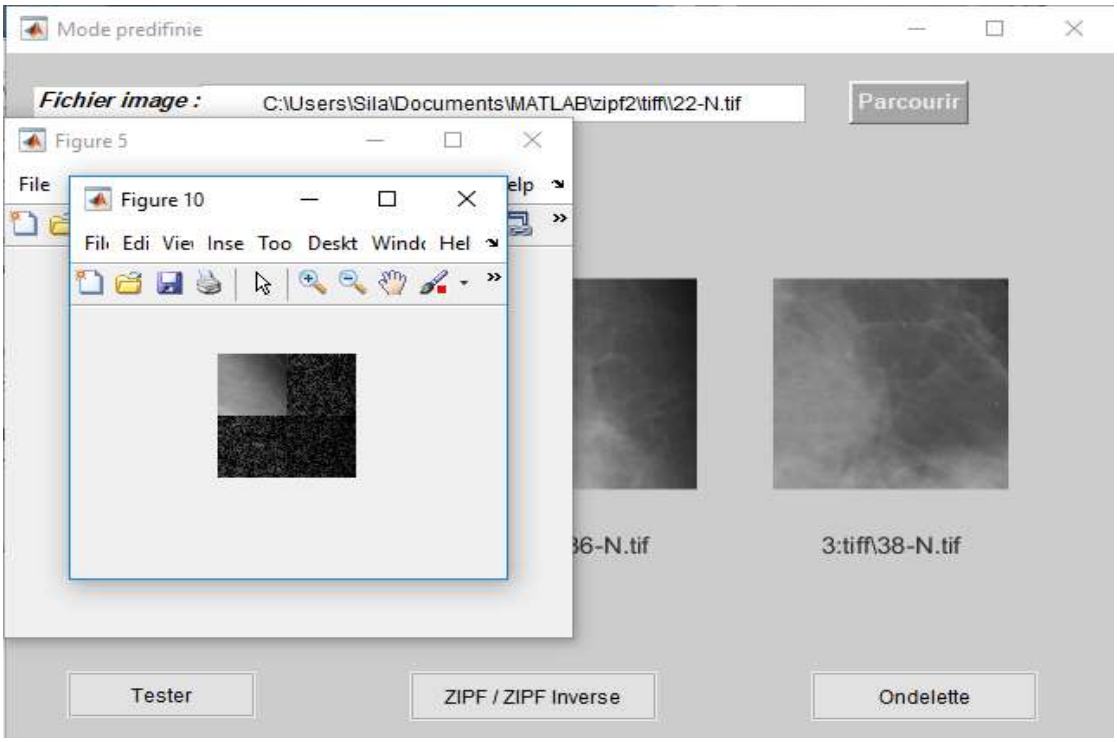


Figure IV.5 Calcul des ondelettes de Haar de la zone d'intérêt requête.

Sur la figure IV.6, nous évoquons un autre exemple de diagnostic fourni par notre système où il réussi à bien diagnostiqué une zone d'intérêt portant une tumeur bénigne.

