Plateforme d'évaluation multi-critères pour les algorithmes de traitement d'images médicales

par

Pierre LAURENT

MÉMOIRE PRÉSENTÉ À L'ÉCOLE DE TECHNOLOGIE SUPÉRIEURE
COMME EXIGENCE PARTIELLE À L'OBTENTION
DE LA MAÎTRISE EN GÉNIE
CONCENTRATION TECHNOLOGIES DE LA SANTÉ
M.Sc.A.

MONTRÉAL, LE 24 JANVIER 2017

ÉCOLE DE TECHNOLOGIE SUPÉRIEURE
UNIVERSITÉ DU QUÉBEC

PRÉSENTATION DU JURY

CE MÉMOIRE A ÉTÉ ÉVALUÉ

PAR UN JURY COMPOSÉ DE:

M. Jacques A. De Guise, Directeur de Mémoire
Département de génie de la production automatisée à l'École de technologie supérieure

Mme Nicola Hagemeister, Co-directeur
Département de génie de la production automatisée à l'École de technologie supérieure

M. Carlos Vázquez, Président du Jury
Département de génie logiciel et des TI à l'École de technologie supérieure

Mme Neila Mezghani, Examinateur Externe
Département des sciences et technologies à la TÉLUQ

IL A FAIT L'OBJET D'UNE SOUTENANCE DEVANT JURY ET PUBLIC

LE 17 JANVIER 2017

À L'ÉCOLE DE TECHNOLOGIE SUPÉRIEURE

REMERCIEMENTS

Je tiens à remercier toutes les personnes qui m'ont aidées dans la mise en place de ce projet.

Je remercie Jacques de Guise, mon directeur de recherche, de m'avoir accueilli et accepté au LIO, d'abord dans le cadre d'un stage puis de m'avoir offert l'opportunité de continuer en maîtrise de recherche. Merci à Nicola Hagemeister pour l'encadrement et l'aide précieuse apportée, en tout temps, dans l'organisation de ce projet. Merci à Carlos Vazquez et Thierry Cresson de m'avoir encadré sur la partie technique et le développement de la plateforme. Encore une fois, merci de m'avoir fait confiance et de m'avoir aidé et orienté tout au long de la maîtrise. Grâce à vous, je pense avoir vécu une aventure à la fois enrichissante sur le plan technique puisque j'ai découvert une passion mais également et surtout sur le plan humain.

Je remercie toutes les personnes qui ont également collaboré de près ou de loin au projet. Merci à Nathalie Bureau, radiologue du CRCHUM, de m'avoir aidé pour obtenir les données nécessaires à ce projet. Merci à Joseph et Julien d'être venus les fins de semaine au laboratoire pour produire ces données. Merci à Meriem et au CRAN de Nancy avec qui j'ai pu collaboré pour développer mon application en 3D. Merci à Gerald de m'avoir toujours aidé lorsque j'avais besoin en informatique. Merci à Caroline pour l'aide logistique et les relectures de traductions. Je remercie également l'entreprise EOS Imaging et en particulier Nasr Makni et Maurice Delplanque pour m'avoir soutenu dans ce projet.

Enfin, j'aimerais remercier chaleureusement tous ceux grâce à qui j'ai passé presque 3 ans de folie en chalets, sorties et discussions plus ou moins philosophiques, plus ou moins scientifiques au laboratoire PA, Flo, Marta, Jérôme, Boris, Sophie(s), Capu, Wendy, Ted, Alex, Nils et en dehors Jo, Manue, Sam, Laurie, Thib, Victor, Lu.

PLATEFORME D'ÉVALUATION MULTI-CRITÈRES POUR LES ALGORITHMES DE TRAITEMENT D'IMAGES MÉDICALES

Pierre LAURENT

RÉSUMÉ

Les algorithmes de segmentation d'images médicales permettent la détection de structures anatomiques. Chaque année de nombreux algorithmes sont développés. Ils aident les cliniciens dans l'interprétation de la pathologie du patient. C'est pourquoi, il est nécessaire de proposer une segmentation la plus proche possible de la réalité. Pour vérifier cela, ils sont validés en utilisant des tests de performances appliqués sur des métriques, en les comparant à une référence. Toutefois, la validation est souvent subjective et chaque algorithme a sa propre méthode de validation. Il n'existe pas de protocole standardisé et il est bien souvent difficile de pouvoir comparer les algorithmes entre eux. Par ailleurs, dans le cas de données sur des patients, les résultats de validation sont conditionnés par le modèle de référence, généralement obtenu par le biais d'un seul expert clinique.

L'objectif de ce mémoire est de proposer une plateforme d'évaluation d'algorithmes de traitements d'images médicales.

Le processus d'évaluation associé à cette plateforme est divisé en deux étapes principales. Dans un premier temps, nous proposons une manière d'obtenir une référence d'étude : soit en générant un étalon-or à partir de simulations mathématiques ; soit en utilisant des modèles obtenus par des experts et en générant un étalon-bronze servant de référence. Dans le cas de l'étalon-bronze, nous proposons un moyen de le caractériser à travers des paramètres de performances. Cela permet d'obtenir un intervalle de confiance de la référence et de normaliser l'étude d'algorithmes qui s'en suit. Par la suite, nous comparons le comportement d'algorithmes à la référence en définissant des critères d'évaluation en fonction des besoins du cahier des charges. Ces critères sont définis à l'aide d'un arrangement de métriques tirées de la littérature. Les résultats sont finalement reportés dans des graphiques de performances de type radar. Ces graphiques permettent une interprétation multi-critères des algorithmes et servent à déceler les points forts et points faibles de chacune des méthodes. Il est ainsi possible en étudiant plusieurs algorithmes de les comparer sur un même graphique et pouvoir avoir une interprétation très rapide des différents comportements des algorithmes. Dans ce sens, la plateforme est un outil permettant non seulement de sélectionner l'algorithme qui répond le mieux aux exigences souhaitées mais également d'avoir une idée sur les points qui mériteraient d'être améliorés pour chacun des algorithmes testés.

Cette plateforme a été mise en place pour des algorithmes de segmentation 2D et 3D ainsi que pour des maillages de reconstruction 3D. Dans la partie résultats du mémoire, nous proposons une évaluation d'algorithmes pour certaines méthodes développées au laboratoire du LIO.

Mots clés: Plateforme d'évaluation, algorithmes d'imagerie médicale, référence

A MULTI-CRITERIA EVALUATION PLATFORM FOR SEGMENTATION ALGORITHMS

Pierre LAURENT

ABSTRACT

Each year, an increasing number of algorithms and methods are developed to segment anatomical structures given X-Rays. These algorithms provide assistance for clinicians diagnosis in order to interpret patients' diseases. They, therefore, require to produce a segmentation as close as possible to reality. To ensure this, they are validated using performances tests applied on metrics, comparing the output to a reference. However, validation is very often subjective and each algorithm presents its own validation process. There is no standard protocol of validation and it is generally very challenging to compare an algorithm to others. Furthermore, in case of data taken on patients, validation are conditioned by the reference model, which is generally obtained by one clinician expert.

In this thesis, we present a platform for evaluating imaging algorithms in medical images.

This platform evaluation process has been divided in two major steps. In the first instance, we propose a way to obtain a valid reference : either by generating a gold standard on mathematical simulations ; either by using experts' manual inputs and generating a bronze standard used for the reference. In the bronze standard case, we define a way of characterizing it using performances parameters. This allows obtaining a confidence interval of the reference and normalizing the subsequent algorithm analysis. Thereafter, we compare the behavior of algorithms to the reference using evaluation criteria defined by requirements. These criteria are an arrangement of metrics commonly found in the literature. Results are ultimately displayed in a radar style graph for performance analysis. These graphs offer a multi-criteria interpretation and are used to determine algorithms strengths and weaknesses. It is therefore possible to study many algorithms and compare them on a single graph that displays an easy to understand interpretation of algorithms. The developed platform is an interesting tool to not only pick the algorithm that best answers to specific requirements but also to have an idea on points that deserve to be improved on tested algorithms.

This platform has been set up for 2D / 3D segmentation algorithms and for 3D reconstruction meshs. In the results section, we show an algorithm evaluation for some methods developed in the laboratory.

Keywords: Evaluation platform, medical imaging algorithms, reference

TABLE DES MATIÈRES

Page

LISTE DES TABLEAUX

Page

LISTE DES FIGURES

Page

LISTE DES ABRÉVIATIONS, SIGLES ET ACRONYMES

2D	Bidimensionnel, 2 dimensions
3D	Tridimensionnel, 3 dimensions
ASD	Average Symmetric Distance, distance moyenne symétrique
Al	Algorithme
Bs	Étalon-bronze
CRAN	Centre de Recherche en Automatique de Nancy
CT	Computed Tomography
Dice	Indice de Dice
EOS	Système d'acquisition de radiographie biplane
Hauss	Distance de Haussdorff
IRM	Imagerie à Résonance Magnétique
JI	Indice de Jaccard
LIO	Laboratoire de recherche en Imagerie et Orthopédie
RMSD	Root Mean Square Distance, distance moyenne symétrique et quadratique
RSD	Relative Surface Difference, différence relative de surface
RVD	Relative Volume Difference, différence relative de volume
SD	Déviation Standard

INTRODUCTION

Le mémoire s'inscrit dans le contexte clinique d'identification automatisée de structures d'intérêt dans des images médicales. Ces images, obtenues à partir de diverses modalités d'imageries (CT-Scan, IRM, EOS^{TM} etc.), permettent de fournir des clichés anatomiques de patients. Ils peuvent se révéler essentiels au clinicien dans l'établissement d'un diagnostic. Ils fournissent également des informations précieuses quant à l'évolution et le stade des différentes pathologies. Comme mentionné par Udupa et al. (2006) dans son cadre de validation d'algorithmes, l'imagerie médicale présente des "scènes réelles", c'est-à-dire des images de patients subissant des protocoles d'imagerie médicale. Dans ce cas, il est impossible d'obtenir une modélisation absolument vraie de la scène évaluée. Le diagnostic est souvent laissé libre à l'interprétation du clinicien. La tâche de lecture d'images radiographiques est souvent longue et fastidieuse pour le clinicien. Certaines régions sont parfois difficiles à identifier et peuvent être une source d'importantes variations intra et inter observateurs. Par ailleurs, le nombre d'images augmentant de manière importante, il faut gérer beaucoup de données. Pour s'aider, le clinicien peut se servir d'algorithmes. Par contre, l'utilisation de ces algorithmes présente des défis techniques. Un des défis principaux de l'algorithmie est de pouvoir proposer une modélisation la plus proche possible de la réalité tout en diminuant de manière importante la durée d'acquisition, la pénibilité de lecture, la variabilité inter et intra opérateurs induite par une interprétation humaine et les erreurs causées par la répétition de tâches. De nombreux algorithmes sont donc développés chaque année dans le but de fournir des modélisations les plus proches possibles de la réalité. De manière à pouvoir être implantées dans des routines cliniques, ces méthodes nécessitent d'être testées, validées et comparées. Plusieurs plateformes et cadre de validation proposent des protocoles pour évaluer ces algorithmes mais aucun ne permet une évaluation quantifiée et intégrée des différentes méthodes de traitement d'images médicales. Dans ce mémoire, nous étudierons les techniques et méthodes permettant d'évaluer et de comparer les algorithmes de traitements d'images médicales. L'objectif principal est de proposer une plateforme d'éva-

luation des algorithmes de segmentation d'images médicales. Le mémoire se divise en cinq chapitres.

