

République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
Université SAAD DAHLAB Blida
Faculté des Sciences
Département d'Informatique



Mémoire de Fin d'Etudes

Mémoire Présenté par :

Menacer Sara Mekerkeb Aberrane Sara

En vue d'obtenir le diplôme de master

Domaine : Mathématique et informatique

Filière : Informatique

Spécialité : Informatique

Option : Ingénierie de logiciel

Thème

Segmentation basée clustering et reconstruction 3D des images

CT-SCAN

Mr. CHIKHON

President

Mr. Benhabiles.H

Promoteur

Mr. Belkacem. H

Encadreur

Année universitaire : 2016/2017

Résumé :

L'objectif de ce travail est de réaliser une plateforme Java d'extraction et de visualisation 3D des zones d'intérêt adéquate à la nature des images CT Scan.

Pour cela nous avons divisé notre travail en deux parties, la première est la détection qui s'appuie sur l'aspect morphologique de l'image et sur les algorithmes de clustering existants dans la littérature (nous avons utilisé l'algorithme d'expectation-maximisation), la fin de cette phase va nous permettre à avoir des clusters labélisés qui représente une anomalie ou un organe humain.

La deuxième se résume à l'empilement des clusters et de visualisation 3D en utilisant les différents toolkit existants.

Mots clés : CT-SCAN, segmentation, clustering, expectation-maximisation, empilement, visualisation 3d.

Abstract :

The objective of this work is to create a Java platform (Netbeans) of extraction and 3D visualization of areas of interest suitable to the nature of CT-Scan images. For this we have devise our work into two parts, the first is the detection which relies on the morphological aspect of the image and on the clustering algorithms existing in the literature (we used the algorithm of Expectation - Maximization), the end of this phase to allow us to have labels clusters that represent an anomaly or a human organ.

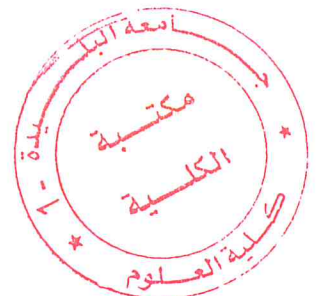
The second is the stacking of clusters and 3d visualization using the various existing tools.

Keywords: CT-SCAN, segmentation, clustering, expectation -maximization, stacking, 3d visualization.

المخلص :

الهدف من هذا العمل هو تحقيق استخراج منصة جافا (نتبيانس) والتصوير الثلاثي الابعاد من المجالات ذات الاهتمام المناسب لطبيعة صور الأشعة المقطعية. لهذا جزأنا العمل إلى قسمين، الأول هو الكشف عن المظهر الصرفي للصورة وخوارزميات التجميع الموجودة في الأدب (استعملنا خوارزمية تعظيم التوقعات)، ونهاية هذه المرحلة تسمح لنا لدينا مجموعات وصفتم تمثل حالة شاذة أو الأعضاء البشرية.

والثاني هو تكديس المجموعات والتصوير الثلاثي الابعاد باستخدام مختلف الأدوات الموجودة. الكلمات المفتاحية : الصور المقطعية، التجزئة ، التجميع، خوارزمية تعظيم التوقعات ، التصوير الثلاثي الابعاد.



REMERCIEMENTS

Nous tenons tout d'abord à remercier ALLAH tout puissant de nous avoir donné le courage et la force pour la réalisation de ce modeste travail.

Nous adressons nos vifs remerciements à Mr Benhabiles notre promoteur enseignant à l'université Saad Dahleb de Blida pour ses conseils, sa collaboration et ses encouragements.

Nous exprimons notre grande reconnaissance et remerciement à nos encadreur Mr. Hachemi Belkacem et Mme. ALIM Fatiha pour leurs aides, leurs sollicitudes, leurs conseils et de nous avoir orienté et mis à notre disposition tous les moyens nécessaires durant notre travail.

Nous remercions très sincèrement notre jury d'avoir bien voulu d'accepter d'examiner notre travail.

Enfin, on remercie toute personne ayant contribué de près ou de loin à la progression de ce projet.

DEDICACE DE SARA

Que ce travail témoigne de mes respects :

A mes parents : Grâce à leurs tendres encouragements et leurs grands sacrifices, ils ont pu créer le climat affectueux et propice à la poursuite de mes études. Aucune dédicace ne pourrait exprimer mon respect, ma considération et mes profonds sentiments envers eux. Je prie le bon Dieu de les bénir, de veiller sur eux, en espérant qu'ils seront toujours fiers de moi.

A ma sœur 'IMENE' et à mes frères 'ZINEDDINE' et 'AYMEN'. Sans oublier Nadjet et Ibrahim.

A mon mari FAIZ: qui va trouver ici l'expression de mes sentiments de respect et de reconnaissance pour le soutien qu'il n'a pas cessé de me porter.

A toute ma famille et ma belle-famille 'grand' et 'petit'.

A ma chère binôme SARA que j'aime énormément et sa famille.

A toutes mes amies : KHADIDJA ,AMINA ,AMIRA, IMENE, SARAH , Ils vont trouver ici le témoignage d'une fidélité et d'une amitié infinie.

A toute personne qui m'a soutenu de près ou de loin.

DEDICACE DE SARA

Je dédie ce modeste travail à :

A mes parents .Aucun hommage ne pourrait être à la hauteur de l'amour Dont ils ne cessent de me combler. Que dieu leur procure bonne santé et longue vie.

Et bien sûr à mon frère Tayeb, à ma sœur Asma ,et à Nabila, sans oublier Mohamed, Younes et Youcef que j'aime.

A ma très chère binôme SARA que j'aime très fort.

A mes chères amies d'être toujours à mes côtés : Chanez , Wissem , Khadija , Meriem ...

Et Enfin à tous ceux qui ont contribué de près ou de loin pour que ce projet soit possible.

TABLE DES MATIÈRES

Introduction générale.....	01
Chapitre 01 : Généralités.....	
1.1. Introduction.....	03
1.2. Les techniques de l'imagerie médicale :	03
1.3. Définition d'une tomodensitométrie	03
1.3. 1. L'utilité des CT-SCAN.....	05
1.3.2. Fonctionnement de la tomodensitométrie	05
1.3.3. Caractéristiques des CT-SCAN.....	06
1.3.4. Différence entre CT-SCAN et IRM	06
1.3.5. Exemples de CT-SCAN	06
1.4. Caractéristiques de l'image	07
1.4.1. Pixel	07
1.4.2. Poids de l'image	07
1.4.3. Transparence	08
1.4.4. Luminance	08
1.4. 5. Contraste	08
1.4.6. Homogénéité	08
1.4.7. Connexité.....	08
1.4.8. Région	09
1.5. Les différents formats d'images	09
1.5.1. Format vectorielle	09
1.5.2. Format matricielle	09
1.6. Conclusion.....	09

Chapitre 02 : Etat de L'art sur le clustering.....

2.1. Introduction.....	11
2.2. Machine Learning c'est quoi	11
2.2.1. Différence entre apprentissage automatique supervisé et non supervisé.....	11
2.2.2. Quelques exemples pratiques de Machine Learning	12
2.2.3. Méthodes de machine learning	12
2.2.4. Clustering.....	14
2.2.4.1. Définition	14
2.2.4.2. Principe général	14
2.2.4.3. Les exigences de Clustering	15
2.2.4.4. Exemples de clustering	15
2.2.4.5. Les types de Clustering	15
2.2.4.6. Exemples d'algorithmes	17
2.2.4.6.1. K-means.....	17
2.2.4.6.2. Méthodes hiérarchiques	19
2.2.4.6.3. Algorithme Fuzzy c-means de clustering	20
2.2.4.6.4. Algorithme d'Expectation-Maximisation.....	21
2.2.4.7. Avantage et Inconvénient des méthodes de clustering.....	25
2.3. Mesure de similarité	26
2.4. Conclusion.....	28

Chapitre 03 : Conception.....

3.1. Introduction.....	29
3.2. Architecture globale du système	29
3.3. Conversion d'image en matrice	31
3.4. Distribution	31
3.5. Histogramme	33
3.6. La méthode EM	35

3.6.1. choix de la méthode EM.....	35
3.6.2. Application de la méthode EM	35
3.7. Conversion du masque calculé dans l'étape précédente	37
3.8. Extraction d'un cluster	38
3.9. Empilement et reconstruction 3D des images ct-scan	40
3.10. Conclusion.....	43
Chapitre 04 : Implémentation, expérimentation et interprétation des résultats.....	
4.1. Introduction.....	43
4.2. Environnement de travail.....	43
4.2.1. Matériel.....	43
4.2.2. Langage.....	43
4.2.3 Environnement de développement	44
4.3. Données utilisées	45
4.4. Description d'application	46
4.4.1. Interface	46
4.4.2. Description des boutons	47
4.5. Conclusion.....	52
Conclusion générale.....	53
Bibliographie.....	54

LISTE DES FIGURES

Figure 1.1 : image qui représente la machine CT-SCAN	04
Figure 1.2. images CT-SCAN des poumons, colonne vertébrale et le foie.	06

Figure 2.1. Schéma qui explique la classification	13
Figure 2.2. Schéma qui explique la régression	13
Figure 2.3. Schéma qui explique le clustering	13
Figure 2.4. Graphique des 3 méthodes	13
Figure 2.5. Illustration de regroupement en clusters	14
Figure 2.6. Les deux types de clustering non-hiérarchique/hiérarchique	16
Figure 2.7 . Exemple d'un problème de discrimination à deux classes, avec une séparatrice linéaire : la droite d'équation $y=x$. Le problème est linéairement séparable	17
Figure 2.8. le dendrogramme de la hiérarchie H de la suite de partitions d'un ensemble {a, b, c,d, e}	20
Figure 3.1. système globale	30
Figure 3.2. Conversion d'une image de la base de données 1 en matrice	31
Figure 3.3. une distribution gaussienne	32
Figure3.4. la distribution d'une image avec $k=3$	33
Figure 3.5. histogramme de l'image 48 BD1	34
Figure 3.6. histogramme de l'image 8 BD	34
Figure 3.7. le déroulement d'EM dans notre programme	36
Figure 3.8. Image originale de BD1	38
Figure 3.9. Figure 3.10. Clustering d'image avec $k=5$	38
Figure 3.10. Extraction de cluster jaune de l'image 48 BD1	39
Figure 3.11. Application de notre programme sur l' image 33 de base de donnée Dataset 10 ...	40
Figure3.12. Empilement des images	41
Figure 3.13. La visualisation 3D d'une image: 48 BD1	41

Figure 4.1 Exemple d'image de la base de données images CVIP Spinal CT image CT --- SCAN (une couche de vertèbre).....	45
Figure 4.2. Exemple d'image de la base de données Dataset10 (une couche transversale de vertèbre).....	46
Figure 4.3.l'interface de la plateforme.....	47
Figure 4.4.Affichage du bouton « Ouvrir image ».....	47
Figure 4.5. Ouverture d'image de dataset4 choisie.....	48
Figure 4.6. Ouverture d'image de dataset10 choisie.....	48
Figure 4.7.Affichage d'image (dataset4) après l'application de la méthode EM.....	49
Figure 4.8. Affichage d'image (dataset10) après l'application de la méthode EM.....	49
Figure 4.9. Affichage du cluster choisi (image de dataset4).....	50
Figure 4.10. Affichage du cluster choisi (image de dataset10).....	50
Figure 4.11. Affichage du graphe.....	51
Figure 4.12. Visualisation 3D(ensemble d'images de dataset4).....	51
Figure 4.13. Visualisation 3D (ensemble d'images de dataset10).....	52

LISTE DES TABLEAUX :

Tableau 2.1. Tableau d'avantages et d'inconvénient des méthodes de clustering.....	26
---	-----------

Introduction générale

Des millions d'images médicales sont produites chaque année dans le monde pour établir un diagnostic ou contrôler une action thérapeutique. Ces images proviennent principalement du tomodensitomètre à rayons x ou scanner CT-Scan, de l'imagerie par résonance magnétique ou IRM, de l'imagerie ultrasonore ou échographie.

Les images médicales fournissent des informations sur la forme et le fonctionnement des organes du corps humain. Malheureusement, ces informations sont extrêmement difficiles à exploiter de manière quantitative et objective.

Le but de l'imagerie médicale est de créer une représentation visuelle intelligible d'une information à caractère médical. L'objectif est de pouvoir représenter sous un format relativement simple une grande quantité d'informations issues d'une multitude de mesures acquises selon un mode bien défini et d'extraire à partir des images acquises, les informations utiles au diagnostic, de révéler des détails difficiles à percevoir à l'œil nu, tout en évitant la création d'artefacts, faussement informatifs. Pour cela, le traitement fait appel à des outils, des algorithmes, qui permettent d'agir sur l'image.

Problématique :

Malgré les avancées technologiques dans le domaine médical et en particulier dans l'imagerie médicale, la segmentation qui est une étape critique du traitement d'image, celle-ci consiste à localiser dans une image les régions (ensembles de pixels) appartenant à une même structure. Les images CT - Scan souffrent d'un grand nombre d'artefacts, le bruit dans les données, dues à la mesure physique, les variations d'intensité au sein d'un même tissu ou d'une même structure, qui rendent leur segmentation automatique difficile.

L'objectif de notre contribution dans ce projet consiste à réaliser une plateforme java d'extraction et de visualisation 3D des zones d'intérêt adéquate à la nature des images CT-scan qui rend la tâche plus facile aux médecins pour faire un diagnostic complet et fiable qui permet la bonne segmentation des images CT-scan pour déterminer la pathologie et déterminer les traitements les plus appropriés.

Le but recherché est d'implémenter certaines méthodes de segmentation afin d'évaluer leurs performances et faire la reconstruction 3D des zones d'intérêts dans l'image. Nous avons opté sur une méthode basée sur le clustering de l'image, ce dernier va nous permettre de labéliser les pixels de l'image pour qu'on puisse extraire notre classe d'intérêt.

Ce travail est structuré en quatre chapitres :

Le chapitre. 1 : Généralités .

Dans ce chapitre nous allons présenter les notions essentielles liées à l'image CT-scan ainsi que les caractéristiques et les formats de l'image.

Introduction générale

Le chapitre. 2 État de l'art sur le clustering .

Dans ce chapitre nous allons présenter un état de l'art sur le clustering et ses différentes méthodes.

Le chapitre 3 :Conception .

Dans ce chapitre nous allons présenter l'application de la méthode d'expectation-maximisation ainsi que l'empilement et la reconstruction 3D des image CT-scan.

· **Le chapitre. 4** : Implémentation, expérimentation et interprétation des résultats.

Le dernier chapitre s'appuie sur les résultats obtenus afin de tester notre application sur une base d'images CT-scan **CVIP Spinal CT**, et **Dataset10**.

Nous achèverons notre mémoire par une conclusion générale et perspective.



CHAPITRE 1
Généralités

1.1. Introduction :

La première application d'imagerie médicale date de 1895 avec l'utilisation des rayons X [0]. C'est le début de la radiographie. Puis, les techniques s'améliorent et se diversifient tout au long du 20^e siècle avec l'arrivée du scanner, de l'échographie puis de l'IRM. L'imagerie médicale est aujourd'hui incontournable dans de nombreuses situations : établir un diagnostic, évaluer la sévérité d'une pathologie, l'efficacité d'un traitement.

Dans ce chapitre, nous allons présenter un aperçu général sur la technique (Radio ou Scanner) plus précisément la **tomodensitométrie** (Computed tomography CT).

1.2. Les techniques de l'imagerie médicale : [1]

La radiologie désigne l'ensemble des procédures ayant pour but le traitement ou le diagnostic d'une affection réalisées par un médecin Radiologue, sous contrôle d'un moyen d'imagerie. (Radio, Echo, Scanner, IRM).

- **Radio ou scanner** sont des rayons X qu'ils traversent le corps et c'est leur atténuation par les os, les organes, les tissus (musculaires, nerveux...) et les vaisseaux qui permet d'obtenir une image.
- **Echographie** sont des ondes, des vibrations qui se propagent dans le corps et sont réfléchies lorsqu'elles rencontrent un obstacle. L'appareil capte ces échos et reconstitue une image en mouvement.
- La résonance magnétique (**IRM**) n'émet aucun rayonnement. Cette technique repose sur le phénomène de résonance des atomes d'hydrogène du corps quand ils sont soumis à un champ magnétique. La machine fonctionne avec un puissant électro-aimant et produit des images en coupe du corps, en deux ou trois dimensions.

1.3. Définition d'une tomodensitométrie :[2]

La technique de tomodensitométrie a été inventée et développée par G.N.Hounsfield au Central Research Laboratories of EMI Ltd, Royaume-Uni en 1972 [2].

Une tomodensitométrie (CT) fait appel à des combinaisons traitées par ordinateur de nombreuses images de rayons X prises à des angles différents pour produire des images en coupe transversale («tranches virtuelles») de zones spécifiques d'un objet numérisé, permettant à l'utilisateur de voir à l'intérieur de l'objet sans le couper.

Le traitement de géométrie numérique est utilisé pour générer une image tridimensionnelle de l'intérieur de l'objet à partir d'une grande série d'images radiographiques bidimensionnelles prises autour d'un seul axe de rotation. L'imagerie médicale est l'application la plus courante de la tomographie par rayons X.

Ces images transversales sont utilisées à des fins diagnostiques et thérapeutiques dans diverses disciplines médicales.

CT produit un volume de données qui peuvent être manipulés afin de démontrer diverses structures corporelles en fonction de leur capacité à bloquer le faisceau de rayons X.

Bien que, historiquement, les images générées soient dans le plan axial ou transversal, perpendiculaire au grand axe du corps, les scanners modernes permettent de reformater ce volume de données dans divers plans ou même de représentations volumétriques (3D) des structures.

Les personnes responsables de l'examen CT sont appelées radiologues ou technologues en radiologie.

Une tomodensitométrie utilise des rayons X pour réaliser des images détaillées des parties du corps et des structures à l'intérieur.

Pendant le test, nous allongeons sur une table qui est attachée au scanner CT.

Le scanner CT est une grande machine en forme de beignet.

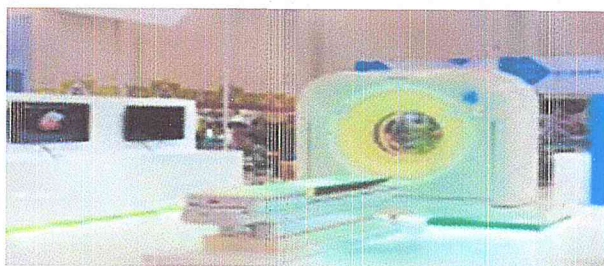


Figure 1.1 : image qui représente la machine CT-SCAN

1.3. 1. L'utilité des CT-SCAN :[2]

Les tomodensitogrammes sont également utilisés pour aider ou vérifier le succès d'une intervention chirurgicale, un exemple de ceci est quand un CT est utilisé pour guider une aiguille dans le corps pendant une biopsie de tissu.