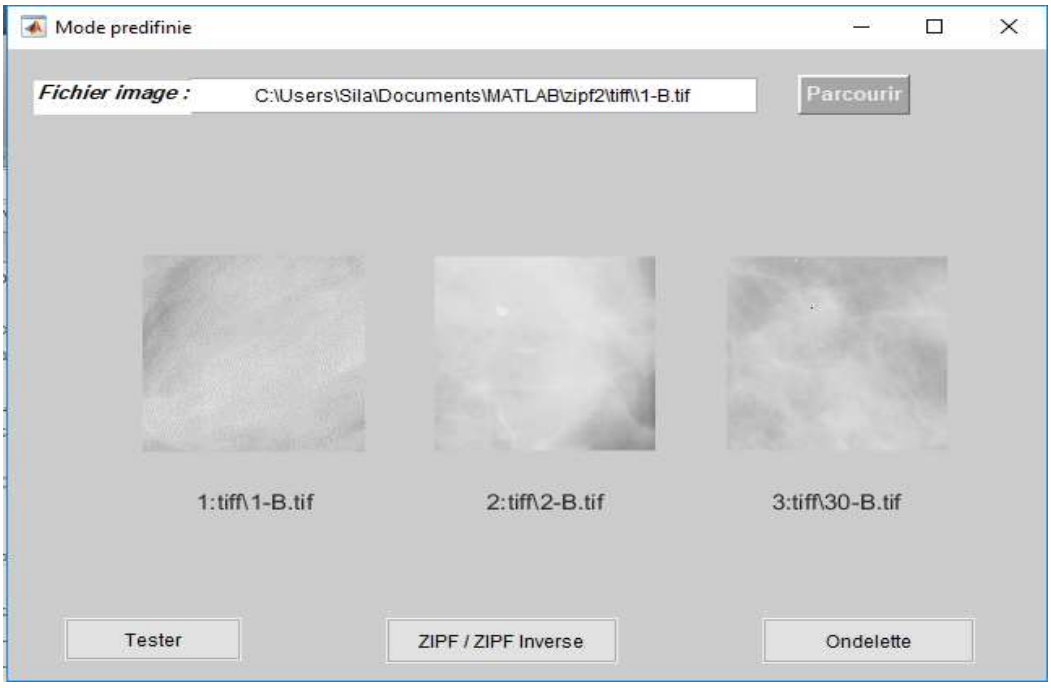


Figure IV.6 Processus de décision médicale issue de l'utilisation de la sortie de Mammo_ZipOnd par la considération de k=3 et d'une zone d'intérêt requête portant une tumeur bénigne.

Sur la figure IV.7, nous présentons un exemple où notre système a échoué à déterminer la classe pathologique d'une zone d'intérêt requête.

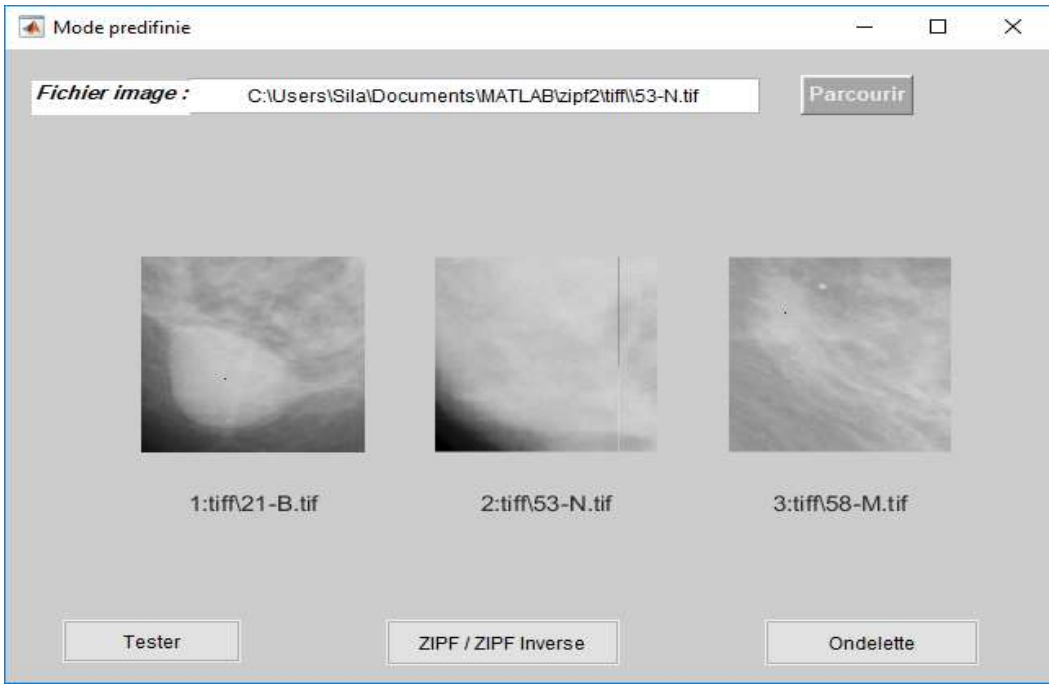


Figure IV.7 Echec du système Mammo-ZipOnd dans le processus de classification.

IV.7. Evaluation des performances

Dans le but d'évaluer les performances de notre système, nous utilisons le critère du taux de bonne classification.

Pour cela, nous partitionnons la base de données MIAS en considérant la validation croisée qui divise l'échantillon de taille n en deux sous échantillons : celui d'apprentissage (supérieur à 60 % de l'échantillon) et celui de test.

La base MIAS a été partitionnée en deux parties, une > 60% de l'échantillon pour l'apprentissage (80 images mammaires) et la deuxième partie (40 images mammaires) pour le test.

Nous avons effectué nos tests en considérant K=3 et nous avons obtenus le taux de classification en divisant le nombre de zones d'intérêts dont la pathologie est correctement identifiée par le nombre total de zones d'intérêts requêtes. Le taux de classification obtenus est 82.5%.

Nous envisageons d'améliorer l'étape de la sélection des descripteurs générés par les ondelettes de Haar afin de surpasser les performances de hamoud et al utilisant les lois de Zipf uniquement et aboutissant à un taux de classification de 97%.