Le chapitre 1 présente la revue de la littérature et l'état de l'art des méthodes et techniques permettant de valider des algorithmes de segmentation d'images. Le chapitre 2 présente la problématique et les objectifs de recherche. Le chapitre 3 présente la méthodologie associée au projet de recherche. Dans un premier temps, nous détaillons l'intérêt de générer une référence d'étude pour fournir une évaluation objective. Dans un second temps, nous expliquons comment comparer les résultats obtenus des algorithmes de segmentation à cette référence d'étude en proposant des critères d'évaluation obtenus à partir de métriques de la littérature. Enfin, nous montrons comment interpréter et analyser les résultats des évaluations. Le chapitre 4 présente l'application spécifique de la plateforme à un cas d'étude 2D avec comme exemple la comparaison de deux algorithmes de segmentation de la tête fémorale à partir d'images EOS^{TM}. Cette étude a fait l'objet d'une présentation orale à la conférence IEE EMBC (cf article IV). Le chapitre 5 présente l'application spécifique de la plateforme à un cas d'étude 3D avec comme exemple l'analyse d'un algorithme de segmentation 3D de tumeurs cérébrales à partir d'IRM. Cette étude a été menée en collaboration avec le CRAN de Nancy.

CHAPITRE 1

REVUE DE LA LITTÉRATURE

Dans cette revue de littérature, nous présenterons d'abord les métriques utilisées pour évaluer les erreurs de distance faites lors de segmentations. Par la suite, nous présenterons des moyens de comparaison des segmentations. Enfin, nous détaillerons les cadres actuellement utilisés pour proposer des validations d'algorithmes et générer les références les plus précises possibles.

1.1 Préliminaires

Nous présentons dans cette section quelques rappels préliminaires et définitions mathématiques qui seront repris dans tout le rapport. Ceux-ci sont utiles à la compréhension complète du présent mémoire.

1.1.1 Rappels et notations

Dans tout le manuscrit, nous utiliserons les termes du traitement d'image et de l'imagerie médicale. Quelques rappels sur les éléments caractéristiques d'une image sont disponibles par la suite :

Image I en 2 dimensions
Une image en 2 dimensions est définie par une matrice I de taille $N = L \times C$, comprenant L lignes et C colonnes.
Chacune des N cases de cette matrice correspond à une unité minimale appelée pixel dont la valeur est codée suivant les normes de l'image (jpg, mask etc.) et permet de définir le contenu de l'image.

Image $I3$ en 3 dimensions
En 3 dimensions, le pixel défini précédemment est remplacé par un cube appelé voxel.

L'image 3D est donc définie par une matrice cubique $I3$ de taille $N = L \times C \times P$ où P représente la profondeur.
Un ensemble de voxels permet de définir une image 3D ou un volume V tel que présenté en figure 1.1.

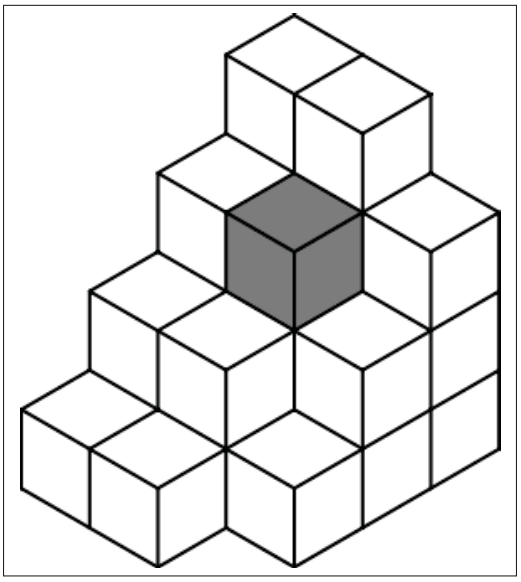

Figure 1.1 Ensemble de voxel d'une image 3D représentant un volume
Un des voxels est mis en évidence en gris

Dans le manuscrit, nous parlerons de segmentation 2D et 3D ainsi que de reconstruction.
Voici un bref rappel sur ces notions :

Segmentation 2D

La segmentation d'images 2D est un processus numérique de traitement d'images qui a pour but de partitionner l'ensemble des pixels de l'image en différents groupes. Chacun de ces groupes (ou régions d'intérêts) correspond à une caractéristique de l'image. D'après Udupa *et al.* (2006), elle permet deux tâches fondamentales :

1. Détection de structures dans les images ;
2. Identification précise de ces structures.

Par exemple, dans la figure 1.2, la segmentation 2D d'image sur une radiographie EOS^{TM} du membre inférieur (en vert) permet de détecter 2 structures d'intérêt. L'identification permettra ensuite d'identifier la structure supérieure comme étant un fémur et la structure inférieure comme un tibia.

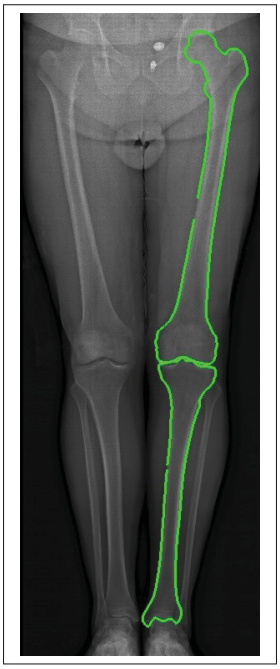

Figure 1.2 Identification du fémur et du tibia dans le membre inférieur sur une image EOS^{TM} en utilisant la segmentation 2D

Segmentation 3D

La segmentation d'images en 3D utilise sensiblement le même principe que la segmentation 2D sauf que l'on rajoute des informations de manière à obtenir la dimension de profondeur. Pour ce faire, plusieurs méthodes existent comme la segmentation d'images 2D selon l'axe de la troisième dimension désirée ou bien la segmentation dans un espace 3D. La mise en commun de ces ensembles d'images à l'aide d'algorithmes d'a priori mathématiques permet d'obtenir un volume 3D de voxels. La figure 1.3 (fig. de gauche) présente un exemple de modèle 3D obtenu par segmentation 3D du foie, proposée par Chartrand (2016).

Reconstruction 3D

Dans ce mémoire, la reconstruction 3D d'images consiste à générer un modèle 3D à partir d'un ensemble d'images prises sous différents angles de vue. Une de ces techniques est visible en figure 1.3 (fig. de droite). Pour plus d'informations sur les techniques de reconstruction 3D utilisées en imagerie médicale dans le cadre de notre étude, se référer à la thèse de Chaibi (2010).

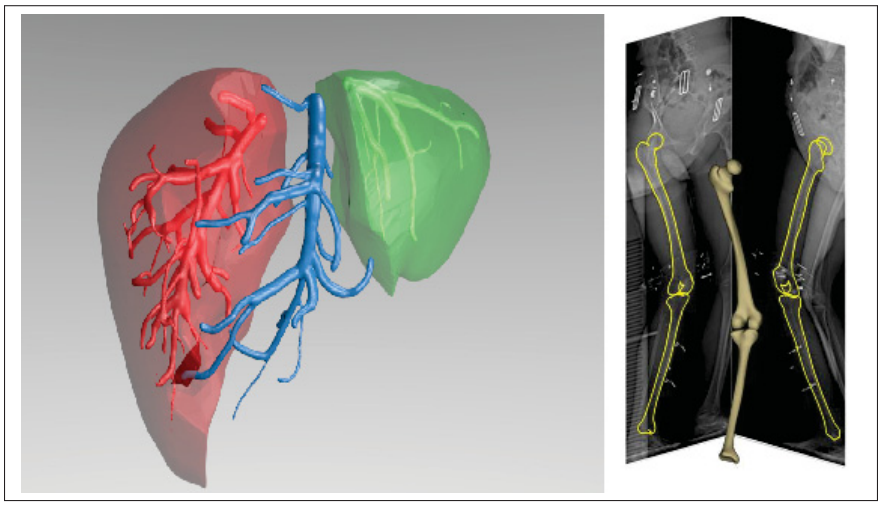

Figure 1.3 Exemple de segmentation (gauche) et reconstruction 3D (droite)
Gauche : Modèle 3D obtenu par segmentation 3D du foie et de ses structures particulières
Tirée de Chartrand (2016)
Droite : Reconstruction d'un modèle 3D
Tirée de Chaibi (2010)

1.1.2 Définitions mathématiques

Nous allons définir les expressions mathématiques suivantes. Celles-ci seront réutilisées dans tout ce chapitre ainsi que dans le chapitre 3.

- U V, formes géométriques définies par l'ensemble de pixels ou voxels $S(U)$, resp. $S(V)$;
- $|S(U)|$, le nombre de pixels ou voxels contenus dans $S(U)$;
- $|U|$, l'aire ou le volume défini par le modèle $S(U)$;
- $d^n(p_i, S(U)) = (\min_{\omega \in S(U)} \|p_i - \omega\|)^n$, la distance n-ième d'un pixel ou voxel p_i à son plus proche correspondant le pixel, resp. voxel, ω de $S(U)$;
- $D^n_U(U,V) = \sum_{p_i \in S(U)} d^n(p_i, S(V))$, la somme des distances n-ième de U à V.

1.2 Métriques d'analyse de performances d'algorithmes

La littérature disponible dans le domaine du traitement d'images utilise de nombreuses métriques permettant de calculer les performances des algorithmes de segmentation d'images médicales. Ces métriques permettent de quantifier les erreurs entre deux modélisations en terme de distance entre des surfaces ou des volumes. Les métriques sont généralement symétriques, ce qui signifie qu'elles sont calculées en prenant en compte comme référence chacun des deux modèles d'analyse ; de manière à mettre en évidence les différences locales de chacun des deux modèles.