CT scan images permettent au médecin de regarder l'intérieur du corps tout comme on regarderait à l'intérieur d'une miche de pain en le tranchant. Ce type de rayons X spécial, dans un sens, prend des «photos» de tranches du corps afin que les médecins peuvent regarder droit à la zone d'intérêt.

Les tomodensitogrammes sont fréquemment utilisés pour évaluer le cerveau, le cou, la colonne vertébrale, la poitrine, l'abdomen, le bassin et les sinus.

CT a révolutionné la médecine parce qu'elle permet aux médecins de voir des maladies qui, dans le passé, ne pouvaient être trouvées que lors d'une intervention chirurgicale ou à l'autopsie car elle est non invasive, sûre et bien tolérée.

1.3.2. Fonctionnement de la tomodensitométrie :[2]

Un tomodensitomètre émet une série de faisceaux étroits à travers le corps humain alors qu'il se déplace à travers un arc, contrairement à une machine à rayons X qui n'envoie qu'un seul faisceau de rayonnement, la photo finale est beaucoup plus détaillée qu'une image radiographique.

A l'intérieur du scanner CT il y a un détecteur de rayons X qui peut voir des centaines de différents niveaux de densité. Il peut voir les tissus à l'intérieur d'un organe solide. Ces données sont transmises à un ordinateur, qui construit une image en coupe transversale 3D de la partie du corps et l'affiche sur l'écran.

Parfois, un colorant de contraste est utilisé parce qu'il apparaît beaucoup plus clairement sur l'écran. Si une image 3D de l'abdomen est nécessaire, le patient peut avoir à boire un repas de baryum. Le baryum apparaît blanc sur le scan alors qu'il traverse le système digestif. Si les images plus bas du corps sont nécessaires, comme le rectum, le patient peut recevoir un lavement baryté. Si les images des vaisseaux sanguins sont la cible, le baryum sera injecté.

La précision et la vitesse de la tomodensitométrie peuvent être améliorées avec l'application de la CT en spirale. Le faisceau de rayons X prend une trajectoire en spirale pendant le balayage - il recueille des données continues sans aucun espace entre les images.

Pour un balayage en spirale de la poitrine, par exemple, le patient sera invité à tenir son souffle pendant quelques secondes.

1.3.3. Caractéristiques des CT-SCAN : [2]

La tomodensitométrie a deux caractéristiques principales qui font tout son intérêt par rapport aux techniques de radiographie conventionnelle:

- elle donne d'excellents résultats tomographiques (avec un « effacement » des autres plans qui est quasi parfait) .
- elle permet de voir directement les tissus mous, avec beaucoup de détails. En effet, sa résolution en densité (=en atténuation) est bien supérieure.

1.3.4. Différence entre CT-SCAN et IRM :[3]

Si on regarde une radiographie standard ou une radiographie (comme une radiographie pulmonaire), il semble qu'on regarde à travers le corps.

La tomodensitométrie et l'IRM sont semblables les unes aux autres, mais fournissent une vision du corps très différente de celle d'un rayon X.

CT et l'IRM produisent des images transversales qui semblent ouvrir le corps vers le haut, permettant au docteur de le regarder de l'intérieur.

L'IRM utilise un champ magnétique et des ondes radio pour produire des images, tandis que le CT utilise des rayons X pour produire des images.

1.3.5. Exemples de CT-SCAN :[4]

- Un CT de la tête ou du cerveau est utilisé pour évaluer les différentes structures du cerveau pour rechercher une masse, un accident vasculaire cérébral, une zone de saignement ou une anomalie des vaisseaux sanguins. Il est aussi parfois utilisé pour regarder le crâne.

- Un CT du cou contrôle les tissus mous du cou et est fréquemment utilisé pour étudier une masse dans le cou ou pour rechercher des ganglions lymphatiques ou des glandes.

- Un CT de la poitrine est fréquemment utilisé pour étudier une anomalie sur une simple radiographie thoracique, il est également souvent utilisé pour rechercher des ganglions lymphatiques élargis.

Ces études sont souvent ordonnées pour vérifier une cause de douleur et parfois pour suivre une anomalie observée sur un autre test, comme une échographie.

- Un test CT de la colonne vertébrale est le plus couramment utilisé pour détecter un disque herniaire ou un rétrécissement du canal rachidien (sténose spinale) chez les personnes souffrant de douleurs au cou, au bras, au dos et / ou aux jambes. Il est également utilisé pour détecter une fracture ou une rupture dans la colonne vertébrale.

- Un CT est utilisé pour évaluer l'embolie pulmonaire (un caillot sanguin dans les vaisseaux pulmonaires) ainsi que pour les anévrismes aortiques.

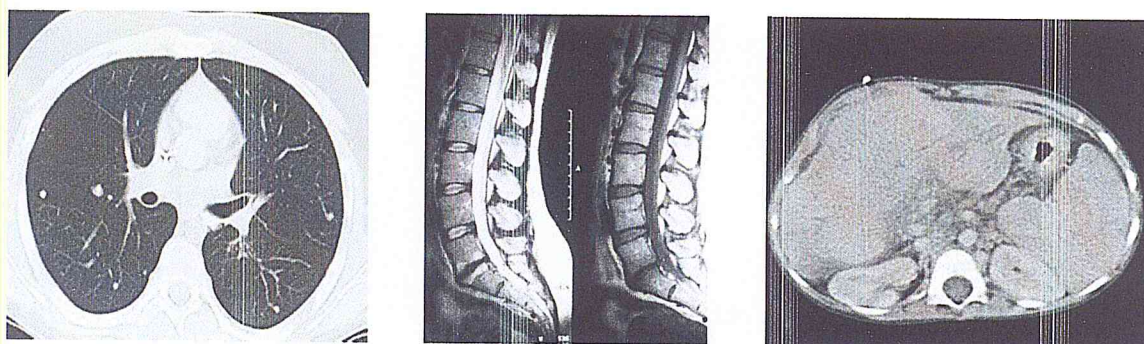


Figure 1.2 : images CT-SCAN des poumons, colonne vertébrale et le foie.

1.4. Caractéristiques de l'image : [5]

1.4.1. Pixel :

Une image est constituée d'un ensemble de points appelés pixels, le pixel (Picture élément) représente ainsi le plus petit élément constitutif d'une image numérique. La quantité d'information que véhicule chaque pixel donne des nuances entre images monochromes et images couleurs. Pour les images 3D le «pixel» est alors appelé un voxel, et représente un volume élémentaire. Des exemples d'images de ce type se rencontrent dans les images médicales. Les images tomographiques axiales sont ainsi des images construites à partir de plusieurs radiographies faites sous des angles de vue différents.

1.4.2. Poids de l'image :

C'est la taille de l'image. Etant donné que cette dernière est représentée sous forme d'une matrice dont les valeurs représentent l'intensité (pixels), le nombre de colonne (W) multiplié par le nombre de ligne (H) donne le nombre total de pixels dans l'image.

Pour une image de 640x480 en couleur :

- Nombre de pixel = $640 \times 480 = 307200$
- Poids de chaque pixel = 3 octets
- Le poids de l'image = $307200 \times 3 = 921600$ octets = 900 Ko.

1.4.3. Transparence : [6]

La transparence est une caractéristique définissant le niveau d'opacité des éléments de l'image, c'est la possibilité de voir à travers l'image des éléments graphiques situés derrière celle-ci.

1.4.4. Luminance : [6]

C'est le degré de luminosité des points de l'image. Elle est définie ainsi comme étant le quotient de l'intensité lumineuse d'une surface par l'aire apparente de cette surface.

1.4.5. Contraste :

Est une propriété intrinsèque d'une image qui désigne et quantifie la différence entre les parties claires et foncées d'une image (elle différencie les couleurs claires des couleurs foncées).

En photographie on le définit le contraste comme la différence entre la densité la plus forte et la plus faible d'une image. Le contrôle du contraste est un élément important de la pratique photographique. Le contraste final de l'image dépend à la fois du sujet, de la nature et du traitement du négatif et du positif.

1.4.6. Homogénéité :

L'homogénéité est une information locale et correspond au caractère uniforme d'une région. Une région dans une image est dite homogène si elle regroupe un ensemble de pixels qui possèdent des caractéristiques similaires ou uniformes. Ces caractéristiques peuvent être par exemple la variance du niveau de gris, la couleur...

1.4.7. Connexité :

En traitant une image, on est souvent amené à se déplacer dans celle-ci. Un déplacement doit souvent obéir à des règles de voisinage, on utilise généralement deux types de voisinage : le voisinage à 4-connexité (4 pixels voisins), et le voisinage à 8-connexité (8 pixels voisins).

Deux pixels seront considérés comme connexes (appartenant au même objet donc) s'ils satisfont deux critères:

- d'une part un critère de similarité (par exemple même niveau de gris).
- s'ils sont adjacents (voisins).

1.4.8. Région :

Une région est un ensemble de pixels connexes et homogènes. Un pixel n'appartient à une région donnée que s'il vérifie les caractéristiques de celle-ci (intensité moyenne, centre de gravité,...). Une région est toujours limitée par un contour.

1.5. Les différents formats d'images :

On peut classer les images en deux formats :

1.5.1. Format vectorielle : [7]

Dans une image vectorielle les données sont représentées par des formes géométriques simples qui sont décrites d'un point de vue mathématique.

Par exemple, un cercle est décrit par une information du type (cercle, position du centre, rayon). Ces images sont essentiellement utilisées pour réaliser des schémas ou des plans.

1.5.2 Format matricielle : [7]

Une image matricielle est formée d'un tableau de points ou pixels. Plus la densité des points sont élevée, plus le nombre d'informations est grand et plus la résolution de l'image est élevée. Corrélativement la place occupée en mémoire et la durée de traitement seront d'autant plus grandes.

Les images vues sur un écran de télévision ou une photographie sont des images matricielles. On obtient également des images matricielles à l'aide d'un appareil photo numérique, d'une caméra vidéo numérique ou d'un scanner.

1.6. Conclusion :

Nous avons présenté dans ce chapitre les notions essentielles liées à la technique tomodynamométrie et nous avons présenté également caractéristiques et les formats de l'image.

L'imagerie CT-scan est une technique d'imagerie médicale très utile pour l'observation de plusieurs parties du corps. Car c'est la méthode d'imagerie qui, dans la plupart des cas, donne les informations les plus riches

. Les images CT Scan souffrent d'un grand nombre d'artéfacts, le bruit dans les données, dû à la mesure physique, les variations d'intensité au sein d'un même tissu ou d'une même structure, qui rendent leur segmentation automatique difficile.

Le chapitre suivant, introduit l'état de l'art sur les différentes méthodes de clustering .



CHAPITRE 2
Etat de L'art sur le clustering

2.1. Introduction :

La segmentation est l'opération fondamentale de l'analyse d'image, elle consiste à partitionner l'image en un ensemble de régions connexes. Ces régions sont exploitées, pour extraire des paramètres de forme, de position et de taille afin de réaliser l'étape de reconnaissance des entités caractérisées.

Dans ce chapitre nous allons commencer par une définition de la machine learning et ensuite nous allons présenter le clustering qui va être l'objet de notre étude et il est aussi l'un des champs de recherche les plus utilisées dans la vision par ordinateur dans le domaine médicales. [08]

2.2. Machine Learning c'est quoi ? [9]

Machine learning (apprentissage automatique) c'est un champ d'étude de l'intelligence artificielle, concerne la conception, l'analyse, le développement et l'implémentation de méthodes permettant à une machine (au sens large) d'évoluer par un processus systématique, et ainsi de remplir des tâches difficiles ou problématiques par des moyennes algorithmiques plus classiques.

L'analyse peut concerner des graphes, arbres, ou courbes (par exemple, la courbe d'évolution temporelle d'une mesure ; on parle alors de données continues, par opposition aux données discrètes associées à des attributs-valeurs classiques) au même titre que de simples nombres. .

Il existe 2 types d'apprentissage :

- Apprentissage supervisé (la classification).
- Apprentissage non supervisé (clustering).

2.2.1. Différence entre apprentissage automatique supervisé et non supervisé : [10]

L'apprentissage supervisé revient à construire un modèle de corrélation entre 2 ou plusieurs variables a priori connues – on "sent" qu'il y a un lien entre telle et telle variable – alors que l'apprentissage non supervisé va prendre en compte l'ensemble des variables d'un problème et va en extraire les corrélations les plus fortes. Ce qui est bien plus puissant car

cela va au-delà des préjugés humains et met en relief des corrélations cachées qu'aucun d'entre nous n'aurait pu imaginer.

2.2.2. Quelques exemples pratiques de Machine Learning : [11]

- La voiture autonome de Google.
- Classification des emails dans Gmail.
- La traduction en temps réel de Skype.
- La reconnaissance vocale Siri d'Apple.
- Détection de fraude dans le monde de la banque.
- Reconnaissance faciale.
- Liste non exhaustive....

2.2.3. Méthodes de machine learning : [12]

Nous pouvons dénombrer 3 méthodes basiques:

- **La Classification:** est une discipline relié de près ou de loin a plusieurs domaines, pour attribuer une définition au terme « classification », il faudrait d'abord définir ses racines, ça vient du verbe "classer" qui désigne plus une action qu'un domaine, ou plutôt une série de méthodes qu'une théorie unifiée.

En mathématique, On appelle classification, la catégorisation algorithmique d'objets. Elle consiste à attribuer une classe ou catégorie à chaque objet (ou individu) à classer, en se basant sur des données statistiques. Elle fait couramment appel aux méthodes d'apprentissage et est largement utilisée en reconnaissance de formes.

D'une manière général en vertu de ces définitions, la classification se définit alors comme une méthode mathématique d'analyse de données, pour faciliter l'étude d'une population d'effectif important, généralement des bases d'observations caractérisent un domaine particulier (animaux, plantes, malades, gènes, . . . etc.), où on les regroupe en plusieurs classes.

La classification se fait par les étapes suivantes :

1. Choix des données.
2. Calcul des similarités entre les n individus à partir des données initiales.
3. Choix d'un algorithme de classification et exécution.
4. L'interprétation des résultats :

- évaluation de la qualité de la classification,
- description des classes obtenues.

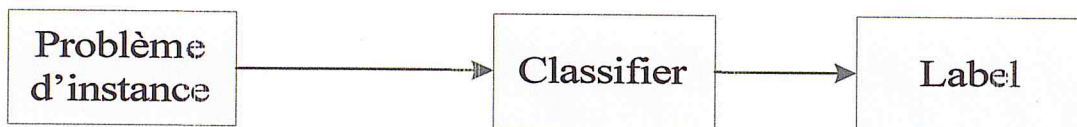


Figure 2.1. Schéma qui explique la classification

- la **Régression**: les données sont liées à d'autres données numériques par une corrélation (une droite, une courbe, une tendance) .

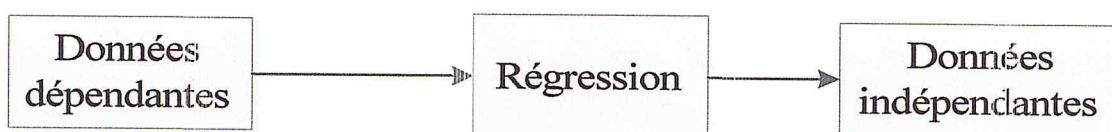


Figure 2.2. Schéma qui explique la régression

- le **Clustering**: ressemble à la classification mais avec des classes non connues.
- Cette dernière est la méthode utilisée dans notre travail, nous allons la détailler par la suite.

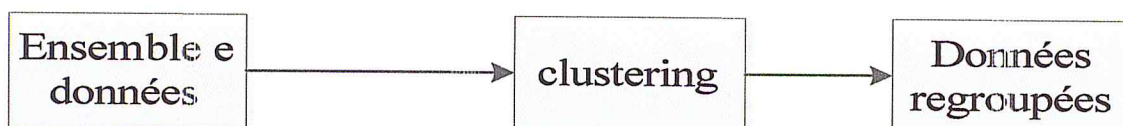


Figure 2.3. Schéma qui explique le clustering

Le graphique ci-dessous illustre les 3 méthodes sur des données similaires:

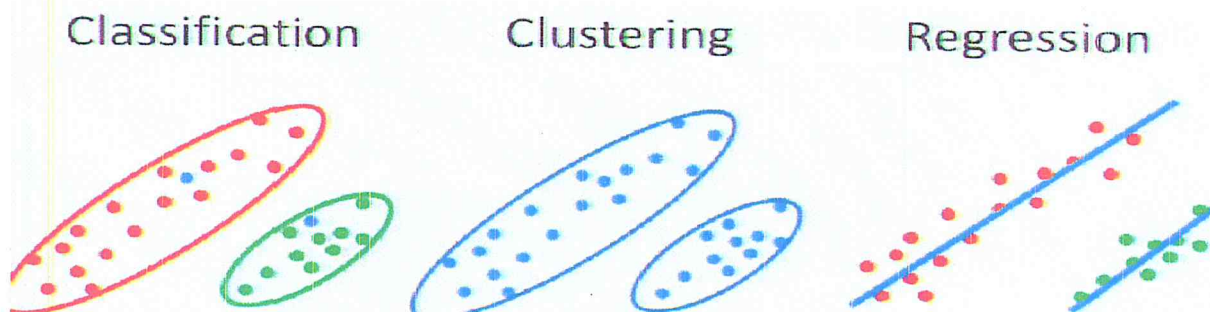


Figure 2.4. Graphique des 3 méthodes

2.2.4. Clustering : [13]

2.2.4.1. Définition

Le Clustering aussi connu sous nom. (Segmentation) est un regroupement en classes homogènes consistant à représenter un nuage de points appelé Cluster.

C'est un traitement sur un ensemble d'objets qui n'ont pas été étiquetés par un superviseur. Ce type de méthodes vise à répondre au problème de : diminution de la dimension de l'espace d'entrée, ou pour le groupement des objets en plusieurs catégories (clusters) non définies à l'avance.