IV.8. Bilan

Il est important de mentionner que ce mémoire vise la caractérisation de la texture des régions d'intérêt par une fusion de deux méthodes : les lois puissance Zipf et Zipf inverse ainsi que la transformée en ondelettes de Haar pour bénéficier de l'apport complémentaire de ces deux techniques d'analyse de la texture pour donner un second diagnostic assisté par ordinateur aux radiologues durant l'analyse des images mammaires.

IV.8 Conclusion

Dans ce chapitre, nous avons présenté le principe sur lequel nous nous somme basés pour la réalisation de notre système « Mammo_ZipOnd » d'aide au diagnostic médical du cancer du sein.

En effet, ce dernier s'appuie sur l'utilisation simultanée des lois puissance : Zipf et Zipf inverse et des ondelettes de Haar pour l'indexation et la recherche par le contenu des zones d'intérêts extraites à partir des images de la mammographie.

Nous pouvons affirmer que notre système aboutit à un taux de classification encourageant de l'ordre de 82.5%.

Conclusion générale

Dans ce mémoire, nous avons abordé le problème de l'aide au diagnostic médical du cancer du sein assisté par ordinateur. En effet, les radiologues font passer un pourcentage élevé de cas de cancer du sein, à cause de la fatigue et la limitation de l'analyse visuelle.

A cet effet, le diagnostic de tumeurs assistées par ordinateur a été proposé comme un deuxième avis aux radiologues pour l'élaboration du diagnostic final.

Notre approche consiste en une fusion des lois puissance : Zipf et Zipf inverse et les ondelettes de Haar pour la caractérisation et la quantification de la texture des zones d'intérêt extraites à partir des mammographies.

Nous nous sommes basées sur l'utilisation des mammographies similaires pour fournir l'aide assisté par ordinateur aux radiologues afin d'effectuer leur diagnostic, notre motivation nous ai parvenu de l'hypothèse que les radiologues répondent mieux à un résultat de diagnostic assisté par ordinateur sous forme d'un affichage des mammographies similaires au cas en cours d'analyse, au lieu des chiffres abstraits fournis par les système de diagnostic basés sur la sortir d'un classifieur.

Dans ce sens, nous avons développé un système d'indexation et de recherche des mammographies par le contenu se basant sur la fusion des descripteurs texturaux générés par l'application des deux méthodes citées précédemment de caractérisation de la texture. Entre autre, une analyse en composantes principales (ACP) a été appliquée aux vecteurs descripteurs pour apaiser la dimension ainsi que générer de nouveaux descripteurs discriminants. Pour le processus de classification, la technique template-matching a été utilisé qui affecte la région d'intérêt requête à la classe pathologique la majoritairement renvoyée par le processus de recherche via l'algorithme des K plus proches voisins.
Suite à la présentation des fonctionnalités principales de notre système d'indexation et de recherche des mammographies par le contenu pour l'aide au diagnostic médical du cancer du sein ainsi que l'évaluation de ce système, nous avons obtenus un taux de bonne classification encourageant de l'ordre de 80%.

Perspectives

Dans nos futurs travaux, nous utiliserons d'autres techniques d'extraction de l'information de la texture à partir de l'analyse d'image par les ondelettes de Haar.

Bibliographie

- Aptoula, E. Lefèvre, S. (2011). Morphological Texture Description of Grey-Scale and Color Images. Advances in Imaging and Electron Physics, 169. pp. 1-74

-Anne Sabourn et joseph salmon Travaux Pratiques 2015 «méthodes des k-plus proches voisins »

-BOUKLI.H Abdelkarim et ABDELALI .W Mémoire.2014 «Cancer du sein pris en charge au niveau du service de gynéco –obstétrique » Université Tlemcen

- Bentifour Kheira et Mimouni Mémoire Master Fatima 22 mai 2016 «REALISATION D'UNE INTERFACE POUR LA CLASSIFICATION DES DONNE MEDICALES» université Tlemcen
- Centre international de recherche sur le cancer. Article .12 décembre 2013 «Dernières statistiques
- mondiales sur le cancer En augmentation à 14,1 millions de nouveaux cas en 2012 : L'augmentation marquée du cancer du sein demande des réponses» Lyon / Genève

-Centre de santé et de services sociaux Alphonse – Desjardins. Article décembre 2014 «Mammographie : examen de base et incidences complémentaire» centre hospitalier affilié universitaire de Levi.

- Edmond Boyer m2iico-ufrima Cour 3 juillet 2006 «vision prétraitement et Analyse d'images». L'analyse d'images regroupe ... - DocPlayer.fr. Edmond.Boyer@imag.fr,

-Eve Mathieu-Dupas Article · January 2010 «algorithme des k plus proches voisins pondérés et application en diagnostic» 42èmes Journées de Statistique, 2010, Marseille, France, France. 2010.