1.2.1 Métriques de distance

De nombreuses métriques permettent de calculer la distance entre deux formes (surfaces ou volumes). Ces métriques sont basées sur l'analyse d'ensembles de points (ensembles de pixels pour les surfaces ou ensembles de voxels pour les volumes) définissant chaque forme. Elles sont disponibles dans Heimann *et al.* (2009). En utilisant les expressions définies en section 1.2, on a :

Distance moyenne entre deux formes U et V

ASD : Average Symmetric Distance, détaillée dans Heimann *et al.* (2009)

Il s'agit de la distance moyenne pixel à pixel (resp. voxel à voxel pour les volumes) entre les deux formes. C'est la métrique de distance la plus répandue dans la littérature de segmentation.

$$ASD(U,V) = \frac{D_U^1(U,V) + D_V^1(V,U)}{|S(U)| + |S(V)|} \tag{1.1}$$

Distance moyenne quadratique entre deux formes U et V

RMSD : Root Mean Square Symmetric Distance, détaillée dans Heimann *et al.* (2009)

Il s'agit de calculer la distance moyenne point à point symétrique entre les deux formes, en l'élevant au carré de manière à discriminer d'avantage les points les plus éloignés entre les formes.

$$RMSD(U,V) = \sqrt{\frac{D_U^2(U,V) + D_V^2(V,U)}{|S(U)| + |S(V)|}} \tag{1.2}$$

ASD et RMSD sont très proches, mais l'élévation au carré du RMSD donne une information supplémentaire puisqu'il discrimine de manière plus importante les différences les plus significatives (cf Heimann *et al.* (2009)) : en effet, plus le ratio $\frac{ASD(U,V)}{RMSD(U,V)}$ est proche de 1, plus la distance d'erreur entre les deux formes est homogène. Ce ratio renseigne donc sur l'homogénéité de l'erreur en distance.

Distance maximale entre deux formes U et V

Hauss : Distance de Haussdorff, détaillée dans Huttenlocher *et al.* (1993)

Il s'agit ici de calculer la distance maximale symétrique entre les deux formes.

$$Hauss(U,V) = \max\left(\max_{m \in S(U)} (d(m,S(V)), \max_{n \in S(V)} (d(n,S(U))))\right) \tag{1.3}$$

avec, $\{m,n\}$ pixel ou voxel de S(U), resp. S(V)

Cette information importante permet de détecter la présence de valeurs aberrantes et surtout de les quantifier. Si la distance de Haussdorff est très importante par rapport à la distance moyenne symétrique définie en équation 1.1, cela signifie qu'il y a des valeurs aberrantes issues de la segmentation. Basé sur la distance de Hausdorff, nous définirons une métrique dans le chapitre 3 permettant d'avoir des informations sur les valeurs aberrantes.

Dans la même logique, il est possible de définir la distance minimale entre 2 ensembles de points (en remplaçant max par min dans 1.3). Il est également possible de définir une fonction de seuillage mettant en évidence les points supérieurs à une certaine distance δ :

$$f(\delta) = \{p_i \in S(U), d^1(p_i, S(V)) > \delta\} \tag{1.4}$$

Cette fonction sera utilisée au chapitre 3 pour définir l'ensemble des valeurs aberrantes.

1.2.2 Métriques d'évaluation de surfaces ou de volumes

Contrairement à la section précédente, nous utilisons ici des métriques sur des modèles générés à partir d'un ensemble de pixels ou voxels. Il faut notamment créer des surfaces fermées ou des volumes à partir d'ensembles de pixels ou voxels d'entrée. Cette section utilise l'intersection et l'union d'objets, illustré en figure 1.4.

Différence relative de volume ou de surface (relativement RVD ou RSD)
RVD : Relative Volume Difference (resp. **RSD** : Relative Surface Difference), détaillée dans Heimann *et al.* (2009)

$$RVD(U,V) = 100 \times \frac{|U - V|}{|V|} \tag{1.5}$$

Il s'agit de la seule métrique présentée dans cette revue qui ne soit pas symétrique. Cette métrique s'utilise en présence d'une référence pour calculer le pourcentage d'aire (ou volume) de

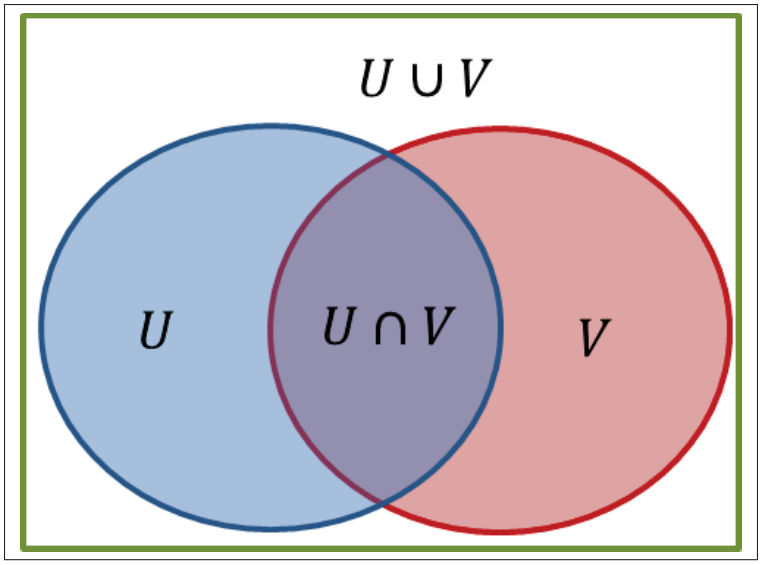

Figure 1.4 Intersection ($U \cap V$) et Union ($U \cup V$) illustrés à partir des deux modèles U et V

différence entre le modèle et la référence. Le résultat est signé (positif ou négatif). Cette métrique s'utilise sur des surfaces fermées. Elle exprime le pourcentage de sur ou sous segmentation du modèle par rapport à la référence. Il est recommandé de toujours l'utiliser couplée à une métrique symétrique.

Indice de similarité de Jaccard

JI : Jaccard Index, détaillé dans Akhondi-Asl & Warfield (2011)

$$JI(U,V) = 100 \times \frac{|U \cap V|}{|U \cup V|} \tag{1.6}$$

Il permet de calculer la similarité entre deux formes en déterminant la surface qu'elles ont en commun par rapport à leur surface totale. La figure 1.5 présente son utilisation sur deux volumes. Il est généralement couplé à la différence relative de volume ou de surface définie en équation 1.5. Il est souvent utilisé pour classer des algorithmes dans différents concours d'imagerie médicale, notamment dans les *grands concours* organisés chaque année [1].

1. Le site http://grand-challenge.org/all_challenges/ répertorie de nombreux concours d'imagerie médicale

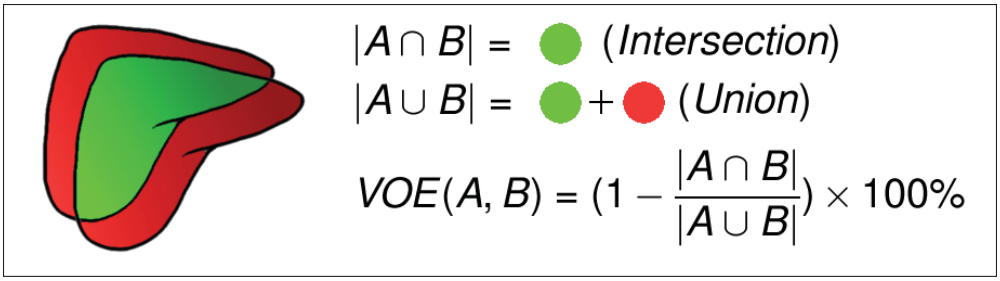

Figure 1.5 Présentation de l'indice de Jaccard en 3D, aussi appelé VOE (Volume Overlap Error)

Indice de similarité de Dice

Dice : Dice Index, détaillé dans Akhondi-Asl & Warfield (2011)

$$Dice(U,V) = 100 \times \frac{2 \times |U \cap V|}{|U| + |V|} \tag{1.7}$$

Se rapprochant de l'indice de Jaccard défini en équation 1.6, il s'agit également d'un indice de similarité entre deux objets et il est utilisé dans de nombreuses contributions scientifiques. La différence principale est que, contrairement à l'indice de Jaccard, il ne vérifie pas l'inégalité triangulaire stipulant que $|U+V| \leq |U| + |V|$. En effet, son dénominateur est supérieur ou égal à la somme des distance $|U+V|$, ce qui peut occasionner un risque de chevauchement des aires. Cela implique que l'indice de Dice va sur-estimer la vraie valeur de similarité. Il est donc intéressant pour comparer des résultats entre eux mais ne permet pas de valider un résultat de similarité.

1.2.3 Répartition des données

Cette section présente différentes méthodes permettant d'interpréter la répartition de données obtenues à partir d'un ensemble de pixels ou voxels. Elles vont donc permettre de définir s'il y a un comportement récurrent au niveau de la répartition de données.

Sensibilité et spécificité

Lorsqu'une décision binaire doit être effectuée (le pixel ou voxel appartient ou non au modèle

désiré), Warfield *et al.* (2004) définit des paramètres de performances binaires : la sensibilité (p_j) et la spécificité (q_j) pour la modélisation j. Pour cette partie, nous définissons :

- $i \in \{1,\ldots,N\}$ un pixel ou voxel d'une image de taille N ;
- $j \in \{1,\ldots,r\}$ une segmentation ;
- D_{ij}, la matrice de taille $N \times r$ de décision pour la segmentation j au pixel ou voxel i, c'est-à-dire la décision de chacun des experts ;
- T, le vecteur référence de taille N contenant pour chaque pixel i la décision binaire, représentant le vecteur de décision final, c'est-à-dire la segmentation supposée être la plus probable ;
- le paramètre de sensibilité p_j pour la segmentation j représente le pourcentage de vrai positif, c'est-à-dire la probabilité d'avoir une bonne décision sachant qu'elle l'est réellement ;

$$p_j = Pr(D_{ij} = 1 | T_i = 1) \tag{1.8}$$

- le paramètre de spécificité q_j pour la segmentation j représente le pourcentage de faux négatif, c'est-à-dire la probabilité d'avoir une décision négative sur un point n'appartenant pas au contour.

$$q_j = Pr(D_{ij} = 0 | T_i = 0) \tag{1.9}$$

Coefficient de fiabilité (R)

Il permet de déterminer la variabilité relative de l'erreur par rapport à la variabilité globale des données

$$R = \frac{\sigma_t^2}{\sigma_t^2 + \sigma_e^2} \tag{1.10}$$

Avec,

σ_t, écart type d'un lot de données

σ_e, écart type de l'erreur lié à ce lot de données

Weir (2005) utilise le coefficient de fiabilité de manière à déterminer l'homogénéité d'un lot de données. Plus le coefficient est proche de 1, plus les données sont homogènes. Un coefficient de 1 signifie que l'ensemble des données a la même valeur.