Un «Cluster» est donc une collection d'objets qui sont «similaires» entre eux et qui sont «dissemblables » par rapport aux objets appartenant à d'autres groupes. On peut voir cette définition clairement graphiquement dans l'exemple suivant :

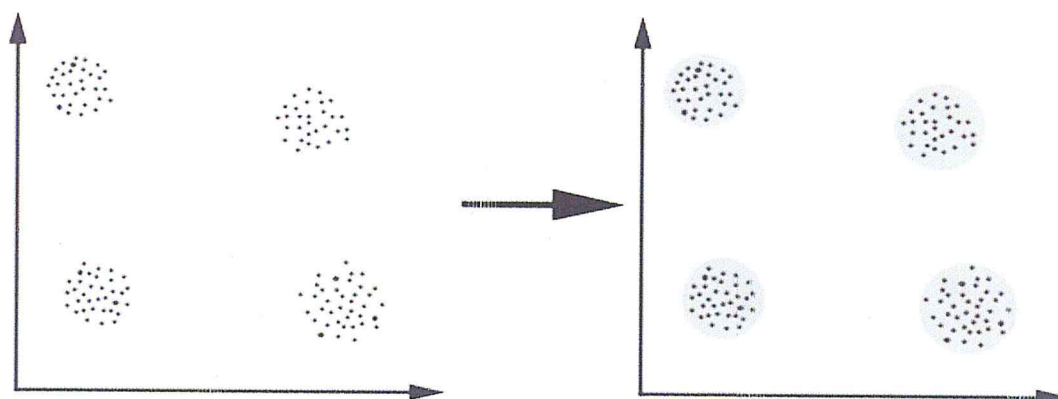


Figure 2.5. Illustration de regroupement en clusters

Dans ce cas, il est très facile pour une personne d'identifier 4 Clusters dans lesquels les données (nuage des points) peuvent être divisées, le critère de similarité est la distance : deux ou plusieurs objets appartiennent au même cluster s'ils sont «proches», bien sûr cela dépend d'une distance donnée (dans ce cas la distance géométrique).

2.2.4.2. Principe général :

Contrairement à la classification (méthodes supervisées), on ne possède pas des connaissances a priori sur les classes prédéfinies des éléments. Donc La division des objets dans les différents groupes (clusters) se procède en se basant sur le calcul de similarité entre les éléments.

Alors que l'objectif des méthodes du Clustering est de grouper des éléments proches dans un même groupe de manière à ce que deux données d'un même groupe soient le plus similaires possible et que deux éléments de deux groupes différents soient le plus dissemblables possible.

Mathématiquement, on a un ensemble X de N données décrites chacune par leurs P attributs.

Donc Le Clustering consiste à créer une partition ou une décomposition de cet ensemble en sous parties (clusters) telle que :

- * Les données appartenant au même groupe se ressemblent,
- * Les données appartenant à deux groupes différents soient peu ressemblantes.

2.2.4.3. Les exigences de Clustering : [13]

Les principales exigences qu'un algorithme de clustering doit répondre sont les suivantes :

- Evolutivité des clusters.
- Traiter les différents types d'attributs.
- Découvrir les clusters de forme arbitraire.
- Exigences minimales pour la connaissance du domaine afin de déterminer les paramètres d'entrée.
- Capacité de composer avec le bruit et les valeurs manquantes traiter les dimensionnalités élevées. L'intelligibilité et la convivialité.

2.2.4.4. Exemples de clustering :

On l'utilise souvent en traitement d'images pour fixer les divers objets qu'elles contiennent (segmentation) : routes, villes, rues, des organes humaines (pour les images médicales).

2.2.4.5. Les types de Clustering : [13]

Il existe deux grands types du clustering :

1. Le clustering hiérarchique : d'agglomération («bottom-up»), on décompose l'ensemble d'individus en une arborescence de groupes.
2. Le clustering non-hiérarchique : de division («top-down»), on décompose l'ensemble d'individus en K groupes,

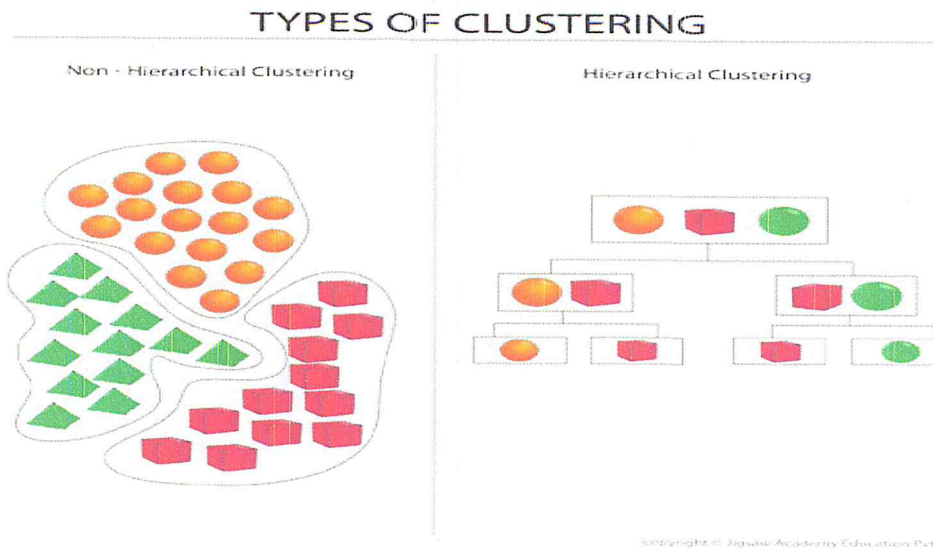


Figure 2.6. Les deux types de clustering non-hiérarchique/hiérarchique.

Cependant dans certains ouvrages on classifie les types des Algorithmes de clustering en 4 groupes à cause des méthodes qui ne respectent plus les normes du premier classement comme le cas de la règle « Chaque objet doit appartenir à un seul groupe. »

Les 4 types sont :

1. Clustering exclusif (k-means).
2. Overlapping Clustering (Fuzzy clustering).
3. Clustering Hiérarchique (CHA).
4. Clustering non hiérarchique(EM).

Dans le premier cas, les données sont regroupées d'une manière exclusive, de sorte que si une donnée certaine appartient à un amas définie alors il ne pourrait pas être inclus dans un autre cluster. Un simple exemple de cela est montré dans la figure ci-dessous, où la séparation des points est définie par une ligne droite sur un plan bidimensionnel.

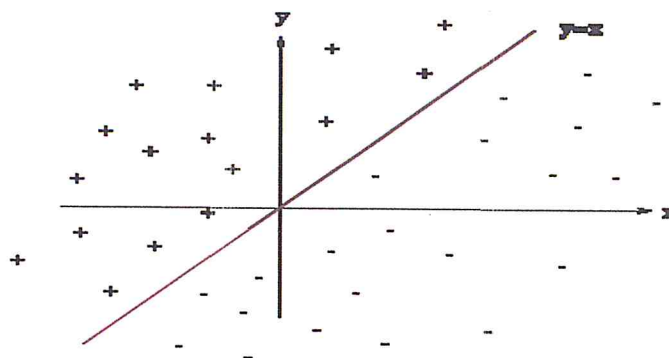


Figure 2.7 . Exemple d'un problème de discrimination à deux classes, avec une séparatrice linéaire : la droite d'équation $y=x$. Le problème est linéairement séparable.

Au contraire le second type, le regroupement overlapping, utilise des ensembles flous aux données de cluster, de sorte que chaque point peut appartenir à deux ou plusieurs groupes avec différents degrés d'appartenance. Dans ce cas, les données seront associées à une valeur d'une composition appropriée.

Un algorithme de clustering hiérarchique est fondé sur l'union entre les deux plus proches clusters c'est-à-dire : consiste à trouver des clusters successifs utilisant des clusters précédemment établis, la première condition est de mettre au début chaque objet dans un cluster distinct et les fusionner en clusters successivement plus grand. Après quelques itérations on atteint le final Cluster voulu qui regroupe tous les sous-clusters (sous-partitions).

2.2.4.6. Exemples d'algorithmes :

Dans ce qui suit, nous analysons le cadrage théorique de chaque méthode, concernant la 4^{ème} méthode (EM) c'est la méthode cible de notre travail.

2.2.4.6.1. K-means

L'algorithme k-means mis au point par McQueen en 1967 [10] un des plus simples algorithmes d'apprentissage non supervisé, appelée algorithme des centres mobiles, il attribue chaque point dans un cluster dont le centre (centroïde) est le plus proche, le centre est la moyenne de tous les points dans le cluster, ses coordonnées sont la moyenne arithmétique pour chaque dimension séparément de tous les points dans le cluster c'est-à-dire chaque cluster est représentée par son centre de gravité.

Algorithme : K-means [14]

- Initialiser k noyaux (centres des classes) d'une manière aléatoire ;

- A chaque itération n balayer l'image :

- Affecté x à la classe k si :

$$\|x_s - k(i)\| = \min_{i=1}^K \|x_s - k(i)\|$$

- Recalculer le centre de chaque classe par :

$$m(k(i)) = \frac{1}{\Omega} \sum_{x \in \Omega} x$$

- Si les centres des groupements à l'itération p sont les mêmes que ceux obtenus à l'itération précédente, alors les groupements constitués à l'itération p et $(p-1)$ sont identiques et l'algorithme a convergé, sinon on retourne à l'étape 2.
-

Principe :

L'idée principale est de définir les k centroïdes arbitraires c_1, c_2, \dots, c_k (k le nombre de clusters fixé a priori, chaque c_i représente le centre d'une classe), Ces centroïdes doivent être placés dans des emplacements différents. Donc, le meilleur choix est de les placer le plus possible éloignés les uns des autres, la prochaine étape est de prendre chaque point appartenant à l'ensemble de données et l'associer au plus proche centroïde, c'est à dire Chaque classe S_i sera représentée par un ensemble d'individus les plus proches de son c_i , Les nuées dynamiques sont une généralisation de ce principe, où chaque cluster est représenté par un noyau mais plus complexe qu'une moyenne.

Lorsqu'aucun point n'est en attente, la première étape est terminée et un groupage précoce est fait. À ce point nous avons besoin de recalculer les k nouveaux centroïdes m_i des groupes issus de l'étape précédente qui vont remplacer les c_i (m_j est le centre de gravité de la classe S_j , calculé en utilisant les nouvelles classes obtenues). Après, on réitère Le processus jusqu'à atteindre un état de stabilité où aucune amélioration n'est possible, nous pouvons constater que les k centroïdes changent leur localisation par étape jusqu'à plus de changements sont effectués. En d'autres termes les centroïdes ne bougent plus.

2.2.4.6.2. Méthodes hiérarchiques [15]:

Le processus basique des méthodes hiérarchiques a été donné par Lance & Williams, 1967 [15]. Ce type de clustering consiste à effectuer une suite de regroupements en Clusters de moins en moins fines en agrégeant à chaque étape les objets (simple élément) ou les groupes d'objets (un Cluster-partition-) les plus proches. Ce qui nous donne une arborescence de clusters. Cette approche utilise la mesure de similarité pour refléter l'homogénéité ou l'hétérogénéité des classes.

Principe :

Son principe est simple, initialement chaque individu forme une classe, soit n classes, donc on cherche à réduire ce nombre de classe n itérativement de sorte que dans chaque étape on fusionne deux classes ensemble (Les deux classes choisies pour être fusionnées sont celles qui sont les plus "proches" en fonction de leur dissimilarité) ou ajouter un nouveau élément à une classe (un élément appartient à une classe s'il est plus proche de cette classe que de toutes les autres) La valeur de dissimilarité est appelée indice d'agrégation. Qui commence dans la première itération faible, et croît d'itération en itération.

Parmi les algorithmes plus connus de ce type : La classification ascendante hiérarchique (CHA) où le mot ascendante est utilisé pour désigner qu'elle part d'une situation dont tous les individus représentent des clusters à part entière, puis on cherche les rassembler en classes de plus en plus grandes. Ainsi Le qualificatif "hiérarchique" désigne le fait qu'elle produit une hiérarchie, (une amélioration a été proposée en 2002 par P. Bertrand, appelée Classification Ascendante 2-3 Hiérarchique).

Algorithme de CHA :

- Initialisation :
 - Chaque individu est placé dans son propre cluster.
 - Calcul de la matrice de ressemblance M entre chaque couple de clusters (ici les points).
- Répéter :
 - Sélection dans M des deux clusters les plus proches C_i et C_j .
 - Fusion de C_i et C_j par un cluster C_G plus général.

- Mise à jour de M en calculant la ressemblance entre CG et les clusters existants

Jusqu'à fusionner les 2 derniers clusters.

Dans la figure suivante, on représente une illustration du principe de CHA et la hiérarchie finale obtenue où Les liens hiérarchiques apparaissent clairement.

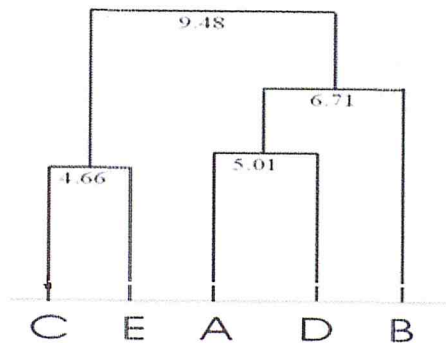


Figure 2.8. le dendrogramme de la hiérarchie H de la suite de partitions d'un ensemble {a, b, c, d, e}

2.2.4.6.3. Algorithme Fuzzy c-means de clustering:[16]

Principe :

Fuzzy C-means (FCM) est une méthode de clustering qui permet à un objet de données d'appartenir à deux ou plusieurs clusters. Cette méthode dérivée de l'algorithme c-means, identique à l'algorithme k-means décrit précédemment, elle a été développée par Dunn en 1973 et améliorée par Bezdek en 1981, est fréquemment utilisée dans la reconnaissance des formes. Il est basé sur la minimisation de la fonction objective suivante :

$$J_m = \sum_{i=1}^N \sum_{j=1}^C u_{ij}^m \|x_i - c_j\|^2 \quad 1 \leq m < \infty$$

où m est un nombre réel (> 1), U_{ij} est le degré d'appartenance de x_i dans le j ème Cluster, x_i est le i ème élément des données mesurées, c_j est le centre d'un cluster et $\| \cdot \|$ est toute norme exprimant la similarité entre les données mesurées et le centre. Ce Partitionnement logique flou (fuzzy) est réalisé grâce à une optimisation itérative de la fonction objectif indiqué ci-dessus, avec la mise à jour de l'appartenance u_{ij} et les centres des clusters c_j .

Algorithme Fuzzy c-means :

- Initialiser la position des centres :
 - $V_i = [v_i^1, v_i^2, \dots, v_i^j, \dots, v_i^n], 1 \leq i \leq c.$
- $l=0$, initialiser la matrice : $U^{(l)}$
- Calculer la position des centres :

$$\circ V_i = \frac{\sum_{k=1}^n (u_{ik})^m x_k}{\sum_{k=1}^n (u_{ik})^m}, 1 \leq i \leq c.$$

- Calculer la matrice : $U^{(l+1)}$.
- Si $\|U^{(l+1)} - U^{(l)}\| < \epsilon$
 - Arrêter l'algorithme
- Sinon
 - $l = l + 1$ et retourner en 2.

2.2.4.6.4. Algorithme d'Expectation-Maximisation :

En français « L'algorithme d'espérance-maximisation », souvent abrégé par EM, Un des premiers articles sur EM a été écrit en 1958 [17] mais la référence pratique qui a formalisé EM et a fourni une preuve de convergence est le document de Dempster, Laird et Rubin en 1977[18], tandis que le livre de [19] est une autre référence populaire et très utile. Son objectif est de trouver le maximum de vraisemblance des paramètres de modèles probabilistes.

On compte une multitude de domaines d'applications de cet Algorithme, à titre d'exemple : il est usuellement utilisé dans la vision artificielle ou encore dans le traitement d'images, plus spécifique en ce qui concerne la segmentation (image médicale, satellitaire...etc.), ou tout simplement en clustering pour regrouper les données homogènes dans un groupe.

Un livre récent consacré entièrement à EM et ses extensions, en plus des applications est : [20] donc l'estimation des paramètres du modèle passe par la maximisation de $Q(\theta/\theta_q)$ pour cela cet algorithme comprend deux étapes essentielles :

1. E-steps (E) : une étape d'évaluation de l'espérance, c'est dans cette étape qu'on calcule l'espérance de la vraisemblance en tenant compte des dernières variables observées,
2. M-steps (M) : une étape de maximisation de vraisemblance qu'on a trouvé à l'étape (E), en tentant d'estimer le maximum de vraisemblance des paramètres.

Et c'est ainsi qu'on itère l'algorithme en utilisant les paramètres trouvés à l'étape (M) pour évaluer à nouveau l'espérance.

On va maintenant résumer l'algorithme par le tableau suivant :

Algorithme 1 : Estimation EM des paramètres

- Initiation θ_0
- Améliorer itérativement l'estimation de θ en alternant entre les deux étapes suivantes :

- Etape E : calculer $Q(\theta/\theta_q) = E_{\theta_q}(\log(p(X, y; \theta))|y)$;

Dans le cas où $Q(\theta/\theta_q)$ n'est pas calculable analytiquement, ou le but d'introduire une perturbation stochastique (SEM), générer un échantillon iid : $x_q(1), \dots, x_q(\tau)$ à partir de $P(x|y; \theta_q)$ et en déduit $Q(\theta/\theta_q)$ de la façon suivante :

$$Q(\theta/\theta_q) = \frac{1}{\tau} \sum_{t=1}^{\tau} P(x_q(t), y|\theta_q)$$

- Etape M : ré-estimer θ en maximisant la fonction Q :

$$\theta_{q+1} = \arg \max Q(\theta/\theta_q)$$

Dans le cas où la maximisation est difficile (GEM), choisir θ_{q+1} tel que :

$$Q(\theta_{q+1}/\theta_q) \geq Q(\theta_q/\theta_q)$$

- Arrêtez $\theta_{q+1} - \theta_q \leq \varepsilon$
-

La Convergence :

L'objectif de convergence est de trouver (mettre à jour) une « meilleure » valeur qui augmente la vraisemblance d'une itération à l'autre, donc il suffit de trouver une fonction comme :

$$\Delta(\theta, \theta_m) := \log(p(X | \theta)) - \log(p(X | \theta_m))$$

On souhaite bien sûr que cette différence soit la plus grande possible, en fait dans [21] on a trouvé qu'une telle fonction est possible appelée Q-fonction, et ce qu'il prouve que l'algorithme EM admet « la croissance de la vraisemblance d'une itération à l'autre ».