- Koudri Mohammed 2011 mémoire finn d'étude «Modèle de mélange Gaussien. Application sur image cytologique »université Tlemcen

-Florian SCOTTE, et all, 2002, édition Marketing S.A, Paris.

- Griff, S. K. Dershaw, D. D. (2002) Oncologic Imaging (Second Edition). Breast Cancer. pp 265-294.

-Grower-Thomas, K. (2009). Hughes, Mansel & Webster's Benign Disorders and Diseases of the Breast (Third Edition). Imaging of the Breast. pp 71-80

- Mitnick. J. (2005). Mammographic Diagnosis of Breast Cancer. Breast Cancer (Second Edition). pp. 211-245
- Maouchich Salima et mbakli lila 2017« L'anxiété chez les femmes atteintes de cancer du sein étude de six cas à CHU de Bejaia » université Bejaia
- Minh-Nguyen, T. Jonathan-Wu, Q. M. (2013). A fuzzy logic model based Markov random field for medical image segmentation. Evolving Systems, 4. pp. 171-181.
- M3.21 – Les images informatiques et leurs traitements IUT d'Arles – DUT SRC Article – 2010-2011

Bibliographie

- Minh-Nguyen, T. Jonathan-Wu, Q. M. (2013). A fuzzy logic model based Markov random field for medical image segmentation. Evolving Systems, 4. pp. 171-181.

  -Weidong Cai, T. Kim, J. Dagan Feng, D. (2008). Content-Based Medical Image Retrieval. Elsevier. pp. 83-113. M3.21 – Les images informatiques et leurs traitements IUT d'Arles – DUT SRC Article – 2010-2011

- Mr. MOUSSATI Omar. Diplôme de Magistère 2016 « Classification des données de biopuces» Université d'Oran.
- N. Miloudi. (2014). «Contribution à l'étude de la vulnérabilité des réservoirs en béton par analyse des composantes principale ». Thèse de doctorat. Université Tizi Ouzou.
- Oumiloud Horiya et Mokeddem Asma 2014 thème «Classification non supervisée : Application de k-means» université Tlemcen.
- Philippe B, Jean-M, Jean p, Didier D, Christine G, Annick M, Sylvie Ph, Rachid Z, Josiane Z, Henri M, Submitted on 23 Jun 2017«Analyse d'images : Filtrage et segmentation» HAL Id: hal-00706168 https://hal.archives-ouvertes.fr/hal-00706168
- Radu HORAUD et Olivier MONG Submitted on 3 May 2011 «Vision par ordinateur outils fondamentaux» HAL Id: inria-00590049 https://hal.inria.fr/inria-00590049
- Caron, Y. (2004). Contribution de la loi de Zipf à l'analyse d'images. Thèse de doctorat. Université de Tour.
- Shen, H. C. Srivastava, D. (1996). Texture Representation and Classification: The Feature Frequency Matrix Approach. Advances in imaging and electron physics, 95.pp. 387-407

  -Souad Meziane Tani, Abdelhafid Bessaid 2011. Techniques d'indexation d'images Médicales par contenu
- Sonka, M. Hlavac, V. Boyle, R. (2008). Image Processing, Analysis, and Machine Vision. International Student Edition. Third Edition. Thomson Learning, part of the Thomson Corporation. USA
- Tuceryan, M. Jain, A. K. (1998). Texture Analysis. The Handbook of Pattern Recognition and Computer Vision (2nd Edition). World Scientific Publishing Co. pp. 207-248
- TERKI Hakima et BENYELLES Khadidja. MASTER 2013«Analyse des images mammographiques en vue de la détection et la caractérisation des Microcalcifications mammaires» Université Tlemcen
- TOUAMI Rachida. Mémoire de Magister.2011«SEGMENTATION DES IMAGES MEDICALES PAR ONDELETTES» Université d'Oran.
- Centre de santé et de services sociaux Alphonse – Desjardins. Article décembre

Bibliographie

- Yassine Ariba – Jérome Cadieux Article 22 mai 2016« manuel Matlab »DépartementsGEI & Mécanique. Icam de Toulouse.
- Zivian, M. T. Gershater, R. (2008). The Accuracy of Diagnostic Radiology.Cancer Imaging: Instrumentation and Applications. pp 109-118.

Liste des abréviations

Abréviation	**Description textuelle**	
ACP	Analyse en Composantes Principales	2
CADe	Computer Aided Detection	3
CADx	Computer Aided Diagnosis	4
CBIR	Content Based Image Retrieval	5
CBMIIR	Content Based Mammogram Image Indexing and Retrieval	6
k-NN	k-Nearest Neighbors	7
MIAS	Mammographic Image Analysis Society	8
ROI	Region Of Interest	9