Analyse de Bland & Altman (1986)

Elle permet une analyse de la concordance entre deux métriques différentes au cours de la même évaluation d'algorithmes. De ce fait, elle met en lumière un comportement récurent et pointe les biais de mesure. Elle est par exemple très utilisée pour déterminer des erreurs de sur ou sous segmentation. La figure 1.6 montre un exemple d'analyse de Bland Altman pour comparer le comportement d'un algorithme par rapport à une référence et déterminer sa sensibilité à sur ou sous segmenter. Sur cette figure, nous avons décidé de tracer la différence d'aire en pourcentage entre le résultat obtenu par l'algorithme et une référence d'étude en fonction de l'aire moyenne des formes finales (en mm^2). La droite bleue représente la moyenne de différence d'aire pour l'algorithme. Ici, on voit que l'algorithme a tendance à avoir une aire de segmentation 10% supérieure à l'aire de la référence, ce qui indique qu'il effectue une sur segmentation. Les droites rouges pointillées représentent l'intervalle dans lequel 95% des données de l'algorithme se retrouvent, en supposant leur distribution normale. Plus cet intervalle est important, plus la répartition des données est hétérogène.

1.2.4 Scores de classification d'algorithmes

De manière à pouvoir classer différents algorithmes, de nombreux concours (ISBI, MICCAI, Kaggle etc. [1]) sont régulièrement organisés. Ils permettent aux différentes unités de recherche de se positionner par rapport à ce qui se fait en terme d'algorithmie ailleurs dans le monde ainsi que d'avoir accès à des bases de données communes pour se comparer. En ce sens, de nombreuses méthodologies sont mises en place pour proposer un classement le plus objectif possible. Lors d'un concours sur la segmentation 3D du foie, Heimann *et al.* (2009) proposait des critères fournissant un score global ϕ sur 100 pour classer les différents algorithmes en compétition. Pour ce faire, il a utilisé 5 métriques (présentées en section 1.2) :

1. L'ensemble de ces concours est disponible à l'url https://grand-challenge.org/All_Challenges/

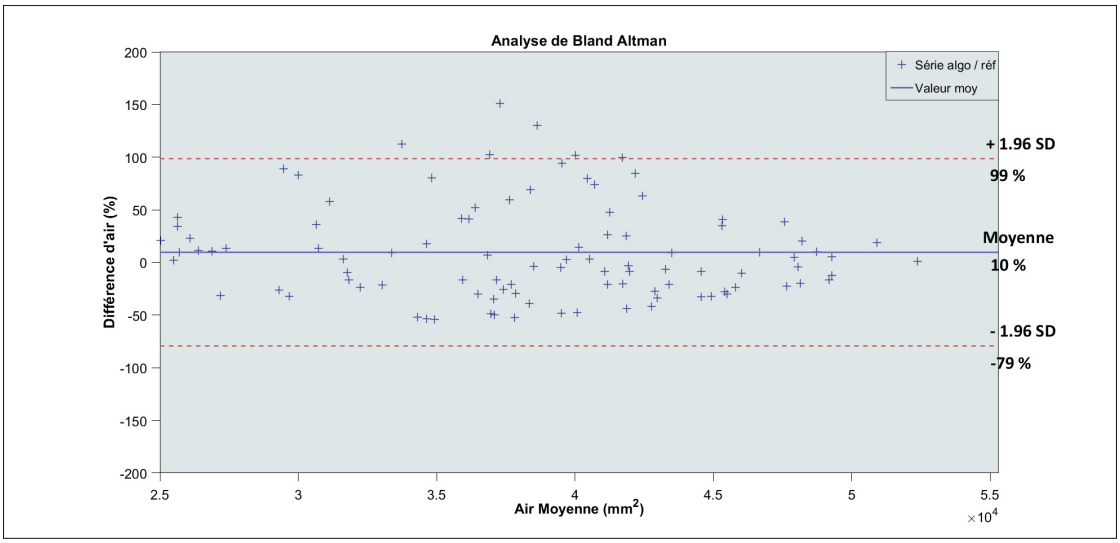

Figure 1.6 Exemple d'un graphique de Bland & Altman (1986) comparant un algorithme de segmentation d'images à une référence. Comparaison entre la différence d'aire (en %) et l'aire moyenne (en mm^2)

1. L'erreur relative de volume (équation 1.5) ;
2. La similarité en utilisant l'indice de Jaccard (équation 1.6) ;
3. La distance moyenne (équation 1.1) ;
4. La distance moyenne quadratique (équation 1.2) ;
5. L'erreur maximale ou distance de Hausdorff (équation 1.3).

Ces métriques sont utilisées pour comparer les sorties des algorithmes en compétition par rapport à une référence de type étalon-or ("gold standard" au sens de Chalana & Kim (1997)), simulée de manière synthétique. L'objectif de ces 5 métriques est de définir un score $\phi_i, i \in \{1,\ldots,5\}$ sur 100. Ce score est calibré de manière à comparer le comportement de l'algorithme par rapport à la segmentation manuelle d'un expert clinique. Pour chacune des métriques, la valeur de l'algorithme est obtenue et elle est notée ε_i. Il est demandé à un expert clinique de manuellement segmenter l'image et la valeur des métriques de l'expert par rapport à la référence $\bar{\varepsilon}_i$ est également rapportée. Cela permet de calibrer le score final en se comparant à l'expert.

L'équation 1.11 montre l'obtention de ce score. Ici un score de 75 suppose que l'algorithme se comporte à peu près comme l'expert.

$$\phi_i = \max(100 - 25\frac{\varepsilon_i}{\bar{\varepsilon}_i}, 0) \tag{1.11}$$

$\bar{\varepsilon}_i$ représente la valeur de la métrique ϕ_i reportée par l'expert définissant un score de 75

L'ensemble de ces scores est ensuite moyenné (en utilisant l'équation 1.12) de manière à fournir un score global ϕ sur 100.

$$\phi = \frac{1}{N}\sum_{i=1}^{N}\phi_i, \text{ ici } N = 5 \tag{1.12}$$

Nota : En fonction des besoins, il est possible de pondérer la moyenne si l'on désire considérer d'avantage l'impact de certaines métriques :

$$\phi = \frac{\sum_{i=1}^{N}\omega_i\phi_i}{\sum_{i=1}^{N}\omega_i}, \omega_i \in \{0\ldots1\} \tag{1.13}$$

Ce score est ensuite utilisé pour classer les algorithmes. Dans le cadre de la segmentation du foie, la figure 1.7 présente le classement de différents algorithmes étudiés. Bien que très utile puisqu'il permet de classer différents algorithmes entre eux et de déterminer ceux qui sont les meilleurs, ce type de classement présente certaines limites :

- les valeurs numériques sont difficiles à interpréter ;
- à travers ces scores, peu d'informations sont données sur les comportements généraux des algorithmes ;
- il est compliqué de pouvoir connaître les points forts et points faibles des différents algorithmes et leurs voies d'amélioration.

Method	Runtime	Overlap error		Volume difference		Avg. distance		RMS distance		Max. distance		Final
	[min]	[%]	Score	[%]	Score	[mm]	Score	[mm]	Score	[mm]	Score	Score
Beichel *et al.* MBR (*high*)	36	5.2 ± 0.9	80	1.0 ± 1.7	91	0.8 ± 0.2	80	1.4 ± 0.4	80	15.7 ± 3.5	79	82 ± 2
Beck and Aurich (*high*)	7	6.6 ± 1.6	74	1.8 ± 2.5	88	1.0 ± 0.3	74	1.9 ± 0.4	73	18.5 ± 4.1	76	77 ± 4
Dawant *et al.* (*med*)	20	7.2 ± 1.2	72	2.5 ± 2.3	86	1.1 ± 0.2	73	1.9 ± 0.5	74	17.1 ± 5.4	77	76 ± 5
Second rater		6.4 ± 1.0	75	4.7 ± 1.8	75	1.0 ± 0.2	75	1.8 ± 0.5	75	19.3 ± 5.6	75	75 ± 4
Lee *et al.* (*low*)	7	6.9 ± 1.4	73	1.3 ± 2.9	88	1.1 ± 0.3	73	2.1 ± 0.5	71	21.3 ± 4.0	72	75 ± 5
Beichel *et al.* CBR (*med*)	31	6.5 ± 1.1	74	1.1 ± 1.9	90	1.1 ± 0.4	72	2.5 ± 1.2	66	23.4 ± 10.5	69	74 ± 9
Wimmer *et al.* (*med*)	4–7	8.1 ± 1.1	68	6.1 ± 2.6	68	1.3 ± 0.2	67	2.2 ± 0.4	69	18.7 ± 4.6	75	69 ± 5
Slagmolen *et al.* (*med*)	60	10.4 ± 3.1	59	3.7 ± 6.2	70	2.0 ± 0.7	50	5.0 ± 2.4	34	40.5 ± 18.2	47	52 ± 19
Beichel *et al.* GC (*low*)	30	14.3 ± 9.4	48	3.1 ± 10.7	62	3.6 ± 3.1	34	7.9 ± 5.9	24	49.2 ± 20.4	38	41 ± 27

Figure 1.7 Comparaison d'algorithmes
Les algorithmes sont testés par rapport à une référence de type étalon-or en utilisant les scores présentés ; cette méthode permet de classer les algorithmes
Tirée de Heimann *et al.* (2009)

1.3 Importance de la référence

Dans cette section, nous présenterons l'importance d'avoir une référence d'étude pour la validation d'algorithmes. Nous verrons que plusieurs études ont été menées pour essayer de fournir la référence se rapprochant le plus possible de la réalité, ceci dans le but d'avoir la validation d'algorithmes la plus objective possible.

1.3.1 Intérêt d'une référence pour valider l'analyse

Pour pouvoir valider les comportements d'algorithmes, ceux-ci sont comparés à une référence considérée comme étant le modèle à obtenir. Bien souvent, comme c'est le cas, par exemple, chez Sun *et al.* (2012), Chav *et al.* (2009) ou encore Chen *et al.* (2014), la référence est obtenue par un seul expert clinique qui segmente manuellement le modèle final à obtenir. Cela signifie que la référence de l'expert est considérée comme étant l'étalon-or de l'analyse. Or, parfois, il est possible que plusieurs experts soient en désaccord entre eux. Dans ce sens, de nombreux algorithmes (comme l'algorithme STAPLE, présenté en section 1.3.2) permettent d'avoir une idée sur les zones d'accord et de désaccord des experts.