Il faut noter que, dans certains cas, l'algorithme peut ne pas bien converger ou converger que vers un maximum local de la vraisemblance...., cela revient aux conditions initiales choisies θ_0 arbitrairement, cependant pour certaines mauvaises valeurs, l'algorithme peut rester gelé en un point selle, alors qu'il convergera vers le maximum global pour d'autres valeurs initiales plus pertinentes. Donc il est recommandé de relancer plusieurs fois l'algorithme avec différentes initialisations pour appréhender ce problème de convergence. En fait, il ya aucun théorème général de convergence pour l'algorithme EM [22], la convergence de la séquence $\{\theta(m)$ à l'itération $m\}$ dépend entièrement aux choix des caractéristiques de $L(\theta)$, du modèle, ainsi que des points initiaux de début $\theta^{(0)}$.

Il y'a une multitude d'études qui ont été fait pour appréhender ce problème chacune traite un domaine particulier, par exemple dans [23], on trouve tous les aspects concernant la convergence dans un modèle de mélange gaussien, ou encore dans [24], on trouvera des discussions sur les méthodes pour le calcul de convergence des différents algorithmes EM.

Principe :

Voyons maintenant comment fonctionne l'algorithme EM pour un mélange de gaussiennes : Soit $X (X_1, \dots, X_n)$ un échantillon d'observations issues d'un mélange de gaussiennes et soit $Z (Z_1, \dots, Z_n)$ la donnée cachée où Z_i détermine la distribution dont est issue x_i (nous utilisons une représentation marginale de la vraisemblance selon les « données cachées Z »), par exemple si on a un mélange de deux gaussiennes bidimensionnelles alors :

$$L(X_i | \{Z_i = 1\}) = N_1(\mu_1; \Sigma_1) ; L(X_i | \{Z_i = 2\}) = N_2(\mu_2; \Sigma_2).$$

$$\text{Avec } P\{Z_i = 1\} = \lambda_1 \text{ et } P\{Z_i = 2\} = \lambda_2 = 1 - \lambda_1.$$

Alors que à ce stade, on ne connaît que : l'échantillon d'observations (exemple : le nuage de points X), on cherche donc, à estimer les paramètres inconnus à savoir :

$$\Phi = (\prod_1, \dots, \prod_g, \theta_1, \dots, \theta_g)$$

La vraisemblance des données complètes est :

$$L(X, Z|\Phi) = \prod_{i=1}^n \sum_{k=1}^g 1\{Z_i=j\} \prod_k f_k(x_i, \theta_i)$$

Où:

f_k : est une densité gaussienne bidimensionnelle de paramètres μ_j et Σ_j ce qui se traduit en log-vraisemblance des données complétés par :

$$\begin{aligned} \log(L(X, Z|\Phi)) = & \sum_{i=1}^n \left[\sum_{k=1}^g 1\{Z_i=j\} (\log(\prod_k)) - \log(2 \prod_k) \right. \\ & \left. - \frac{1}{2} \log(\det(\sum_k)) - \frac{1}{2} (x - \mu_k)^T \sum_k^{-1} (x - \mu_k) \right] \end{aligned}$$

Maintenant à chaque itération, l'étape E nécessite de définir la distribution a posteriori de Z_j connaissant X_i et Φ . On définit :

$$p(Z_{i=j} | X_k = x_i, \Phi^{EM}) = \frac{\prod_k^{EM} f_k(x_i, \theta_k^{EM})}{\sum_{j=1}^g \prod_k^{EM} f_j(x_i, \theta_j^{EM})}$$

Cette formule donne donc la possibilité de calculer la probabilité conditionnelle pour que l'objet x_i appartienne à la gaussienne qui correspond au groupement (distribution) numéro j donnant les paramètres FEM ainsi estimés. Autrement dit c'est la probabilité a posteriori pour que le point X_i soit issu de la distribution $f_k = N(\mu_k, \Sigma_k)$ connaissant Φ^{EM} . La maximisation de log-vraisemblance en Φ , ne présente aucune difficulté majeure, et conduit aux estimateurs suivants :

$$\mu_k^{new} = \frac{1}{N} \sum_{i=1}^N p(Z_k | X_i, \Phi^{old})$$

$$\Sigma_k^{new} = \frac{\sum_{i=1}^N X_i p(Z_k | X_i, \Phi^{old})}{\sum_{i=1}^N p(Z_k | X_i, \Phi^{old})}$$

$$\sum_k^{new} = \frac{\sum_{i=1}^N p(Z_k | X_1, \Phi^{old}) (X_1 - \mu_k^{new})(X_1 - \mu_k^{new})^T}{\sum_{i=1}^N p(Z_k | X_1, \Phi^{old})}$$

En Grosso modo, on peut résumer ces étapes par : on démarre l'algorithme avec une ignorance absolue des données cachées Z et en initialisant θ (pour chaque composante «gaussienne») à une valeur θ_0 d'une manière totalement arbitraire, potentiellement très loin de la réalité.

L'algorithme se sert donc de θ_0 pour estimer Z , puis se sert de meilleure estimation de Z (lors de E-step) pour réestimer les paramètres en une valeur θ_1 plus pertinente.

À l'itération suivante, on évalue donc l'estimation des données cachées Z puisque cette nouvelle estimation se base cette fois sur θ_1 . Et cette meilleure estimation sur Z conduit à son tour à une meilleure précision sur θ_2 , et ainsi de suite jusqu'à atteindre la convergence.

Et au final en plus que l'algorithme nous fournit une meilleur estimation de θ , il estime aussi les variables cachées (latentes) Z qui montre bien une autre utilisation de l'algorithme EM, à savoir : la complétion des données manquantes.

Usage de l'algorithme :

On utilise souvent l'algorithme EM pour la classification de données, l'apprentissage automatique, ou la vision artificielle. On peut également citer son utilisation en imagerie médicale dans le cadre de la reconstruction tomographique.

2.2.4.7. Avantage et Inconvénient des méthodes de clustering:

Chaque méthode de clustering a des avantages et des inconvénients, qui sont résumé dans le tableau suivant.

L'algorithme	avantages	incovénients
K-means	<ul style="list-style-type: none"> - l'algorithme de k-means est très populaire du fait qu'il est très facile à comprendre et à mettre en œuvre. -applicable à des données de grandes tailles, et aussi à tout type de données(même textuelles),en choisissant une bonne notion de distance. 	<ul style="list-style-type: none"> - le nombre de classe doit être fixé au départ. - le résultat dépend de tirage initial des centres de classes.
CHA	<ul style="list-style-type: none"> -ne nécessite pas de connaître le nombre de clusters a priori -il n'y a pas de fonction d'initialisation 	<ul style="list-style-type: none"> - on ne peut pas annuler une action effectuer. - complexité $O(n^2)$
Fuzzy c-means de clustering	<ul style="list-style-type: none"> - donne des résultats mieux que k-means. 	<ul style="list-style-type: none"> - Spécification Apriori du nombre de cluster.
Expectation –Maximisation (EM)	<ul style="list-style-type: none"> - donner des résultats extrêmement utiles pour des données dans le monde réel. 	<ul style="list-style-type: none"> - algorithme est de nature très complexe

Tableau 2.1. Tableau d'avantages et d'inconvénients des méthodes de clustering.

2.3. Mesure de similarité : [25]

Pour comparer l'homogénéité ou la ressemblance, la similarité entre deux objets (points, images, classes...etc.), il faut pouvoir mesurer la similarité (ou la dissimilarité) entre eux. Nous allons décrire maintenant des mesures de similarité pour prouver la similarité entre les objets.

La similarité est une partie importante de la définition d'une méthode de clustering, elle consiste en effet à définir et formaliser une mesure de similarité adaptée aux caractéristiques des données. Si les composantes des vecteurs de données d'instance sont toutes dans les mêmes unités physiques alors il est possible que la distance euclidienne est suffisante pour réussir à grouper les données similaires. Cependant, même dans ce cas, la distance euclidienne peut parfois être trompeuse.

Il est à noter qu'il ya deux concepts pour exprimer la notion de proximité entre les objets à classifier :

1. Mesure de dissimilarité DM : plus la mesure est faible plus les points sont similaires (distance).
2. Mesure de similarité SM : plus la mesure est grande, plus les points sont similaires.
3. On parle souvent de « distances » en désignant une mesure de similarité, lorsque ces mesures ont les propriétés de non-négativité, réflexivité, symétrie (la distance entre l'objet A à B est la même que la distance de B à A) et qui respectent l'inégalité triangulaire. Il existe un grand nombre de mesures de similarité, dans ce qui suit, nous présentons quelques-unes des fonctions entre deux objets $d(x_1 ; x_2)$.

Fonctions de similarité :[25]

- **La distance euclidienne** : (aussi appelée la distance à vol d'oiseau) Un rapport de clusters analyses en psychologie de la santé a conclu que la mesure de la distance la plus courante dans les études publiées dans ce domaine de recherche est la distance euclidienne ou la distance au carré euclidienne.

$$d^2(x_1, x_2) = \sum_i (x_{1i} - x_{2i})^2 = (x_1 - x_2)(x_1 - x_2)'$$

- **La distance de Manhattan** : (appelée aussi taxi-distance)

$$d^2(x_1, x_2) = \sum_i |x_{1i} - x_{2i}|$$

- **La distance de Mahalanobis** : corrige les données pour les différentes échelles et des corrélations dans les variables, L'angle entre deux vecteurs peuvent être utilisés comme mesure de distance quand le regroupement des données de haute dimension. Voir l'espace produit scalaire.

$$d^2(x_1, x_2) = (x_1 - x_2)C^{-1}(x_1 - x_2)'$$

(C=covariance)

- **Distances entre distributions** : La similarité entre distributions consiste à déterminer si deux distributions peuvent être issues de la même distribution de probabilités.

Le test statistique du χ^2 permet de décider si deux vecteurs \vec{x} et \vec{y} sont engendrés par la même distribution. La version symétrique du test est :

$$\chi^2(\vec{x}, \vec{y}) = \sum_{i=1}^n \frac{(x_i - y_i)^2}{x_i + y_i}$$

Cependant que pour les données de grandes dimensions, il y a une distance spécifique très utilisée :

La métrique Minkowski : Pour les données dimensionnelles, c'est la mesure populaire

$$d_p(x_i, x_j) = \left(\sum_{k=1}^d |x_{ik} - x_{jk}|^p \right)^{1/p}$$

Où d est la dimensionnalité des données. La distance euclidienne est un cas particulier où $p = 2$, alors que Manhattan $p = 1$. Néanmoins, il n'existe pas de directives générales théoriques pour la sélection d'une mesure à une application donnée.

2.4. Conclusion :

Dans ce chapitre nous avons tenté de définir les différents algorithmes de clustering les plus connus dans la littérature, ces méthodes permettent une conception des algorithmes de segmentation non supervisée. Nous avons choisi l'algorithme EM pour essayer de réaliser un clustering des images CT Scan.

Dans le chapitre suivant, nous entamerons la partie pratique de ce mémoire en commençant par l'implémentation de l'application de la méthode EM.



CHAPITRE 3

Conception



3.1. Introduction :

Le clustering s'appuie sur des algorithmes plus ou moins complexes, tels que les algorithmes de k-moyennes, ou les algorithmes de maximisation de l'espérance.

Parmi ces algorithmes nous allons choisir celle d'expectation-maximisation (EM) dont l'algorithme est un algorithme itératif qui permet de trouver les paramètres de vraisemblance maximum d'un modèle probabiliste lorsque ce dernier dépend de variables latentes non observables.

Ce chapitre est divisé en deux sections, la première est consacrée à la présentation de l'architecture de notre système ,l'application d'EM sur les images SC-Scan ainsi que la discussions sur le choix de cette méthode . Dans la deuxième section, nous allons présenter l'empilement et la reconstruction 3D des images CT-Scan.

3.2. Architecture globale du système :

Dans le cadre de notre travail, nous nous sommes intéressés au clustering des images CT-scan :

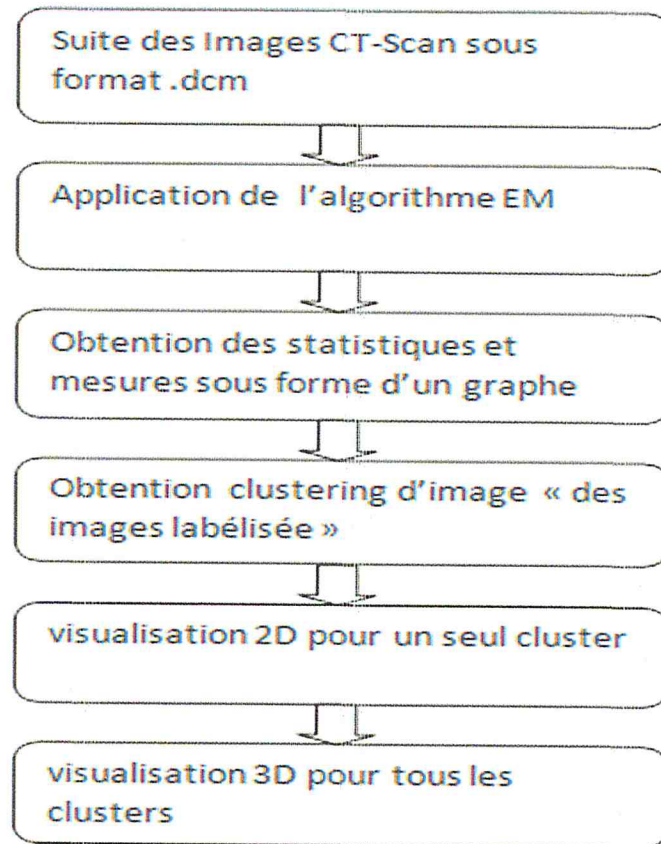


Figure 3.1. Système globale

- En entrée on a les images CT-Scan sous format DCM , pour ses images on doit définir le nombre de classe k .
- Après, on applique l'algorithme EM sur chaque image et on obtient comme résultats : des statistiques et mesures sous forme d'un graphe pour chaque image, et une image labélisée « clustering d'image».
- Sur chaque image labélisée on peut extraire un seul cluster et on fait sa visualisation 2D.
- Pour la reconstruction 3D des images CT-Scan, on fait l'empilement sur l'ensemble des images labélisées et on procède à sa visualisation 3D pour tous les clusters.

3.3. Conversion d'image en matrice :

L'image CT-Scan est une image matricielle d'extension (.dcm) (image DICOM), après la lecture elle devient une matrice de points (un tableau à deux dimensions). Ces points sont appelés des pixels (le plus petit élément de l'image), où chaque pixel admet une couleur.

Après la lecture des images nous avons besoin de les convertir en matrice pour cela nous avons suivi les étapes suivantes :

1. lire l'image.
2. parcourir l'image (lignes et colonnes).
3. Prendre la valeur de chaque pixel et la mettre dans la matrice.

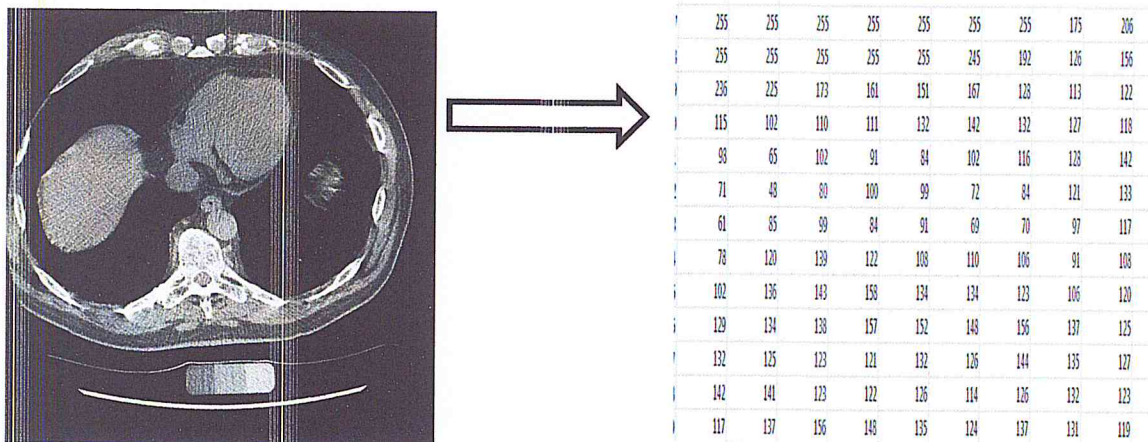


Figure 3.2. Conversion d'une image de la base de données 1 en matrice.

L'image de la figure 3.3 est composé d'un ensemble de pixels après la conversion ,on obtient une matrice ou dans chaque case contient la valeur de pixel de l'image.

3.4. Distribution : [26]

Une distribution est un objet qui généralise la notion de fonction et de mesure, en probabilité, c'est une loi (loi normale, loi de poisson.) Qui décrit les probabilités de chaque valeur d'une variable aléatoire discrète, dans notre travail nous nous intéressons à la distribution gaussienne.

En effet, en probabilité on dit qu'une variable aléatoire réelle X est une variable gaussienne, si elle suit une loi normale gaussienne d'une moyenne « μ » et de l'écart-type « σ » strictement positif, si seulement cette variable aléatoire réelle X admet comme densité de probabilité la fonction $p(x)$ définie comme suit : pour tout nombre réel x :

$$p(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

On la note par :

$$X \sim N(\mu, \sigma^2)$$

Qui veut dire la variable : aléatoire X suit la loi normale des paramètres « μ » et « σ^2 ».

Cette distribution a l'avantage qu'on peut connaître grâce aux paramètres (l'écart-type, la moyenne) la dispersion d'un ensemble de données, en fait, le savoir de la moyenne et l'écart-type permet de déterminer l'intervalle dans lequel on trouve 95 % de la population. D'ailleurs on sait qu'on trouve 95 % de la population dans l'intervalle $[\mu - 2\sigma, \mu + 2\sigma]$ et on trouve 68 % de la population dans l'intervalle $[\mu - \sigma, \mu + \sigma]$.

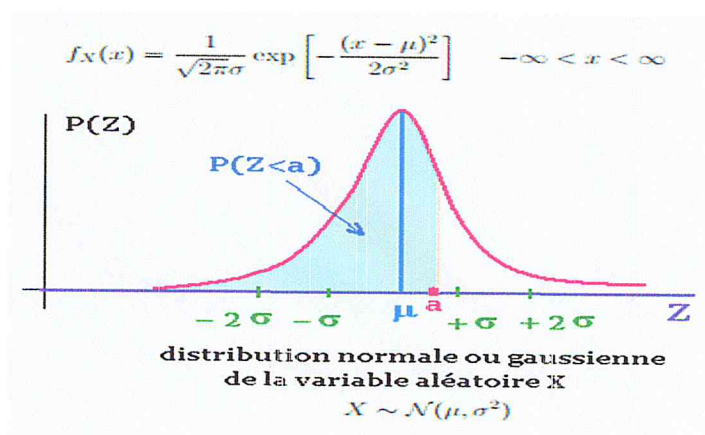


Figure 3.3. une distribution gaussienne

Dans notre travail, pour dessiner cette distribution, nous avons créé ce programme :

Algorithme Distribution

- Pour i allant de 0 à k :
 - calculer la distance $d : d = x - m$;

- calculer l'amplitude amp : $amp = \frac{p}{\sqrt{2\pi v}}$;
- calculer la distribution y : $y = amp * e^{(-\frac{1}{2} * \frac{d^2}{v})}$;

end.