1.3.2 Algorithme STAPLE de Warfield *et al.* (2004)

Très utilisé en imagerie médicale, l'algorithme STAPLE permet la création d'une référence à partir de plusieurs segmentations d'entrée. Ces segmentations sont représentées par des masques de décisions binaires de l'image à analyser : la valeur 1 représente un pixel 2D ou un voxel 3D considéré comme appartenant au contour de l'objet d'étude, la valeur 0 appartenant au reste de l'image. Il permet de générer une "réalité terrain", c'est à dire de produire une carte de probabilités où chaque pixel (resp. voxel) de l'image finale est associé à une probabilité d'être ou non sur le contour. La figure 1.8 détaille le principe de l'algorithme pour l'obtention de la réalité terrain. Il fournit également un rapport des performances des experts en donnant leurs scores de sensibilité et de spécificité (voir section 1.2.3 pour plus de détails sur ces paramètres). Pour ce faire, nous utilisons les définitions de la section 1.2.3 en utilisant j segmentations manuelles de chacune des images, fournissant ainsi une matrice de décision D par image. La carte de probabilité obtenue est représentée par les paramètres de performance permettant d'estimer la matrice T. Cela donne les paramètres de performance de sensibilité p_j et spécificité q_j pour chacun des experts. Cet algorithme a été le sujet de nombreuses améliorations dont la possibilité de faire varier spatialement la performance des paramètres, par Commowick *et al.* (2012). Dans ce mémoire, nous utiliserons la version développée par Warfield *et al.* (2004).

Une illustration de l'algorithme STAPLE pour des images synthétisées est disponible en figure 1.9. Pour cette exemple, sept segmentations binaires d'une forme ont été effectuées. Il fallait remplir en noir (ici la valeur 1) la forme désirée et laisser l'arrière plan en blanc (à valeur de 0). L'algorithme STAPLE fournit ensuite une carte de la forme la plus probable et les valeurs des probabilités par pixel qui lui sont associées. Sont également fournis les paramètres de spécificité et de sensibilité associés à chacune des segmentations.

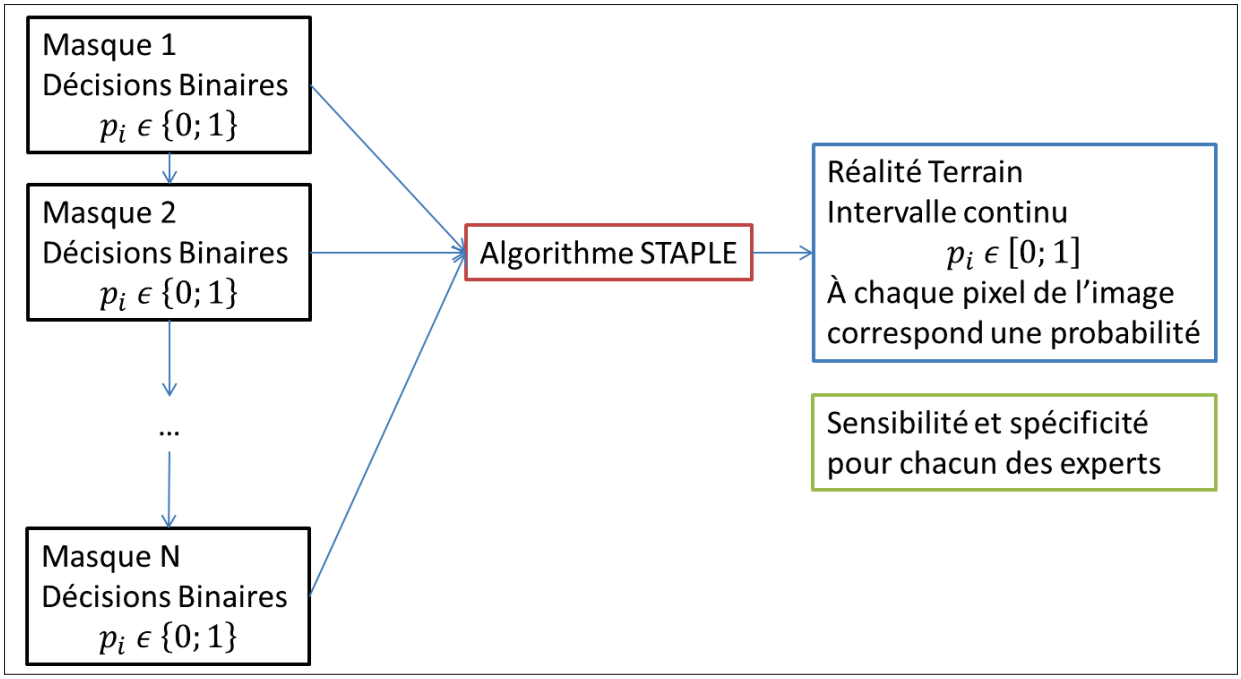

Figure 1.8 L'algorithme STAPLE prend en entrée plusieurs références d'experts. Celles-ci présentent des décisions binaires (1 pour un voxel situé sur le contour, 0 sinon) En sortie, il fournit une carte de réalité terrain comprenant, pour chaque pixel, la probabilité qui lui est associée. Il fournit également des métriques de spécificité et sensibilité pour chacun des experts

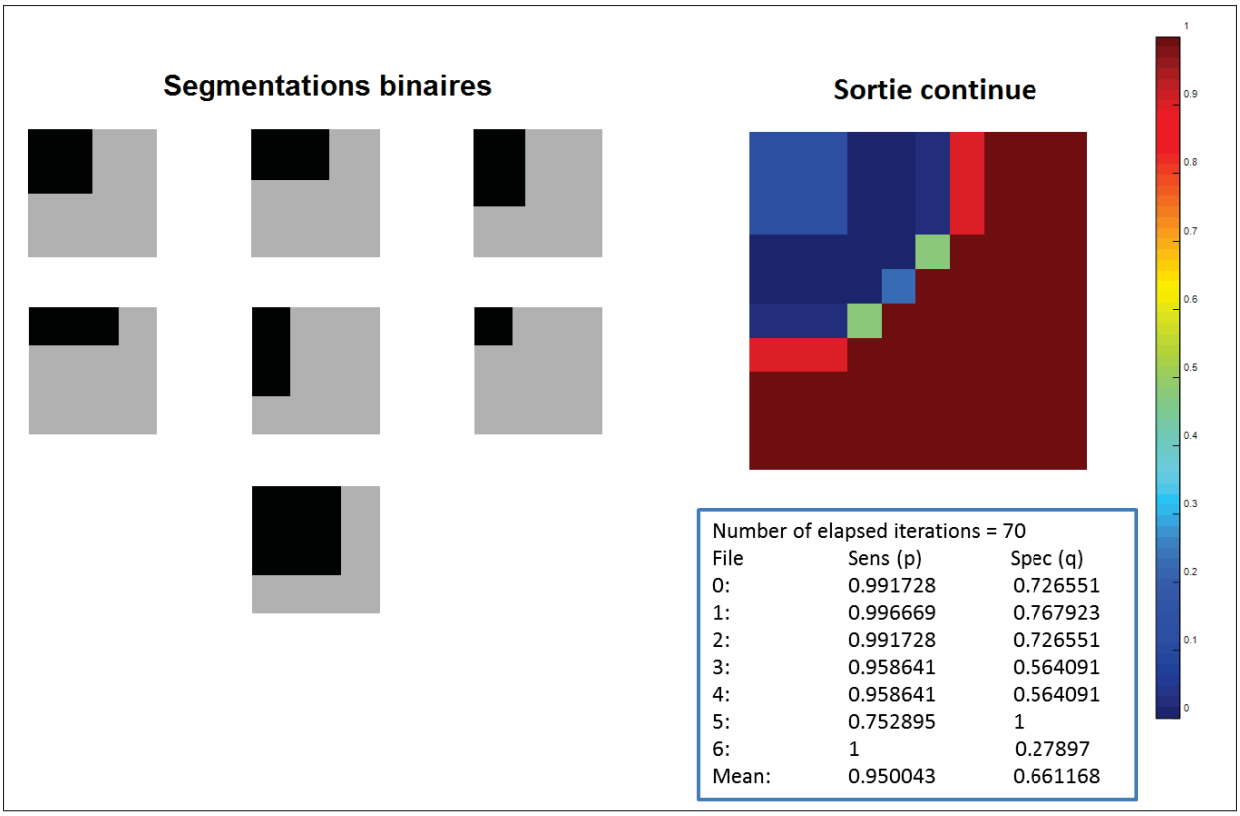

Figure 1.9 Sept segmentations binaires sont présentées (gauche).
L'algorithme STAPLE permet de fournir en sortie une carte de probabilités (droite).
À chacun des pixels est attribuée une densité de probabilité normalisée dont la valeur est visible ici grâce au code de couleur.
L'algorithme fournit également des paramètres de performances associés à chacune des segmentations (encadré bleu)

1.4 Cadres de validation disponibles

Des cadres de validation sont disponibles dans la littérature du traitement d'images ou de l'imagerie médicale. Le précurseur du domaine en imagerie médicale est Udupa *et al.* (2006) avec un cadre de validation. L'intérêt est de proposer un processus de validation standardisé qui puisse s'adapter au maximum de cas d'études. À terme, l'idée serait de faciliter la comparaison d'algorithmes.

1.4.1 Cadre de Udupa *et al.* (2006)

Dans son cadre de validation, Udupa *et al.* (2006) détaille les étapes nécessaires à la mise en place d'un protocole standardisé de validation en imagerie médicale :

Générer une référence
Dans le cas de scènes réelles, il faut proposer une réalité terrain qui soit le plus proche possible de la réalité. Pour ce faire Udupa *et al.* (2006) propose trois solutions :

- contourage manuel
  il s'agit de demander à des experts de contourer manuellement les régions définissant la référence et d'en générer une synthèse qui soit la référence d'étude ;
- fantôme mathématique
  il s'agit de modéliser une référence à partir d'une équation mathématique dont on connaît les paramètres. L'avantage de cette méthode est que l'on possède toutes les informations nécessaires sur la référence. L'inconvénient est qu'on génère une référence de manière mathématique, il ne s'agit donc pas d'une scène réelle ;
- scènes simulées
  il s'agit de partir de fantômes mathématiques tels que présentés précédemment et de définir des scènes en utilisant par exemple des a priori probabilistiques ou des notions d'apprentissage machine pour définir une scène. Comme précédemment, l'avantage est l'obtention de

nombreuses scènes avec toutes les informations de référence. L'inconvénient est que l'on ne peut pas savoir si ces scènes sont conformes à la réalité.

Définir des métriques de validation

En fonction des exigences du cahier des charges et de ce qui doit être testé dans l'algorithme, il faut définir des métriques de validation. Les métriques les plus courantes sont présentées dans la section 1.2.

Comparer les résultats en sortie d'un algorithme à la référence en utilisant les métriques

Il est possible par la suite de collecter tous les résultats obtenus par les métriques et de faire des analyses statistiques pour évaluer un comportement global (biais de sur ou sous segmentation, erreur de segmentation dans certaines régions etc.).

1.4.2 Génération d'étalon-bronze par Jannin *et al.* (2002)

Lorsqu'on se retrouve dans un cas comme l'algorithme STAPLE, la référence n'est plus obtenue par un expert mais générée par un algorithme à la suite d'un post traitement, Jannin *et al.* (2002) suggère de ne plus parler d'étalon-or mais bien d'étalon-bronze ("bronze standard"). Dans ce cas, il s'agit d'une estimation générée. L'intérêt est alors de caractériser la précision de la nouvelle référence, c'est-à-dire de déterminer pour chaque région la sensibilité de l'étalon-bronze. L'avantage de cette approche est que la référence est caractérisée et l'intervalle de confiance dans lequel elle peut être évaluée est défini.

1.4.3 Un système standardisé avec des critères de validation par Jannin *et al.* (2002)

En 2002, Jannin émet l'idée de proposer un protocole standardisé d'évaluation des algorithmes d'imagerie médicale. Selon lui, le problème est qu'il est impossible de comparer différents algorithmes qui ont des méthodes de validation complètement différentes. Pour ce faire, il propose des protocoles standardisés, qui utilisent des critères de validation qui évalueraient la qualité d'un algorithme sous certains critères d'intérêts :

Précision

Distance entre une mesure et la réalité.

Fiabilité

Précision d'un processus, c'est-à-dire la capacité de pouvoir répéter des mesures en obtenant le même résultat. Elle mesure l'homogénéité des données, c'est-à-dire la capacité de pouvoir prédire le comportement de l'erreur de l'algorithme. En effet, si la fiabilité est de 100%, alors l'erreur en précision de l'algorithme est la même pour toute l'image.

Robustesse

Capacité d'un algorithme à détecter la bonne région d'intérêt quelles que soient les contraintes présentes dans l'image à analyser. Une analyse de robustesse est complète lorsque l'on a défini une base de données de tests qui représente l'étude menée. On peut également parler de robustesse à certains facteurs comme la robustesse au bruit, au contraste etc.

D'autres critères peuvent être ajoutés suivant les exigences d'un cahier des charges, comme par exemple la réversibilité des processus, le temps d'exécution, la complexité de l'algorithme ou des exigences plus spécifiques à certains cas particuliers.

Il est par la suite nécessaire d'expliciter la production de ces critères de validation à l'aide de métriques qui fournissent des résultats quantitatifs, comme, par exemple, celles présentées en section 1.2. Si ces critères ont été mentionnés, ils n'ont jamais été définis de manière standardisée.

1.4.4 Cadre de Bayarri *et al.* (2007)

Bayarri *et al.* (2007) met en place un cadre de validation d'algorithmes informatiques censé répondre à la question *"Est-ce que le modèle informatique représente suffisamment la réalité ?"*. Son cadre s'appuie sur 6 étapes, sensiblement proches de celles de Udupa *et al.* (2006) :

1. Compréhension des données d'entrées et de leurs incertitudes ;

2. Détermination des critères d'évaluation ;
3. Collecte de données ;
4. Approximation du modèle ;
5. Analyse du modèle en sortie ;
6. Amélioration du modèle.

1.4.5 Cadre de Khooshabi (2013)

Le cadre développé propose une méthode pour générer un étalon-bronze dans le but de diminuer les erreurs inter et intra opérateurs. La solution développée se base sur l'algorithme STAPLE de Warfield *et al.* (2004). La figure 1.10 détaille le processus du cadre généré. Pour cette solution, Khooshabi (2013) développe une plateforme C++ qui utilise la librairie itk. C'est le premier cadre développé et disponible pour l'utilisation via une interface graphique, comprenant des cas d'exemples ainsi que des motivations concernant le choix des métriques. Par ailleurs, le cadre est par la suite testé sur des images synthétiques : rectangles, cercles et ellipsoïdes.

Il s'agit donc d'un travail d'intérêt puisqu'il présente, pour la première fois à notre connaissance, un outil développé utilisant un cadre de validation standardisé et présentant des applications sur des exemples simples. Les métriques utilisées sont la distance de Haussdorff (éq. 1.3), l'ASD (éq. 1.1), les indices de Jaccard (éq. 1.6) et de Dice (éq. 1.7) et leurs écarts type associés.

Le cadre de Khooshabi (2013) est un excellent travail préliminaire dans la proposition d'un cadre de validation d'algorithmes de segmentation d'images médicales. Toutefois, les applications présentées sont uniquement basées sur des formes géométriques simples et non des cas cliniques. Il n'y a également pas de lien entre l'étalon-bronze généré et le comportement de l'algorithme. Par ailleurs, aucune plateforme de visualisation n'a été développée.

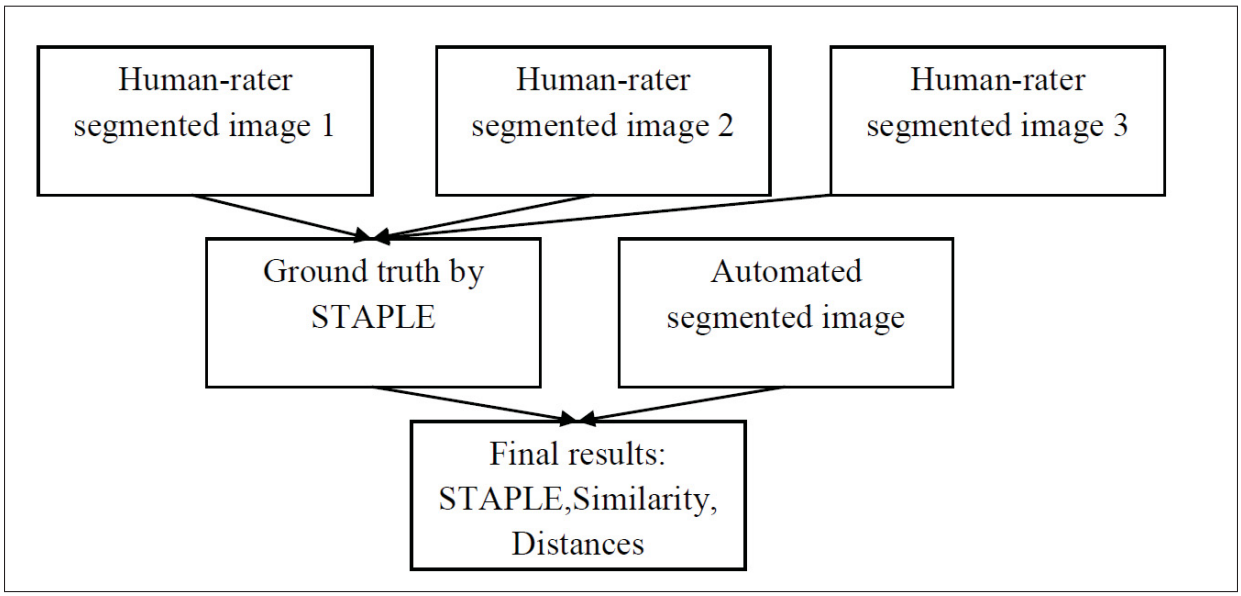

Figure 1.10 Présentation du cadre
Tirée de Khooshabi (2013)

1.4.6 Limitation des processus de validation d'algorithmes

Dans ce chapitre, nous avons présenté l'état de l'art des méthodes de validation d'algorithmes d'imagerie médicale actuellement disponibles. À ce jour, certains protocoles ont été proposés mais aucun ne présente de méthode standardisée permettant de traiter les cas de "scènes réelles". Les sections consacrées à la validation sont souvent composées d'une partie en fin d'article où certaines métriques de la littérature sont évaluées par rapport à une référence d'expert (Balestra *et al.* (2014), Chen *et al.* (2014), Chav *et al.* (2014)). Chaque méthode de validation ainsi que le choix des métriques est subjectif au choix des auteurs. Il devient donc difficile, voire impossible de pouvoir comparer les algorithmes entre eux. Dans ce sens, de nombreuses études proposent des méthodes pour mettre en place un protocole standardisé de validation. Par ailleurs, certains concours ont été mis en place pour classer les algorithmes les plus récents. Ces travaux sont intéressants, mais malheureusement très peu utilisés lors de la validation de nouveaux algorithmes. Par ailleurs, ils proposent souvent des tableaux de métriques avec des scores calibrés qui sont généralement compliqués à interpréter. En effet, bien qu'ils aient une utilité non négligeable puisqu'ils permettent de classer les algorithmes en fonction de leurs performances, ils ne renseignent pas sur les voies d'améliorations ainsi que les points forts et points faibles de ces algorithmes.

CHAPITRE 2

PROBLÉMATIQUE ET OBJECTIFS

2.1 Problématique

Les algorithmes de segmentation d'images médicales permettent la détection de structures anatomiques. Chaque année, nombre d'entre eux sont développés dans le but d'améliorer et d'automatiser cette détection. Pour valider ces algorithmes dans le cas de contexte clinique, ils doivent être appliqués sur des images de "scènes réelles" dont le comportement est comparé à une référence. Il s'agit très souvent de les comparer à une seule référence manuelle générée par un unique expert clinique (Sun *et al.* (2012), Chav *et al.* (2009) ou encore Chen *et al.* (2014)). Le fait de n'avoir qu'une seule référence faite par un expert peut biaiser la validation de l'algorithme puisqu'il est impossible de connaître l'intervalle de confiance de cette référence. Par ailleurs, chaque algorithme de segmentation est validé en utilisant ses propres critères de validation, généralement en calculant un lot de métriques définies, comme celles présentées en section 1.2. Il est ainsi difficile de pouvoir comparer différents algorithmes entre eux puisqu'aucune validation objective et commune n'existe. Malgré le fait que certains concours offrent des classements (Heimann *et al.* (2009), Bernard *et al.* (2014), Zheng *et al.* (2016)) entre différents algorithmes pour une problématique donnée, il est toutefois très difficile d'obtenir une comparaison objective des différents algorithmes développés. Par ailleurs, ces résultats ne permettent pas de proposer des voies d'amélioration. En effet, les algorithmes sont uniquement classés et les résultats de validation sont toujours présentés sous la forme d'un tableau comprenant de nombreuses valeurs numériques dont l'interprétation est souvent délicate. À partir de valeurs numériques seulement, il est également difficile d'avoir une vue d'ensemble du comportement global des algorithmes et de pouvoir en extraire leurs points forts et leurs points faibles.