Nous avons fait une boucle allant de 0 à k (k le nombre de classe).Après dans cette boucle on calcule la distance d entre x(la valeur de chaque pixel) et la moyen m. En suite on calcule la valeur de l'amplitude amp. A la fin on utilise les deux valeurs précédentes(d,amp) pour calculer la distribution y. Dans la figure suivante on représente la distribution d'une image avec k =3.

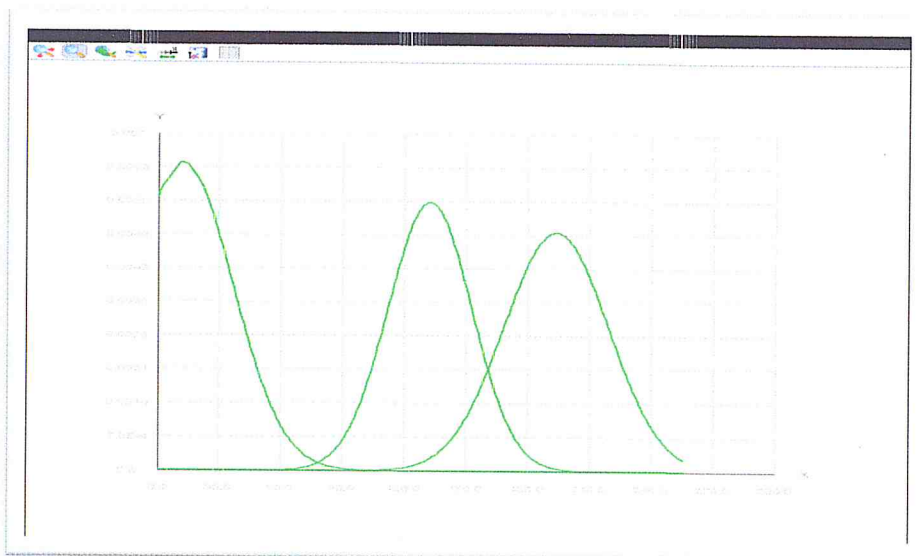


Figure3.4. la distribution d'une image avec k=3.

3.5. Histogramme :

En imagerie numérique, l'histogramme représente la distribution des intensités (ou des couleurs) de l'image. C'est un outil fondamental du traitement d'images, avec de très nombreuses applications.

Dans notre algorithme, l'histogramme représente une classe importante pour modéliser le vecteur de l'image, en calculant des probabilités des différentes intensités (niveaux de gris).

Pour faire l'histogramme de chaque image nous avons appliqué cette algorithme :

 Algorithme Histogramme

- Calculer la convolution : pour filtrer l'histogramme ;
- Calculer la probabilité d'occurrence d'un pixel du niveau de gris p :

$$p = \frac{h}{n} \quad n : \text{le nombre total de pixels.}$$

end

Nous avons calculer la convolution , après nous avons calculer la probabilité d'occurrence d'un pixel du niveau de gris p ou h représente le nombre d'occurrence d'un pixel. Dans la figure suivante on représente l'histogramme d'une image après l'application de notre programme.

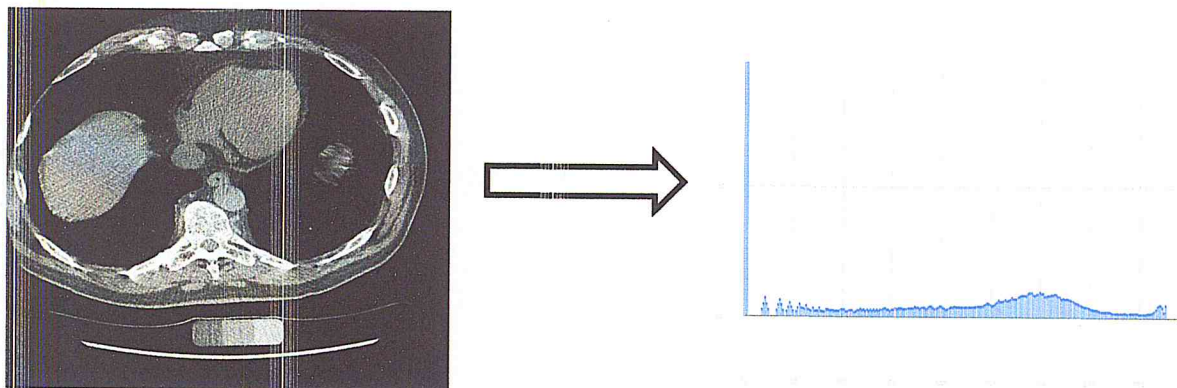


Figure 3.5. histogramme de l'image 48 BD1

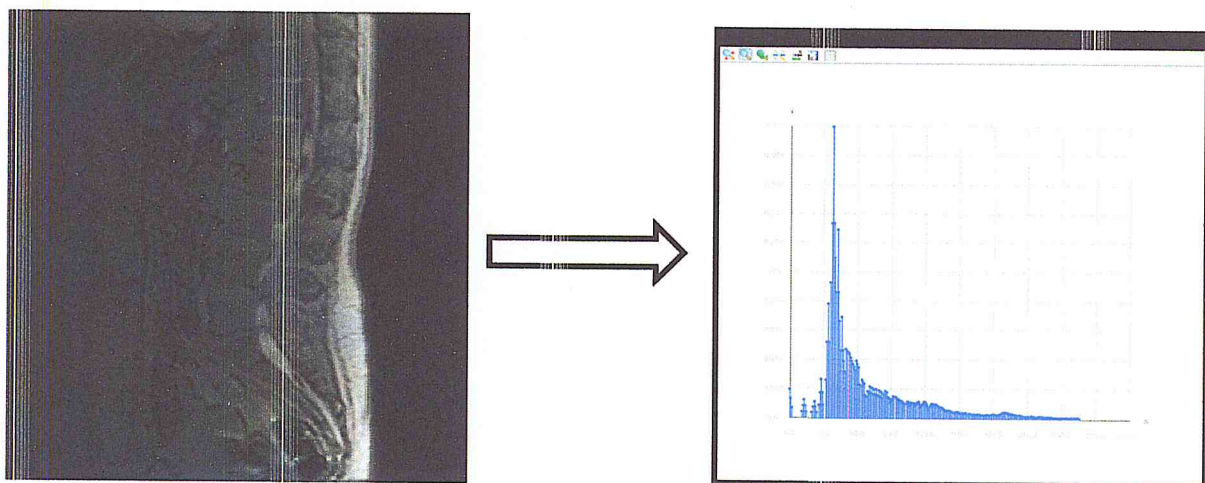


Figure 3.6. histogramme de l'image 8 BD2

Pour réaliser la représentation d'histogramme, et tracer les courbes nous avons utilisé la librairie :

- `org.math.plot.Plot2DPanel` : est une bibliothèque java légère conçue pour permettre de compléter facilement les données 2D et 3D dans un panneau java.

3.6. La methode EM :

3.6.1.choix de la méthode EM :

Nous avons choisis cette méthode, car elle est régulièrement utilisée pour résoudre un large éventail de problèmes d'estimation, en effet, EM touche un grand nombre de problèmes d'estimation dans divers domaines : traitement du signal, traitement de l'image, notamment en imagerie médicale.

Cette méthode est adaptée pour traiter des grandes masses des données de grandes dimensions comme les images dicom, pourtant elle marque quelques faiblesses.

3.6.2. Application de la méthode EM :

Comme nous avons pu le voir dans le chapitre précédent la méthode EM (expectation-maximisation) est une technique de clustering qui contient deux étapes (E step, M step), dans ce chapitre nous allons voir l'application de cette méthode sur les images CT-scan pour cela nous avons présenté un schéma qui explique le traitement effectué :

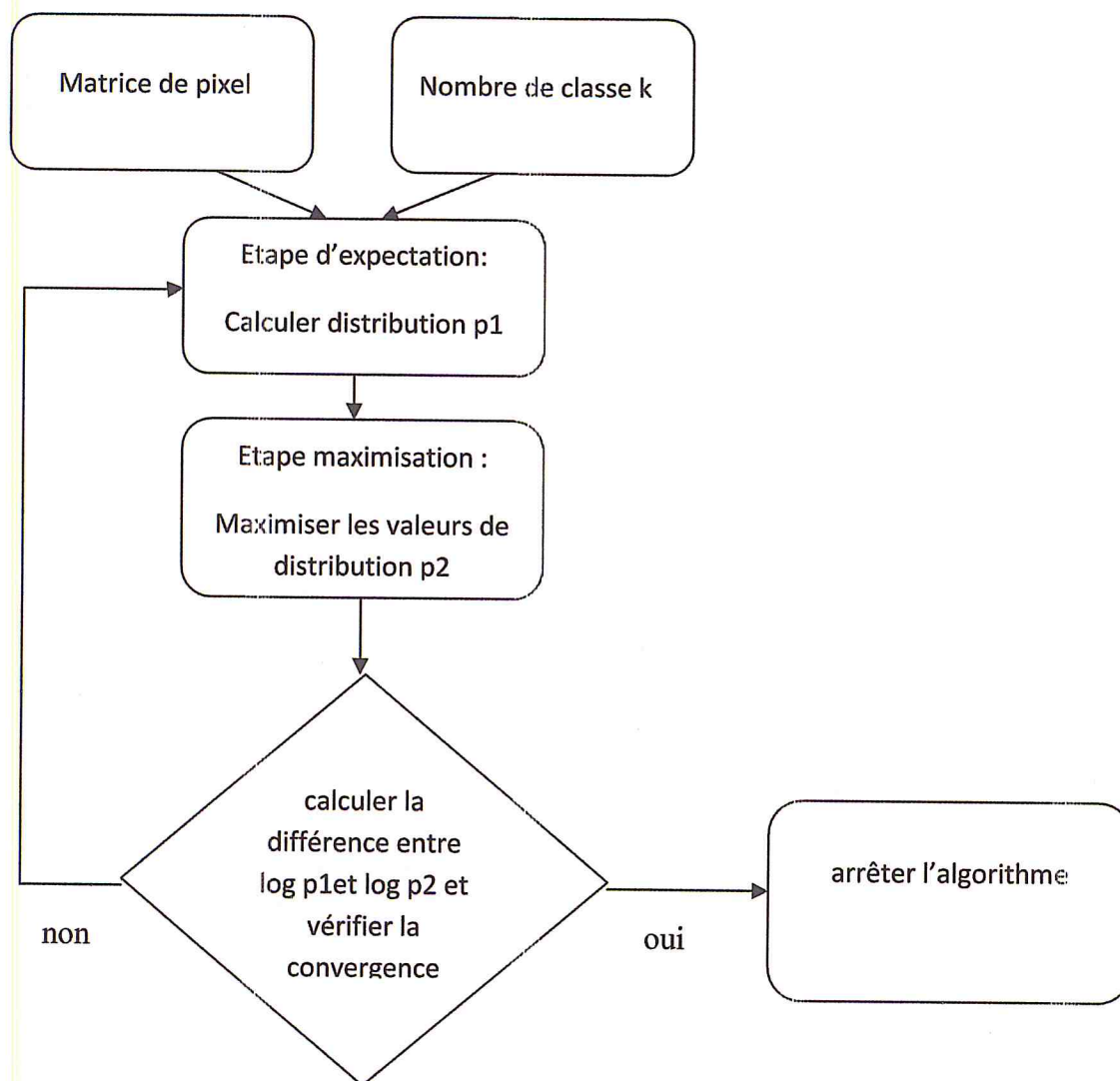


Figure 3.7. le déroulement d'EM dans notre programme

Dans notre programme pour appliquer la méthode EM, il faut mettre comme entrée la matrice de pixels et le nombre de classe k . Après dans l'étape expectation, on calcule la probabilité de la vraisemblance p_1 . Ensuite, dans l'étape maximisation, on fait la maximisation des probabilités p_2 . Après on calcule la différence entre $\log p_1$ et $\log p_2$, si la différence est inférieure à epsilon, on dit que l'algorithme est convergé et on sort de la boucle.

L'algorithme suivante représente le programme principal de notre travail ainsi que l'application d'EM :

 Algorithme EM

- Créer Histogramme de l'image : $h = \text{Histogram}(\text{image})$;
- Initialiser les paramètres : μ : la moyenne, v : la variance, p : la proportion ;
- Etape expectation :
 - calculer la probabilité de la vraisemblance : $prb1 = \text{Distribution}(\mu, v, p, x)$;
 - calculer la somme de probabilité $sommep = \sum prb1$;
 - calculer log de probabilité ($\log prb1$) ;
- Etape maximisation :
 - pour j allant de i à k (nombre de classe) :
 - calculer les nouvelles valeurs de : μ, v, p ;
 - end.
 - calculer la probabilité prb : $prb2 = \text{Distribution}(\mu, v, p, x)$;
- Calculer la condition d'arrêter :
 - calculer la somme de probabilité $sommep = \sum prb2$;
 - calculer log de probabilité ($\log prb2$) ;
 - si la différence entre $\log prb1$ et $\log prb2$ est très petite, on s'arrête.
- Dessiner les graphes de :
 - L'histogramme, les distributions, la somme de distribution.
- Calculer le masque de l'image ;

Avec la matrice de pixel on crée l'histogramme. Ensuite on applique la méthode EM, on initialise les paramètres (μ, v, p), on calcule l'étape d'expectation et l'étape de maximisation et la condition d'arrêter. Si on s'arrête, on dessine les graphes de l'histogramme, les distributions, la somme de distribution, et on calcule le masque de l'image, ce dernier, c'est une matrice qu'elle contient les valeurs entre 0 à $k-1$ avec k le nombre de classe, elle représente les clusters.

3.7. conversion du masque calculé dans l'étape précédente :

Le masque calculé par la méthode EM est sous forme d'une matrice pour mieux comprendre les résultats de cette méthode nous avons converti la matrice en image qui détermine les clusters, le programme suivant montre les différentes étapes de la conversion :

Algorithme

- Créer un objet d'image m ;
Pour chaque élément ij de la matrice :
 si (élément ij = la classe k) alors couleur=couleur k ;
 finsi,
 Couleur de pixel ij de l'objet image = couleur k ;
fait ;
 - Afficher l'image
-

On crée un objet d'image m qui est sous forme d'une matrice vide qu'elle a la taille de la matrice masque .Comme on a dit dans le paragraphe précédent ,les valeur de la matrice « masque » allant de 0 à k-1. On parcourt toute la matrice et on donne à chaque valeur une couleur c'est-à-dire on obtient k couleur.Après on met la valeur de la couleur k dans m.

Les figures suivantes representent le clustering d'image après l'application de notre programme :

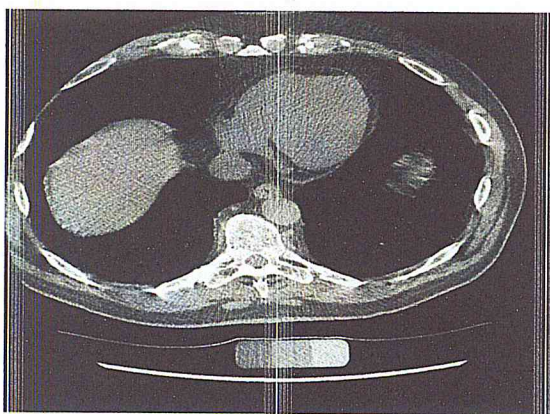


Figure 3.8. Image originale de BD1

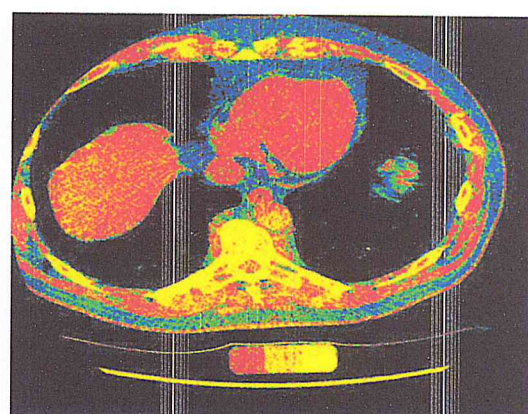


Figure 3.9. Clustering d'image avec k=5

3.8. Extraction d'un cluster :

L'extraction d'un cluster se fait en retirant le cluster choisi et mettre les autres clusters en noir, comme le montre le programme suivant :

Algorithme Extraction d'un cluster

- Créer un objet d'image m ;
 - Pour chaque élément ij de la matrice :
 - si (élément ij = la classe k chercher) alors
 - Couleur = couleur k ;
 - sinon (si élément ij = ! la classe k chercher)
 - couleur = couleur2 ;
 - finsi,
 - Couleur de pixel ij de l'objet image = couleur ;
 - fait ;
 - Afficher l'image.
-

On crée un objet d'image m qui est sous forme d'une matrice vide qu'elle a la taille de la matrice masque . Sur la matrice « masque » on peut extraire un seul cluster, on parcourt toute la matrice , si la valeur de la classe cherché égale à la valeur d'un element de la matrice on donne une couleur k , et les autres valeurs on les donne la couleur noir. Après on met la valeur de la couleur k et la couleur noir dans m.

la fégure suivante represente l'image d'un seul cluster, dans notre exemple le cluster jaune :

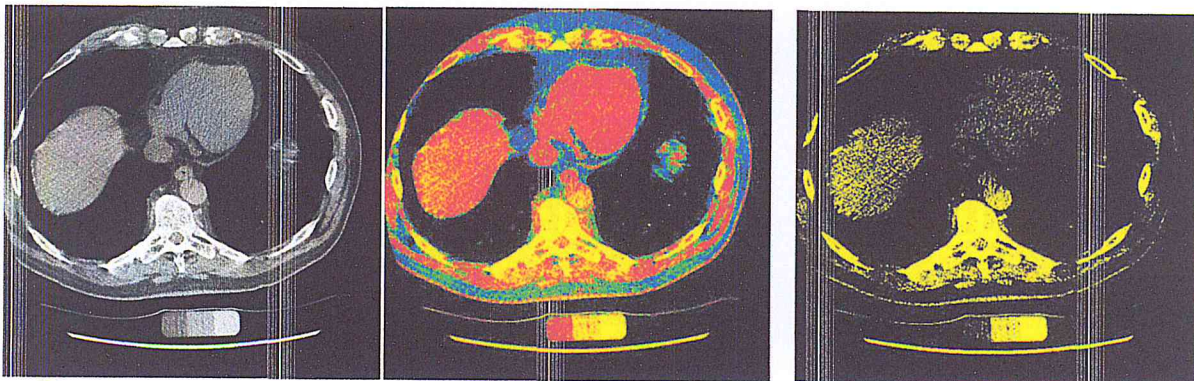


Figure 3.10. Extraction de cluster jaune de l'image 48 BD1

Maintenant nous allons présenter un exemple de l'application de notre programme sur une image de la base Dataset10 avec un nombre de classe $k=4$:

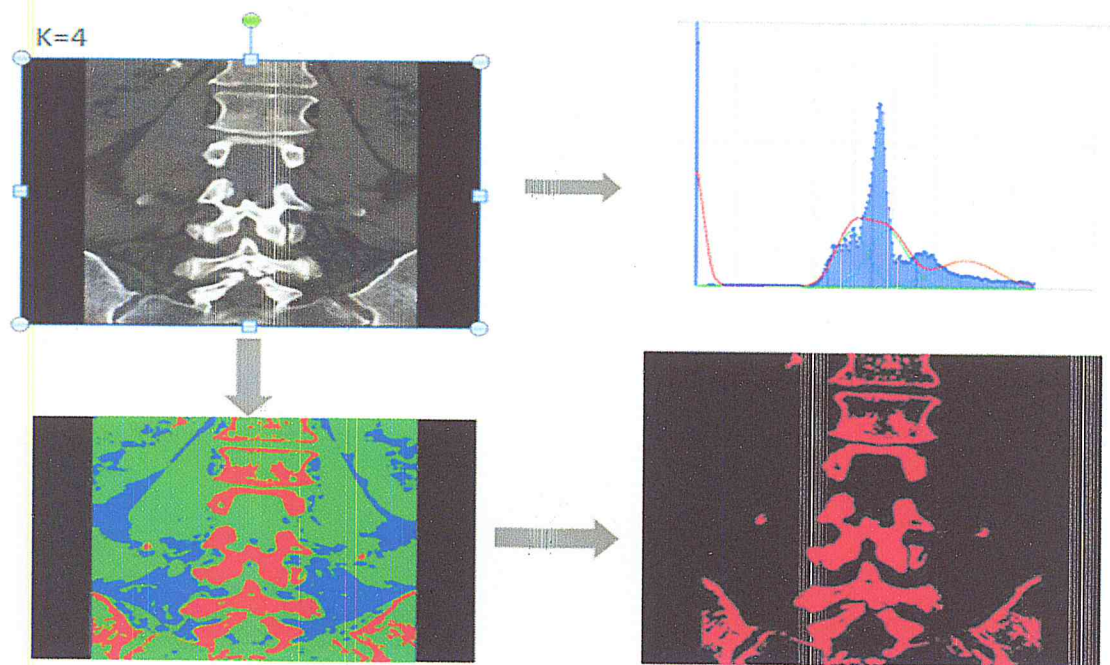


Figure 3.11. Application de notre programme sur l' image 33 de base de donnée Dataset 10

On applique sur l'image original la méthode EM, on obtient : premierement le graphe qui represente l'historgramme de l'image et les probabilité « les gaussiennes ».Deuxiemement , une image labélisée avec 4 cluster(rouge,bleu,vert,noir).A la fin, on extraire un seul cluster qui est en rouge .

3.9. Empilement et reconstruction 3D des images ct-scan:

La géométrie des modèles 3d de l'anatomie humaine sont de plus en plus pertinente dans le monde numérique. les modèles d'images 3d construits à partir de plusieurs tranches 2d de la tomодensitométrie(ct) peuvent aider les médecins à comprendre la complexité de l'anatomie présents dans les tranches.

Dans la partie précédente nous avons fait la segmentation basé sur clustering des images CT et on a obtenue comme résultat des clusters, dans cette partie nous allons faire l'empilement et la reconstruction 3d de ses clusters.

Principe de l'empilement:

Le principe de la reconstruction de formes tridimensionnelles à partir de piles d'images 2d se fait par des techniques de rendu volumique qui sont basé sur une segmentation.

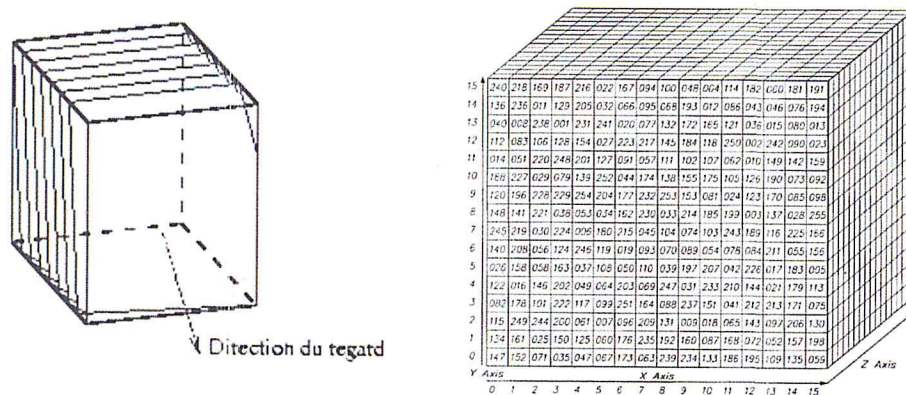


Figure3.12. Empilement des images

Etape de la reconstruction :

- lire les slices 2D.
- construire un cube de pixel à partir de l'ensemble des slices.
- créer un univers virtuel 3D pour la visualisation 3D.

La scène 3D :

Une zone de l'écran qui permet de visualiser la scène 3D

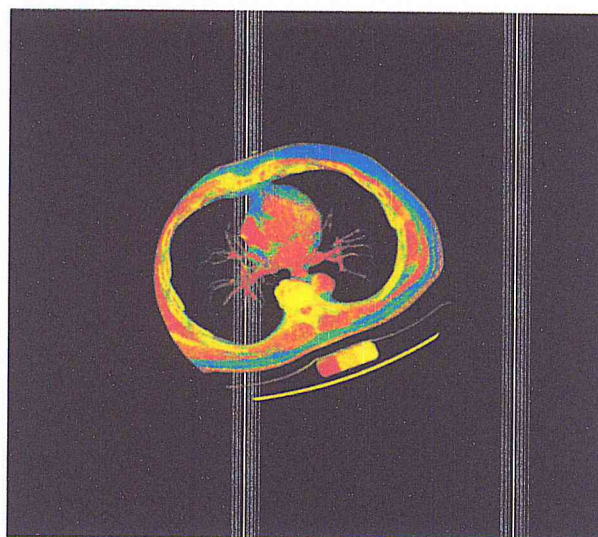


Figure3.13.La visualisation 3D d'une image 48 BD1

3.10. Conclusion :

Dans ce chapitre nous avons donné une description de notre méthode de clustering des images CT-Scan ainsi que le concept de la méthode EM, qui utilise ces images comme entrée, ensuite nous avons introduit les résultats de cette méthode qui terminent avec l'extraction du cluster. Dans la deuxième partie nous avons parlé sur l'empilement et la reconstruction 3D des images CT-SCAN.

Dans le chapitre suivant, nous expliquerons la dernière étape de notre travail qui est l'implémentation, expérimentation et interprétation des résultats.

CHAPITRE 4

**Implémentations, expérimentation
et interprétation des résultats**

4.1. Introduction :

L'objectif de ce chapitre est de présenter l'application développée, les traitements illustrés par des images CT-SCAN dans le but de clustering et reconstruction 3d issues de la base de données **CVIP Spinal CT et Dataset10**.

Notre système permet le clustering des images CT-SCAN de la base **CVIP Spinal CT et Dataset10** suivant la méthode EM, qui est l'une des clusters hiérarchiques les plus populaires dans le clustering des images CT-SCAN grâce à son rendement élevé, sa capacité à traiter des données de grande dimension, et sa flexibilité dans la modélisation de diverses sources de données .

4.2. Environnement de travail

4.2.1 Matériel

Pour le développement de notre application, nous avons utilisé un microordinateur, menu essentiellement de :

- Intel Core i5 3632QM (2.20GHz)
- 4GB Memory
- 1600 x 900
- Windows 7 (32 bits)

4.2.2 Langage

Java est un langage de programmation informatique orienté objet créé par James Gosling et Patrick Naughton, employés de Sun Microsystems, avec le soutien de Bill Joy (cofondateur de Sun Microsystems en 1982), présenté officiellement le 23 mai 1995 au SunWorld. [27]

Le langage Java reprend en grande partie la syntaxe du langage C++, très utilisé par les informaticiens.

Java3D est une bibliothèque logicielle de haut niveau facilitant la création, la visualisation et la manipulation de scènes 3D. Java3D fournit une API en java et bénéficie donc de la portabilité de ce langage. Java3D est porté sur différentes plateformes. Il utilise (de manière transparente pour l'utilisateur) des bibliothèques graphiques de plus bas niveau telles qu'[OpenGL](#) ou [DirectX](#) pour réaliser le rendu 3D.

L'utilisation de Java3D nécessite que les packages `javax.media.j3d`, `javax.vecmath`, et `com.sun.j3d.utils` soient installés sur notre système.[28]

<http://www.java3d.org/introduction.html>

4.2.3 Environnement de développement :

NetBeans est un environnement de développement intégré (EDI), placé en open source par Sun en juin 2000 sous licence CDDL (Common Development and Distribution License) et GPLv2. En plus de Java, NetBeans permet la prise en charge native de divers langages tels le C, le C++, le JavaScript, le XML, le PHP et le HTML, ou d'autres (dont Python et Ruby) par l'ajout de greffons. Il offre toutes les facilités d'un IDE moderne (éditeur en couleurs, projets multi-langage, refactoring, éditeur graphique d'interfaces et de pages Web).[29]

Compilé en Java, NetBeans est disponible sous Windows, Linux, Solaris (sur x86 et SPARC), Mac OS X ou sous une version indépendante des systèmes d'exploitation (requérant une machine virtuelle Java). Un environnement Java Développement Kit JDK est requis pour les développements en Java.

NetBeans constitue par ailleurs une plate-forme qui permet le développement d'applications spécifiques (bibliothèque Swing (Java)). L'IDE NetBeans s'appuie sur cette plate-forme.

Abstract Window Toolkit (AWT) est une bibliothèque graphique pour Java, faisant partie de Java Foundation Classes (JFC). Cette bibliothèque a été introduite dès les premières versions de Java ; depuis Java 2, la bibliothèque de gestion de fenêtre officielle est Swing. Toutefois, AWT sert encore de fondement à Swing, dans la mesure où de nombreuses classes Swing héritent de classes AWT.

AWT emploie les composants natifs de la plate-forme, alors que Swing utilise des composants en pur Java. [30]

Avec l'apparition de Java 8 en mars 2014, **JavaFX** devient la bibliothèque de création d'interface graphique officielle du langage Java, pour toutes les sortes d'application (applications mobiles, applications sur poste de travail, applications Web), le développement de son prédécesseur Swing étant abandonné (sauf pour les corrections de bogues).

JavaFX est désormais une pure API Java (le langage de script spécifique qui a été un temps associé à JavaFX est maintenant abandonné).

JavaFX contient des outils très divers, notamment pour les médias audio et vidéo, le graphisme 2D et 3D, la programmation Web, la programmation multi-fils etc. [31]

4.3. Données utilisées :

Nous avons utilisé 2 bases de données internationales contenant des images médicales. Ces données sont téléchargées à partir du site :

<http://spineweb.digitalimaginggroup.ca/spineweb/index.php?n=Main.Datasets>

la première est « Dataset4 : CVIP Spinal CT », qui représente des images TDM de 5 patients (formats dicom et BMP) sur segmentation 3D des os vertébraux à partir d'images CT volumétriques à l'aide de coupures graphiques [32] [33].

Cette base de données contient 296 images pour 5 patients, ces images sont traitées et segmentées, tous les scans ont des vérités terrestres avec une taille originale (c.-à-d., 512x512).

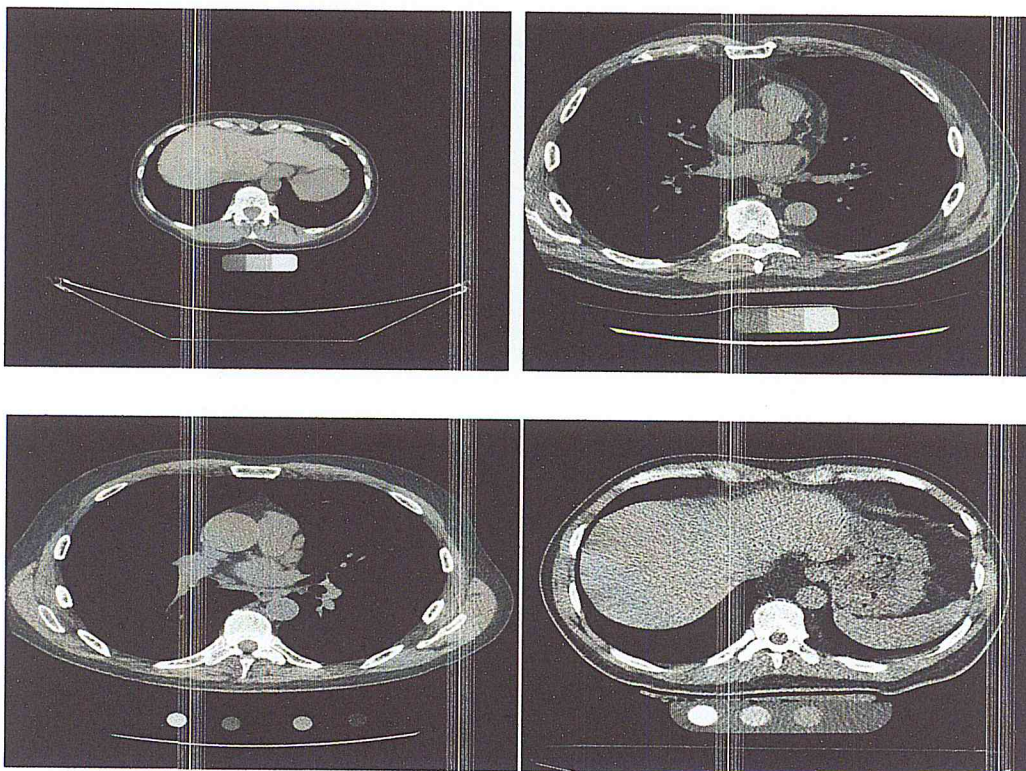
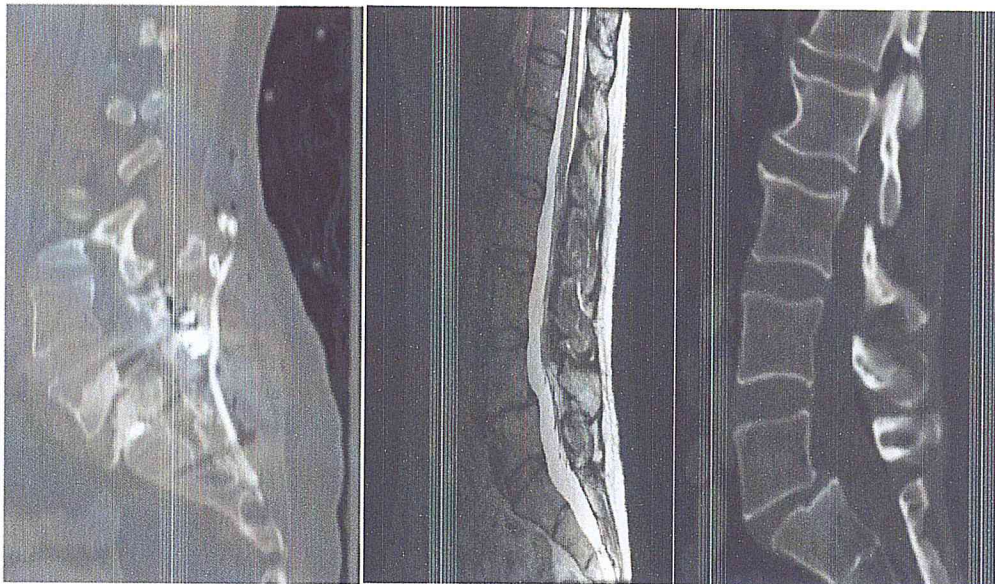


Figure 4.1 Exemple d'image de la base de données images CVIP Spinal CT image CT - SCAN (une couche de vertèbre)

La deuxième est **Dataset10** : « **Multi-Modality Vertebra Recognition in Arbitrary Views using 3D Deformable Hierarchical Model** » qui représente l'analyse de la structure de la colonne vertébrale qui nécessite la compréhension des emplacements des vertèbres, des poses et des noms dans les images MR et CT [34].

L'ensemble de données contient des images MR + CT de 20 sujets. Pour chaque vertèbre, l'emplacement du centre de la vertèbre 3D (à la moelle épinière) et l'orientation 3D de la vertèbre sont annotés manuellement en tant que vérité au sol.

Cette base de données contient plus de 159 images.



**Figure 4.2. Exemple d'image de la base de données Dataset10
(une couche transversale de vertèbre)**

4.4. Description d'application :

4.4.1. Interface :

Notre interface est composée de 4 boutons :

- Ouvrir image.
- Méthode EM.
- Choisir le cluster.
- Graphe.