2.2 Objectifs

L'objectif de ce mémoire est de proposer une plateforme d'évaluation et de comparaison multi-critères objective d'algorithmes de segmentation d'images médicales. L'idée est de pouvoir évaluer et comparer entre eux plusieurs critères relatifs aux algorithmes.
Cette plateforme est divisée en deux parties :

Génération de référence
Pour évaluer les algorithmes, il est important d'avoir une référence d'étude. Une référence peut être générée à partir d'images synthétiques dont on connaîtrait les paramètres. Dans ce cas, la référence est de type étalon-or. Si l'on veut se rapprocher d'un cas clinique en utilisant des images réelles, il faut alors générer la référence la plus proche possible de la réalité. Dans ce cas, nous demandons à plusieurs experts de segmenter manuellement les images. Nous appelons alors cette référence un étalon-bronze pour reprendre la définition de Jannin *et al.* (2002). Cette référence est non seulement générée mais également caractérisée pour chaque région anatomique : il s'agit donc d'analyser en tout endroit la précision qu'a l'étalon-bronze en fonction de l'interprétation des experts ; ceci de manière à normaliser l'analyse qui s'en suit.

Comparaison et évaluation des algorithmes
Les algorithmes de segmentation sont comparés en utilisant des critères d'évaluation fournissant un score de comparaison qui sont un agencement de métriques quantitatives de la littérature. L'ensemble de ces résultats est présenté dans un graphe de type radar pour rendre l'interprétation finale plus intuitive.

La figure 2.1 représente le diagramme de principe de la plateforme d'évaluation multi-critères.

Dans ce travail de maîtrise, nous utiliserons la plateforme pour évaluer dans un premier temps deux algorithmes de segmentation 2D de la tête fémorale à partir d'images EOS^{TM}. Par la suite, nous montrerons des applications de la plateforme à des algorithmes d'imagerie en 3D ; à savoir la segmentation 3D de tumeurs cérébrales à partir d'IRM.

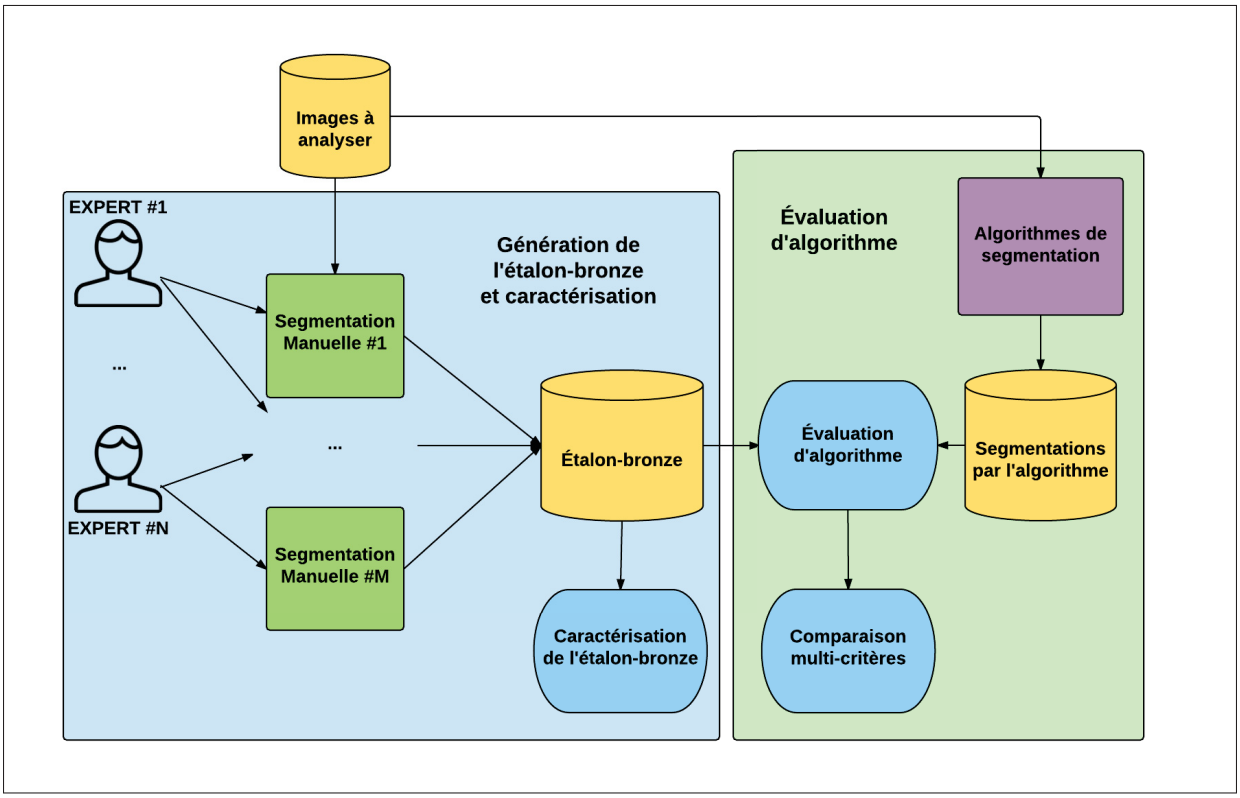

Figure 2.1 Diagramme représentant le principe de la plateforme d'évaluation multi-critères

CHAPITRE 3

MÉTHODOLOGIE GÉNÉRALE

3.1 Génération de référence

Si l'on reprend le cadre de Udupa *et al.* (2006), 2 cas d'études se présentent.
Dans le cas d'images synthétisées, la référence est alors un étalon-or puisqu'elle est générée de manière automatique, par exemple à l'aide d'équations mathématiques. Ces références d'études sont intéressantes puisque leur comportement est parfaitement connu et il n'est donc pas utile de les caractériser. Toutefois, elles ne représentent qu'en partie la réalité.

Dans le cas de contourages manuels, nous avons décidé de proposer une méthode pour générer une référence d'étude à partir de plusieurs interprétations d'experts. Par ailleurs, en plus de générer une référence, nous allons la caractériser : pour chacune des régions de la référence générée, l'intervalle de confiance des segmentations expertes est évalué et des paramètres sont extraits. Ceux-ci vont permettre de normaliser l'étude des algorithmes qui s'en suit. Par ailleurs, la référence étant générée par un algorithme tiers et régionalisée en fonction de sa précision, elle est définie étalon-bronze, comme mentionné par Jannin *et al.* (2002).

3.1.1 Algorithme STAPLE et *post-traitement* pour la construction de l'étalon-bronze

Pour générer un étalon-bronze, il faut utiliser plusieurs segmentations manuelles réalisées par des experts. Ces segmentations sont ensuite introduites en entrée de l'algorithme STAPLE de Warfield *et al.* (2004). Cet algorithme génère une carte de probabilités "réalité terrain" de la référence. Pour chaque pixel de la carte de probabilités générée en sortie de l'algorithme, une valeur normalisée indique la probabilité que ce pixel soit ou non situé sur la frontière de l'objet (pour plus d'informations sur l'algorithme STAPLE, se référer à la section 1.3, sous-section 1.3.2 du rapport). À partir de la carte de probabilités, nous générons un étalon-bronze en prenant le chemin continu parcouru par le maximum de probabilité. Pour une région donnée, la distribution de probabilité normalisée est tracée. La région contenant le maximum de probabi-

lité est retenue comme étant l'étalon-bronze pour cette région. Par la suite, des paramètres sont extraits de manière à caractériser cet étalon-bronze (voir section 3.1.2).

3.1.2 Caractérisation de l'étalon-bronze

En plus d'être généré, l'étalon-bronze est caractérisé. Une analyse de la carte de probabilités permet de déterminer un intervalle de confiance relatif aux segmentations manuelles d'entrée. Pour chacune des régions de la référence, des paramètres de précision sont obtenues de manière à déterminer l'intervalle de confiance des experts à ces endroits. Ces paramètres vont permettre par la suite de normaliser l'évaluation d'algorithmes en fonction de l'intervalle de confiance de la référence. Cela permet d'éviter d'obtenir des références qui seraient biaisées.
Les 3 paramètres extraits de l'analyse sont les suivants :

- ν_{Bs} : limite de précision
  taille de la région où la probabilité normalisée est supérieur à 68% (soit 2σ d'une loi normale) ;
- σ_{Bs} : écart type de ν_{Bs}, prise sur toutes les régions d'étude ;
- $SD(|Bs|)$: écart type de l'aire de l'étalon-bronze, pris sur les segmentations d'entrée de l'algorithme STAPLE (définition de l'aire d'un objet en section 1.1.2).

3.2 Évaluation multi-critères

À partir de la référence (ou bien étalon-or, ou bien étalon-bronze généré et caractérisé), nous allons maintenant proposer un cadre pour évaluer les algorithmes. L'idée est de pouvoir proposer une méthodologie qui va permettre non seulement de valider les algorithmes mais également de les évaluer et les comparer entre eux. L'intérêt est de comprendre les forces et les faiblesses d'un algorithme, ce qui rendra son amélioration d'autant plus simple. L'idée d'une telle plateforme est, à terme, de pouvoir proposer un outil simple, rapide et facile à interpréter pour se comparer à ce qui se fait déjà dans la littérature lors de la production d'un nouvel algorithme.

3.2.1 Mise en place de critères d'évaluation

Dans ce travail de recherche, nous avons défini des critères d'évaluation permettant de peaufiner l'analyse des algorithmes en apportant un score pour l'évaluation. L'objectif d'un critère d'évaluation est de pouvoir analyser le comportement d'un algorithme en apportant une approche quantitative et en fournissant un score défini à partir de métriques que l'on peut trouver communément dans la littérature. Un critère est défini de la manière suivante :

- il s'agit d'un score obtenu à partir de métriques quantitatives ;
- il utilise des métriques qui sont normalisées par les paramètres de l'intervalle de confiance de la référence d'étude (cf section 3.1.2) ;
- le score fourni est sur 100 ;
- un score de 100 ou supérieur à 100 signifie que l'algorithme se comporte dans le même intervalle de confiance que la précision caractérisée de la référence. Si le score est supérieur à 100, cela veut dire que la différence algorithme / référence est inférieure à l'intervalle de confiance lui-même. Dans ce cas, le score final est maintenu à 100.

Dans le cadre de l'évaluation d'algorithmes de segmentation 2D de détection de la tête fémorale (que nous présentons en section 4), nous avons défini cinq critères à partir des métriques présentées dans le chapitre 1.