➤ Visualisation 3D

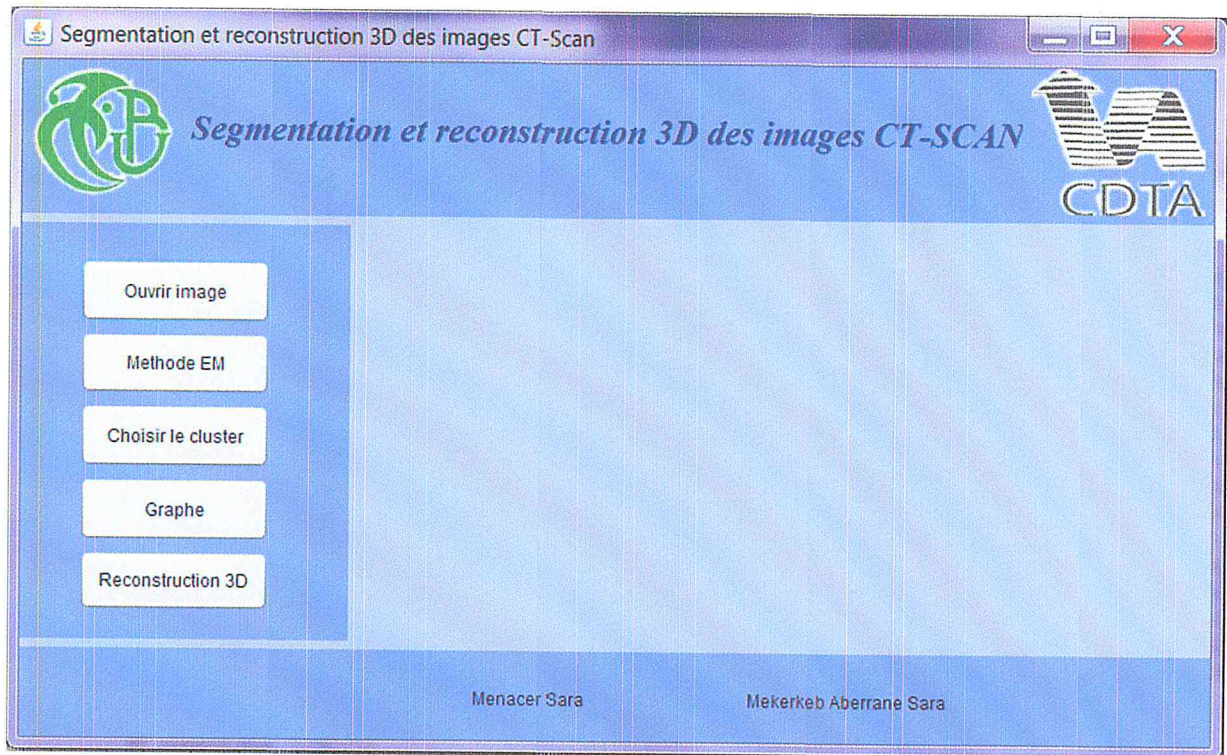
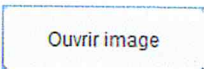


Figure 4.3.l'interface de la plateforme

4.4.2. Description des boutons :

-  : En appuyant sur ce bouton une fenêtre s'ouvre pour choisir une image qu'on veut l'ouvrir.

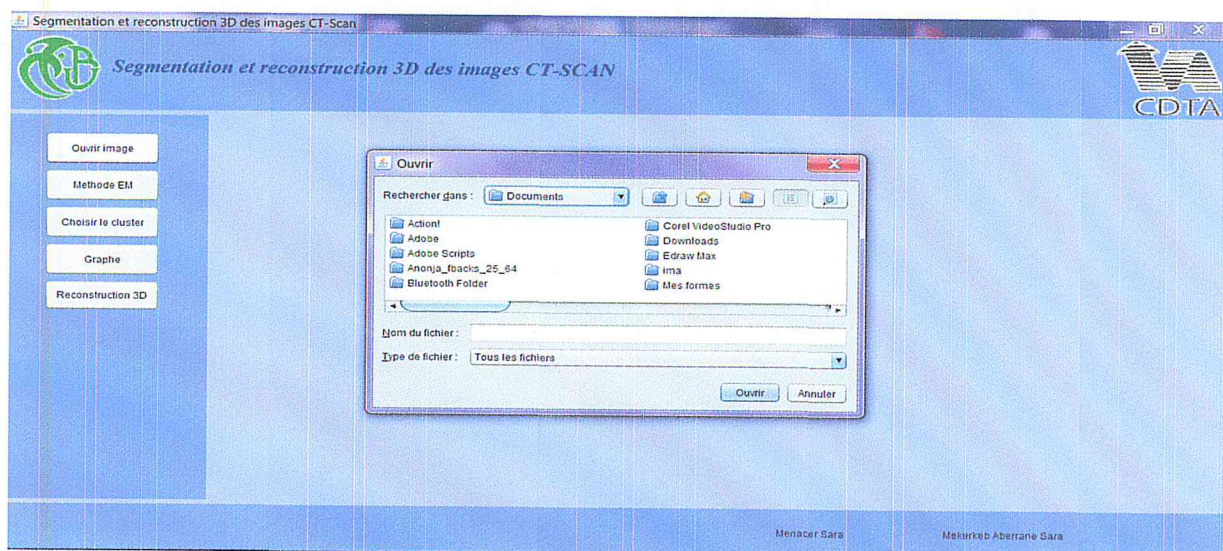


Figure 4.4.Affichage du bouton « Ouvrir image »

Une fois nous choisissons l'image, elle apparaitre sur notre interface comme suit :

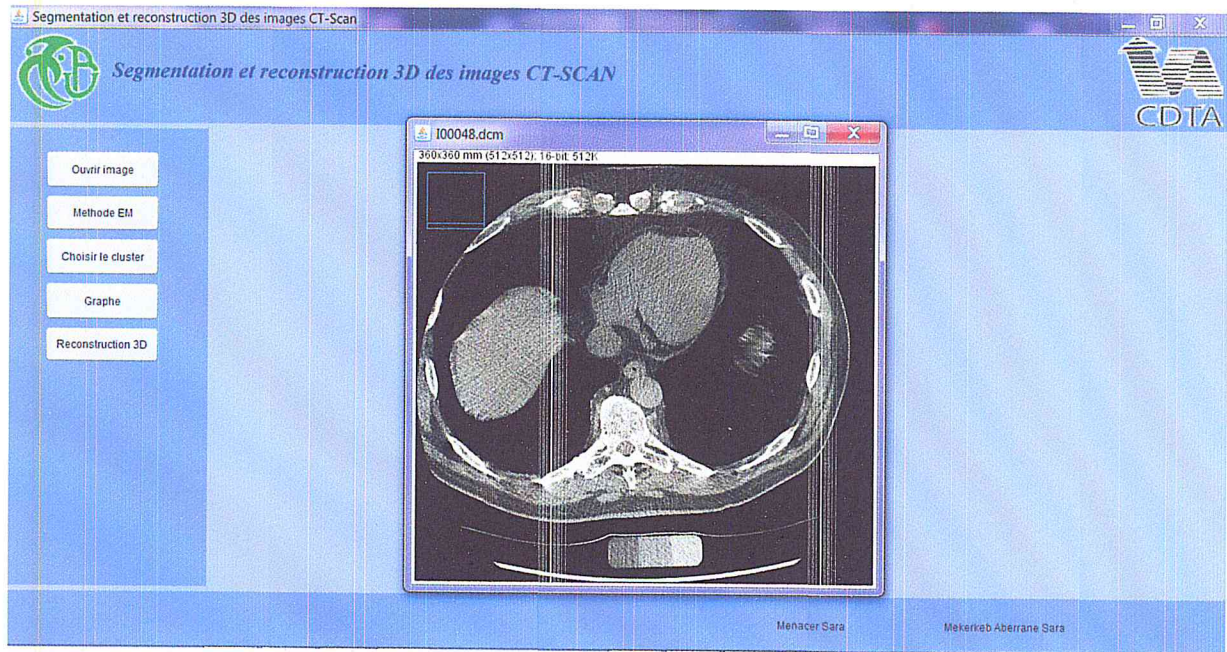


Figure 4.5. Ouverture d'image de dataset4 choisie

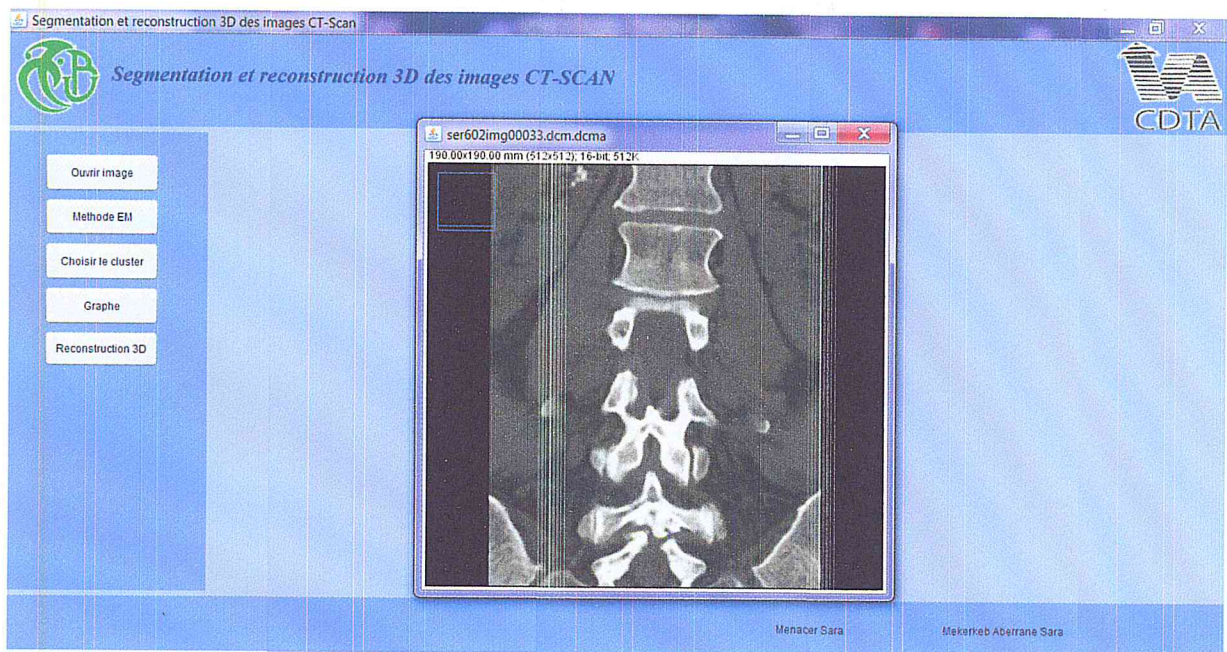


Figure 4.6. Ouverture d'image de dataset10 choisie

- : en cliquant sur ce bouton la méthode EM sera appliquée sur l'image que nous avons choisie, et nous aurons son clustering.

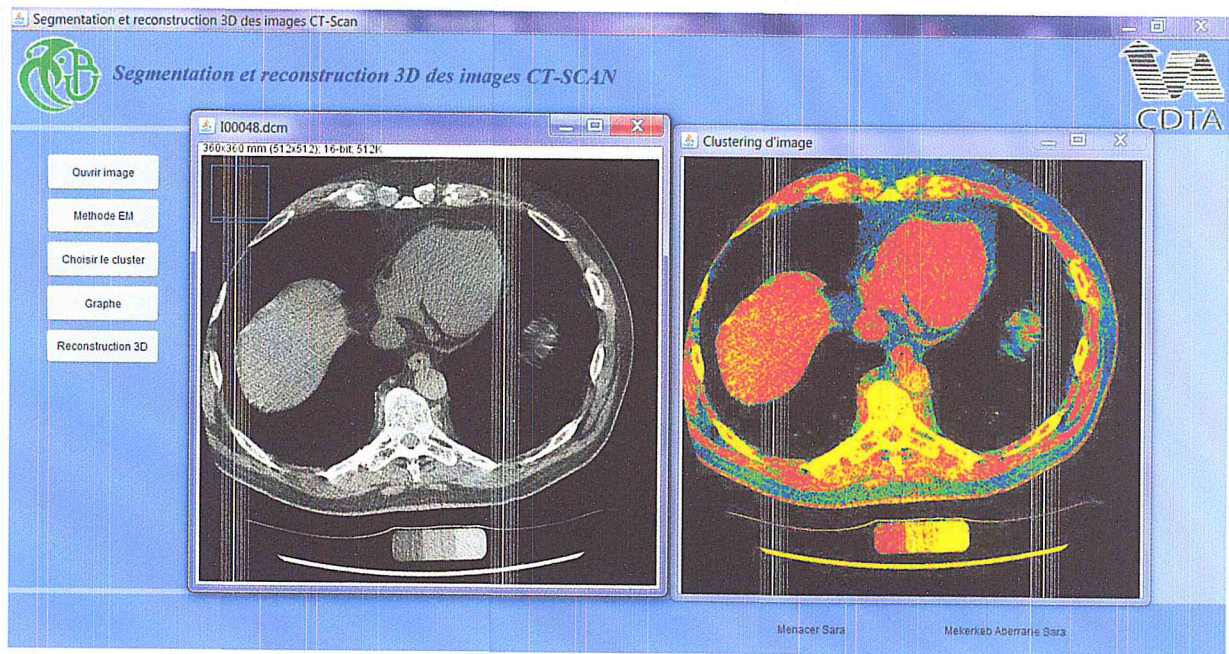


Figure 4.7. Affichage d'image (dataset4) après l'application de la méthode EM

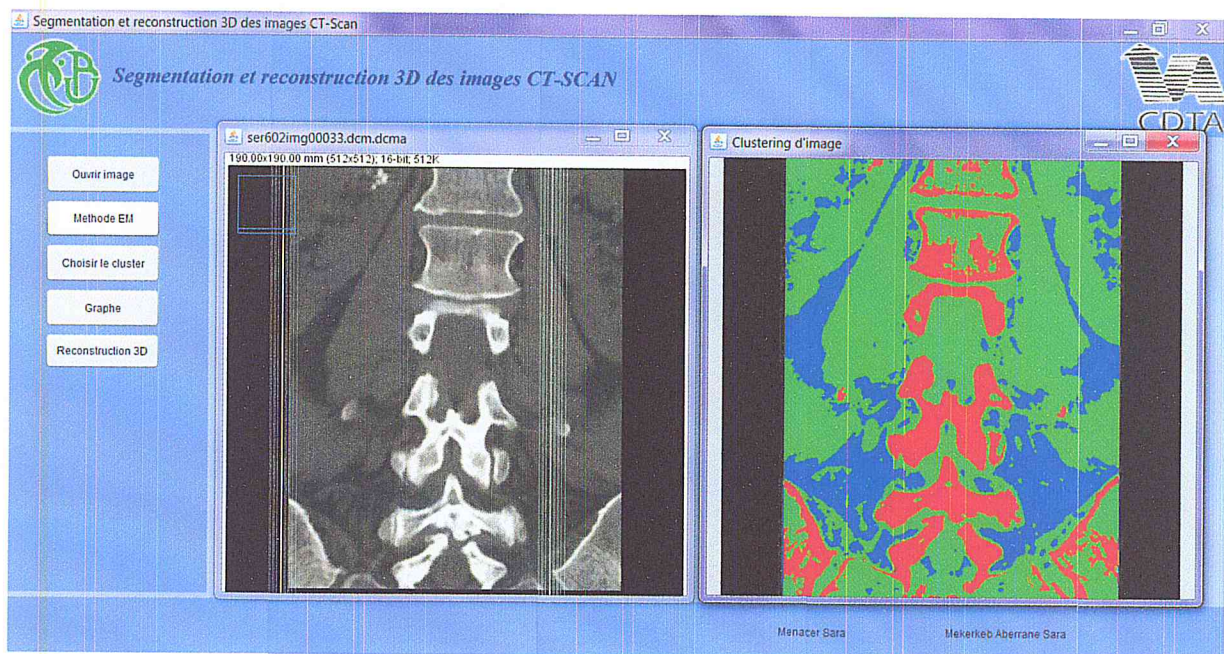


Figure 4.8. Affichage d'image (dataset10) après l'application de la méthode EM

- : en appuyant sur ce bouton nous aurons juste le cluster qu'on veut voir.

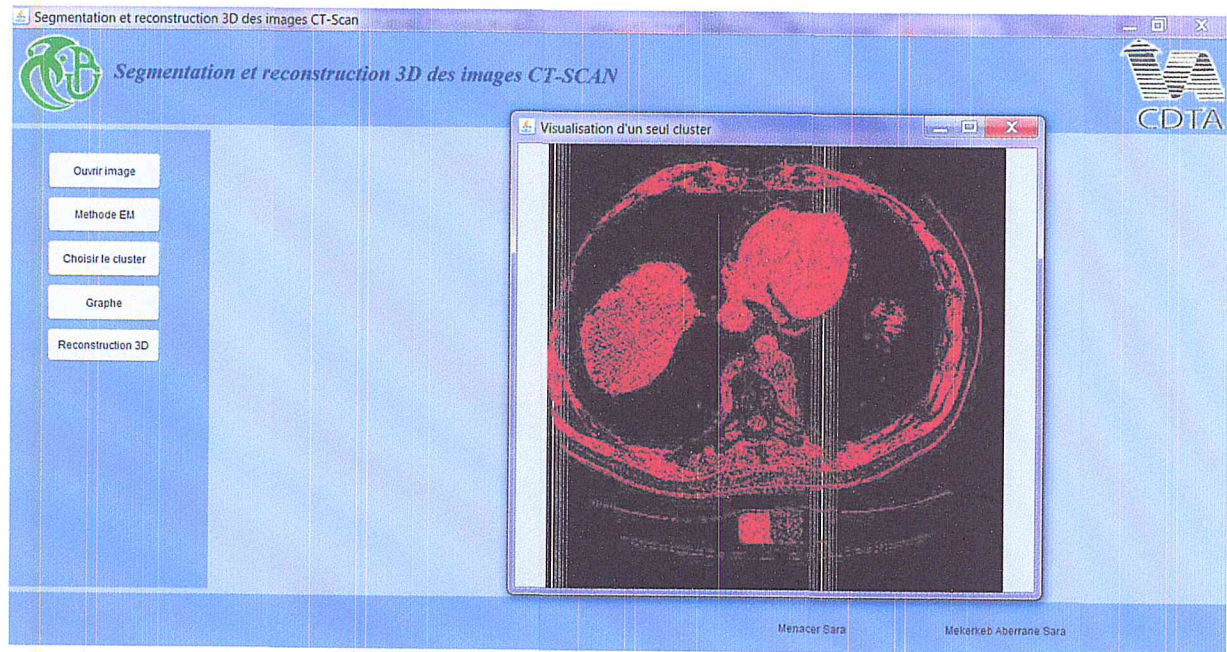


Figure 4.9. Affichage du cluster choisi (image de dataset4)

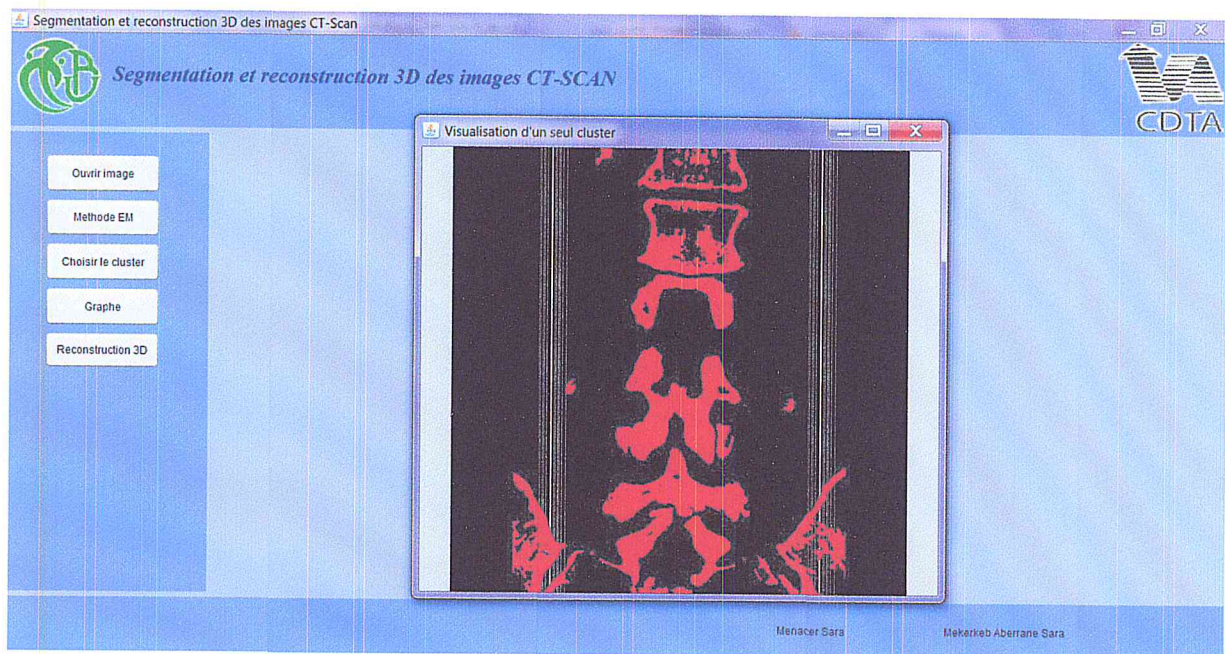


Figure 4.10. Affichage du cluster choisi (image de dataset10)

- : ce bouton affiche le graphe contenant l'histogramme de l'image, les distributions ainsi que leur somme.



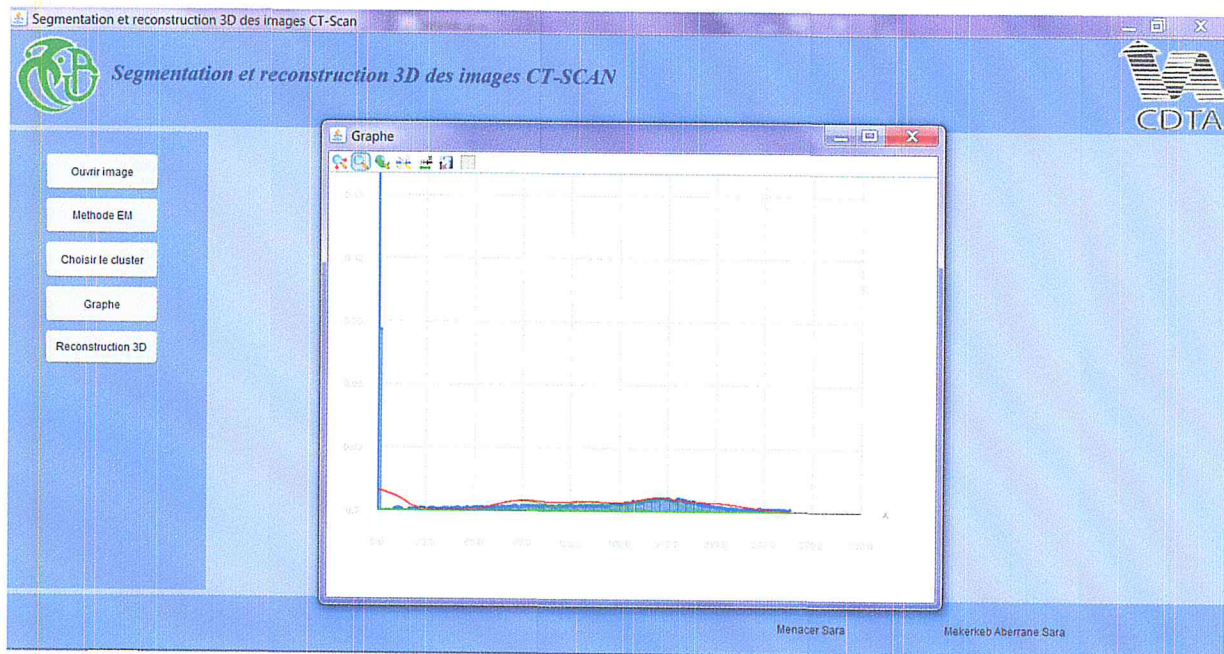


Figure 4.11. Affichage du graphe

- **Reconstruction 3D** : ce bouton affiche la visualisation 3D d'un ensemble des image CT-SCAN :

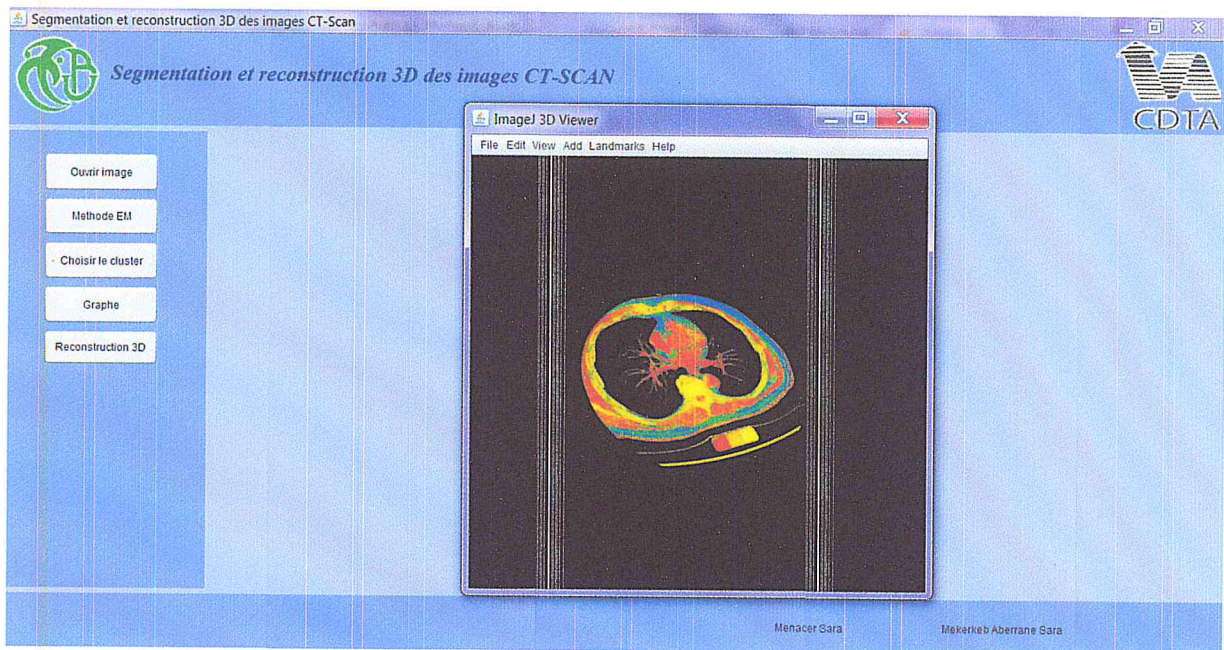


Figure 4.12. Visualisation 3D(ensemble d'images de dataset4)

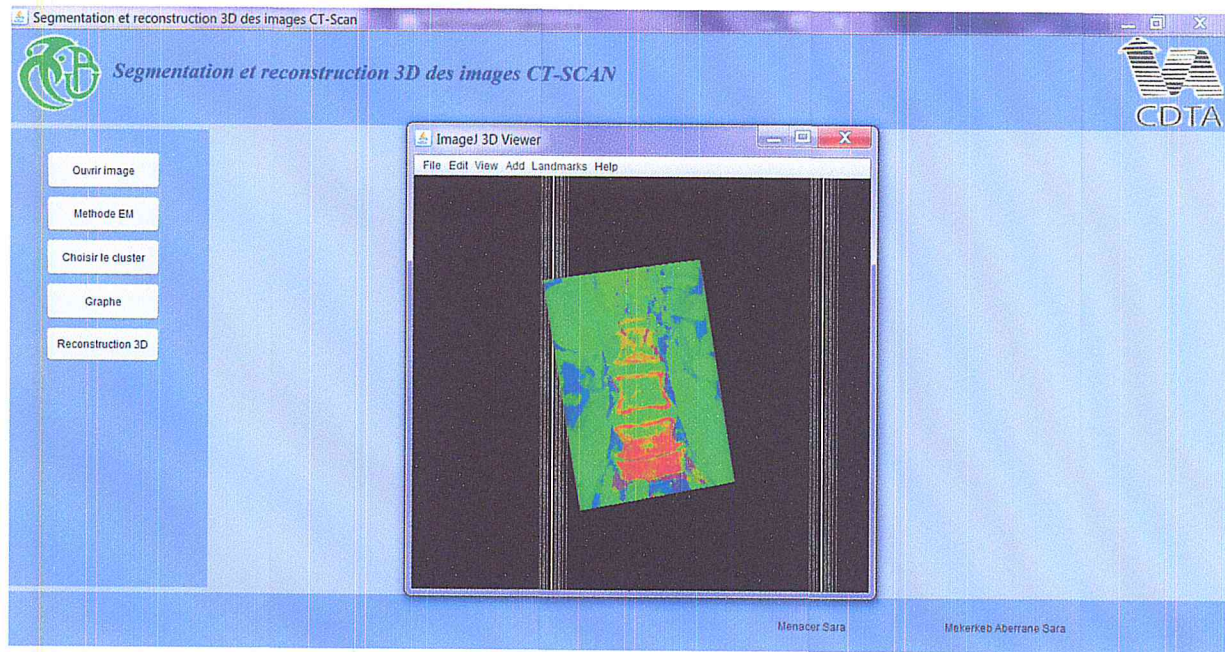


Figure 4.13. Visualisation 3D (ensemble d'images de dataset10)

4.5. Conclusion :

Nous avons présenté dans ce chapitre la plateforme que nous avons réalisée avec l'explication de chaque bouton, ainsi que les résultats obtenues en appliquant la méthode EM.

Nous avons démontré que le clustering des images CT-scan est basée sur la méthode EM.

Conclusion générale

La méthode EM et à travers ses diverses variantes reste une des méthodes qui ont connu un succès d'usage ces dix dernières années, dans l'univers des méthodes classificatoires de DATA MINING, ce n'est pas de hasard qu'elle a marqué son succès, mais grâce à son adaptabilité aux données (manquantes, grande masse de données-images —....), à sa simplicité de modification, à sa robustesse en ce qui concerne le travail complexe qu'elle fasse simultanément (estimation des paramètres et détermination des classes en même temps), à sa force d'estimation pour les grandes dimensions... tous ces critères font d'elle une méthode digne d'être améliorée et moins d'être à bout de ces fonctionnalités, l'objectif de notre travail est le clustering à partir des images CT-Scan par cette méthode.

Le travail présenté dans ce mémoire a révélé des résultats encourageants qui indiquent que le système de clustering des images CT-Scan proposé a réussi à simplifier les résultats.

Avec un bon prétraitement, ce système peut permettre à aider les médecins en leur fournissant une opinion d'orientation afin qu'ils puissent poser un diagnostic avec certitude.

La première partie a été consacrée aux généralités de notre travail qui présente la description des notions essentielles liées à la technique d'imagerie la tomodensitométrie.

Dans la deuxième partie, nous avons passé en revue l'état de l'art sur le clustering avec ces méthodes en expliquant spécialement la méthode (EM) appliquée dans notre travail.

La troisième partie est consacrée à l'application de la méthode (EM) sur notre base de données, nous avons présenté la méthodologie du travail accompli, ainsi que les résultats obtenus.

Dans la dernière partie de notre travail, nous avons implémenté ce travail qui permet le clustering d'une image CT-Scan (Dicom).

À l'issue de ce travail, quelques perspectives sont dégagées on peut citer :

1. Rendre le nombre de classes (K) automatique.
2. Améliorer les performances de cette méthode (augmenter la capacité de mémoire en empilement).
3. Améliorer la visualisation 3D.

Bibliographie

Références bibliographiques:

- [0] Le site : <https://sites.google.com/site/limageriemedicale/radiographie/historique>: premiere imagerie medicale 1895,consulté le 10/12/2016.
- [1] Le site : <http://www.radiologie-macon.com/fr/les-techniques-d-imagerie>, consulté le 28/11/2016.
- [2]Hounsfield GN, « *Computerized transverse axial scanning (tomography). 1. Description of system* », *Br J Radiol*, vol. 46, n° 552, 1973, p. 1016-22.
- [3] Christian Nordqvist. 2016. CT Scan (CAT Scan): How Do They Work?. Consulté le 28/11/2016 (<http://www.medicalnewstoday.com/articles/153201.php>).
- [4] Le site : <https://www.inserm.fr/thematiques/technologies-pour-la-sante/dossiers-d-information/imagerie-fonctionnelle-biomedicale>, consulté le 05/01/2017 .
- [5] Le site : http://www.ac-nice.fr/ia06/iencagnes/file/tice/tuto/Image_numerique.pdf, consulté le 05/01/2017.
- [6] Le site : http://foad.refer.org/IMG/pdf/D226_Chapitre-5.pdf, consulté le 05/01/2017
- [7] Le site : <http://www.crdp.ac-grenoble.fr/image/general/general.htm>, consulté le 05/01/2017.
- [8]Le site : <https://sites.google.com/site/dataclusteringalgorithms/file-cabinet>, consulté le 07/02/2017.
- [9]Le site : <http://robotics.stanford.edu/people/nilsson/MLBOOK.pdf>, consulté le 22/02/2017.
- [10] Shai Shalev-Shwartz and Shai Ben-David, Understanding Machine Learning , by Cambridge University Press,2014.
- [11]Le site : http://cs.du.edu/~mitchell/mario_books/Introduction_to_Machine_Learning_-_2e_-_Ethem_Alpaydin.pdf ,consulté le 09/03/2017
- [12]Le site : <https://www.cs.ubc.ca/~murphyk/MLbook/pml-intro-22may12.pdf>, consulté le :20/03/2017

Bibliographie

- [13] [MacQueen, 1967] : J. B. MacQueen (1967) : "Some Methods for classification and Analysis of Multivariate Observations, Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability", Berkeley, University of California Press, no 1, pp281-297.1967.
- [14] Richard O. Duda, Peter E. Hart, David G. Stork, Pattern Classification, Wiley interscience, 2001
- [15] [Karypis & al., 1999] : Karypis, G., Eui-Hong, H., et Kumar, V. Chameleon : Hierarchical Clustering Using Dynamic Modeling. Computer, no 32(8) :68-75, 1999.
- [16] [Lance & Williams, 1967] : Lance, G.N., & Williams, W.T. : A general theory of classificatory sorting strategies : I. Hierarchical systems. Computer Journal, no 9, pp 373-380, 1967.
- [17] [Gesu, 1988] V. di Gesu. « Mathematical Morphology and Image Analysis : A Fuzzy Approach ». Workshop on Knowledge-Based Systems and Models of Logical Reasoning, Reasoning, 1988.
- [18] [Day, 1969] : N.E. Day : Estimating the Components of a Mixture of Normal Distributions. Biometrika, no 56, pp 464-474, 1969.
- [19] [Dempster and al., 1977] : A.P. Dempster, N.M. Laird , and D.B. Rubin :Maximum-vraisemblance from incomplete data via the EM algorithm", J.Royal Statist. Society, B39, pp1-38, 1977
- [20] [Sean Borman , 2009] : Sean borman , The Expectation Maximization Algorithm A short tutorial , 2009
- [21] [Bailey & Elkan, 1995] : T. L. Bailey and C. Elkan , "Unsupervised learning of multiple motifs in biopolymers using expectation maximization," Machine Learning, vol. 21, pp. 51–80, 1995.
- [22] [McLachlan et Krishnan, 1997] : G. McLachlan, and T. Krishnan, The EM algorithm and extensions. Wiley series in probability and statistics. John Wiley & Sons.1997.
- [23] [McLachlan et Krishnan, 2008] : Geoffrey. McLachlan and Thriyambakam. Krishnan. The EM Algorithm and Extensions (2nd edition). Wiley, New Jersey,2008.»

Bibliographie

- [24] [McLachlan & chang, 2004] : G.J. McLachlan and S.U. Chang. Mixture modelling for cluster analysis. Statistical Methods in Medical Research, no 13, pp 347–361,2004.
- [25] [Gower & legendre, 1986] : Gower, J. C. & P. Legendre : Metric and Euclidean properties of dissimilarity coefficients, Journal of Classification, vol 3, pp. 5-48 .1986.
- [26] [Day, 1969] : N.E. Day : Estimating the Components of a Mixture of Normal Distributions. Biometrika, no 56, pp 464-474, 1969.
- [27]Le site : en Arnold, James Gosling et David Holmes, Le Langage Java, 2001 (ISBN 978-2-7117-8671-8)
- [28]Le site : <http://www.java3d.org/introduction.html> , consulté le 10/06/2017
- [29] Le site : <http://netbeans.org/community/articles/interviews/yarda-tulach.html> ,consulté le 02/06/2017
- [30] Le site :
<https://www.cis.upenn.edu/~bcpierce/courses/629/jdkdocs/api/java.awt.Toolkit.html>, visité le 26/03/2017
- [31] Le site : Documentation officielle d'Oracle pour JavaFX 8, consulté le 28/05/2017.
- [32] Melih S. Aslan, Ahmed Shalaby et Aly A. Farag. « Méthode de segmentation cliniquement souhaitée pour les corps vertébraux ». Dans Imagerie biomédicale (ISBI), 2013 IEEE 10ème symposium international sur, pp. 840-843. IEEE, 2013.
- [33] Melih S. Aslan, Asem Ali, Ham Rara, Ben Arnold, Aly A. Farag, Rachid Fahmi et Ping Xiang. « Une nouvelle segmentation 3D des os vertébraux à partir d'images cT volumétriques à l'aide de coupures graphiques ». Dans Advances in Visual Computing, p. 519-528. Springer Berlin Heidelberg, 2009.
- [34] Le site : <http://spineweb.digitalimaginggroup.ca/spineweb/index.php?n=Main.Datasets>, consulté 28/05/2017.