Précision (A)

Il s'agit de calculer la capacité de l'algorithme à se rapprocher de la forme d'intéret. Elle est définie en utilisant l'erreur moyenne symétrique entre la courbe générée par l'algorithme et la référence et est normalisée en utilisant la limite de précision v_{Bs} de la référence, définie en section 3.1.2.

$$A = 100 \times \frac{v_{Bs}}{ASD(Al)} \tag{3.1}$$

Homogénéité de l'erreur (R)

Il s'agit de définir si l'erreur entre la forme produite par l'algorithme et la référence est homogène, c'est-à-dire déterminer si l'algorithme a un comportement uniforme en terme d'erreur. Cela permet de détecter aisément le biais que pourrait avoir un algorithme. Le calcul est inspiré du coefficient de fiabilité de Weir (2005), en utilisant l'écart type de l'erreur générée par l'algorithme σ_{Al} normalisée par l'écart type de la précision de la référence σ_{Bs}.

$$R = 100 \times \frac{2 \times \sigma_{Bs}^2}{\sigma_{Bs}^2 + \sigma_{Al}^2} \tag{3.2}$$

Une illustration des comportements de précision et d'homogénéité de l'erreur telles que nous les entendons dans ce travail de maîtrise sont présentés en figure 3.1.

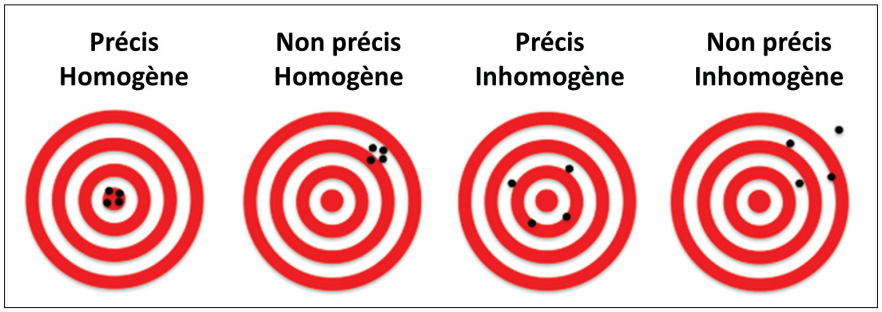

Figure 3.1 Illustration de la différence entre la précision et l'homogénéité ; extrait de Zheng *et al.* (2016)

Robustesse (Φ)

Il s'agit de déterminer la capacité d'un algorithme à identifier un objet d'intérêt. Il faut pour cela essayer de déterminer les métriques permettant au mieux de savoir si la forme finale obtenue par l'algorithme correspond à la référence. Pour ce faire, nous utilisons l'indice de similarité de Jaccard couplé à la différence relative de surface.

$$\Phi = JI(Al, Bs) \times (100 - |RSD(Al, Bs)|) \tag{3.3}$$

Dans le but d'avoir l'analyse de robustesse la plus complète possible, il est important d'avoir la base de données d'images à segmenter la plus représentative de l'étude à mener, avec si possible l'éventail des différents cas.

Sensibilité à la sur / sous segmentation (Θ)

Il s'agit de déterminer si un algorithme a une tendance globale à sur ou sous segmenter la forme d'intérêt. Pour se faire, nous calculons le nombre d'échantillons de notre base de données dont la différence d'aire avec la référence dépasse un critère de sensibilité. Par défaut, nous prenons ce critère égal à 1.96 fois l'écart type de l'aire de la référence ; soit l'intervalle où se situeraient 95% des segmentations expertes, en supposant qu'elles soient normalement distribuées (en référence à Bland & Altman (1986)).

$$\Theta(\alpha) = 100 \times \frac{Card(\{S(Al) : |\,|Al| - |Bs|\,| < \alpha\})}{Card(\{S(Al)\})} \tag{3.4}$$

$$\text{avec, } \alpha = 1.96 \times SD(|Bs|) \text{ par défaut}$$

Couplé à la robustesse Φ, il est possible d'obtenir des informations sur le comportement global de l'algorithme en permettant d'identifier si les erreurs faites par l'algorithme sont principalement des erreurs de sur ou sous segmentation ou bien des erreurs de segmentation globale.
Le tableau 3.1 présente les différents comportements d'algorithmes issus de l'analyse couplée des coefficients Φ et Θ.

Tableau 3.1 Informations globales sur le comportement des algorithmes en couplant le score de robustesse et celui de la sensibilité à la sur / sous segmentation

$\Theta(\alpha)$ \ Φ	**Faible**	**Important**
Faible	Tendance à mal segmenter	Tendance à sur / sous segmenter
Important	Tendance à générer un biais récurrent	Tendance à bien segmenter

Sensibilité aux valeurs aberrantes (Δ)

Il s'agit de déterminer si un paquet isolé de valeurs aberrantes a un impact important sur le comportement de l'algorithme. Pour ce faire, il faut dans un premier temps définir à partir de quelle limite une valeur est considérée aberrante. Pour cela, nous utilisons l'erreur moyenne symétrique quadratique *RMSD* définie en section 1.2. Par la suite, nous calculons l'impact de ces valeurs aberrantes en analysant le poids de la variance de la distance entre les valeurs aberrantes par rapport à la variance des distances totales. Ainsi, un paquet de valeurs aberrantes et voisines aura un impact bien plus négatif sur le comportement final de l'algorithme que des points isolés et mal placés.

$$\Delta = 100 \times (1 - \frac{\delta(O)}{\delta(Al)}) \tag{3.5}$$

$$\text{avec, } \begin{cases} \delta(U) = \sum\limits_{p_i \in U} \sum\limits_{\substack{p_j \in U \\ p_j \neq p_i}} \dfrac{d^2(p_i, S(Bs))}{d^2(p_i, p_j)} \\ \\ O = \{p \in Al, d(p, S(Bs)) > RMSD(Al)\} \end{cases}$$

La figure 3.2 présente une illustration de la sensibilité aux valeurs aberrantes Δ. Dans cet exemple, nous voyons que celle-ci augmente pour un paquet de valeurs aberrantes proches localement. Cela permet de détecter si un ensemble de données produites par l'algorithme va avoir un fort impact négatif sur les résultats. En analysant la cause de ce biais (par exemple une

mauvaise détection), cela permet généralement d'identifier une lacune dans le développement de la méthode.

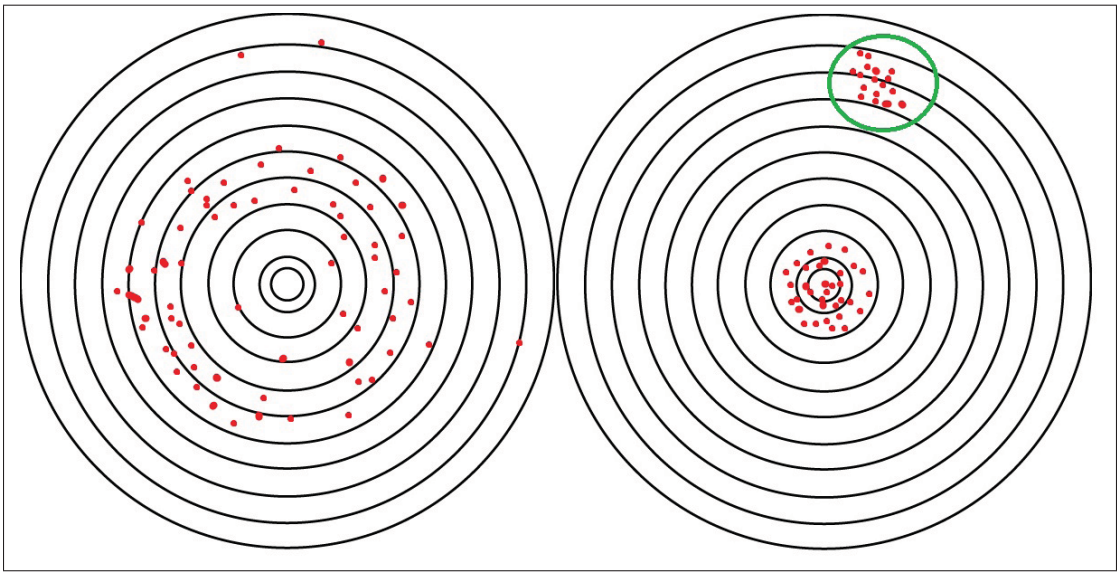

Figure 3.2 Présentation de la sensibilité aux valeurs aberrantes
Gauche : valeurs aberrantes discrètes qui ont un impact très faible sur la détection finale faite par l'algorithme.
Droite : ensemble de données (en vert) biaise de manière importante le comportement final de l'algorithme

Il est important de remarquer que ces critères ont été choisis puisqu'ils semblent être les plus appropriés à l'exemple choisi dans notre travail, soit analyse d'algorithmes de segmentation de la tête fémorale. D'autres critères pourraient être privilégiés en fonction de cahier des charges différents, comme par exemple la compléxité des algorithmes ou encore le temps d'éxécution. En fonction des exigences, il est également possible de fournir un score final et de pondérer certains critères en fonction de l'intérêt (en utilisant l'équation 1.13). Par exemple, utiliser d'abord un algorithme plus robuste pour initialiser la détection et le coupler ensuite avec un algorithme plus précis pour obtenir la forme finale.

3.2.2 Création d'un graphique radar pour la présentation des résultats

Dans la littérature, les résultats sont généralement présentés sous forme d'un tableau de valeurs numériques tel que c'est le cas dans le tableau 1.7 du chapitre 1. Dans ce cas, il peut être compliqué d'interpréter les données et de pouvoir comparer le comportement des algorithmes dans certaines situations. Le graphique radar (ou diagramme de Kiviat [1]) est une bonne alternative pour présenter les résultats. Il permet directement d'avoir un a priori sur les résultats. Les avantages du graphique radar sont les suivants :

- il permet de mettre instantanément en évidence les points forts et points faibles d'un algorithme ;
- il permet de comparer plusieurs algorithmes ensemble.

La figure 3.3 présente un cas où trois résultats de segmentation générés synthétiquement auraient le même score final en utilisant les critères du tableau 1.7 (combinaison ASD, RMSD, distance de Haussdorff, indice de Jaccard et différence relative en surface).Toutefois leur comportement est vraiment différent. Le graphique radar permet ici de mettre en évidence les points forts et points faibles de chacune des méthodes développées. Sur cette figure, nous avons synthétisé trois lots de données (rouge) par rapport à une référence (jaune). Une analyse de critères d'évaluation par le biais d'un graphe radar permet directement d'identifier les pistes d'améliorations de chacun des lots de données :

- le lot 1 présente un décalage (problème de recalage) entre le lot synthétisé et la référence. Il est donc très peu précis et fiable mais la forme finale est très proche de la forme référence ;
- le lot 2 semble globalement précis mais présente un paquet de données qui a un très mauvais impact sur la forme finale ;
- le lot 3 est peu précis mais se rapproche très fortement de la forme finale.

1. Plus d'informations sur le diagramme de Kiviat : https://fr.wikipedia.org/wiki/Diagramme_de_Kiviat

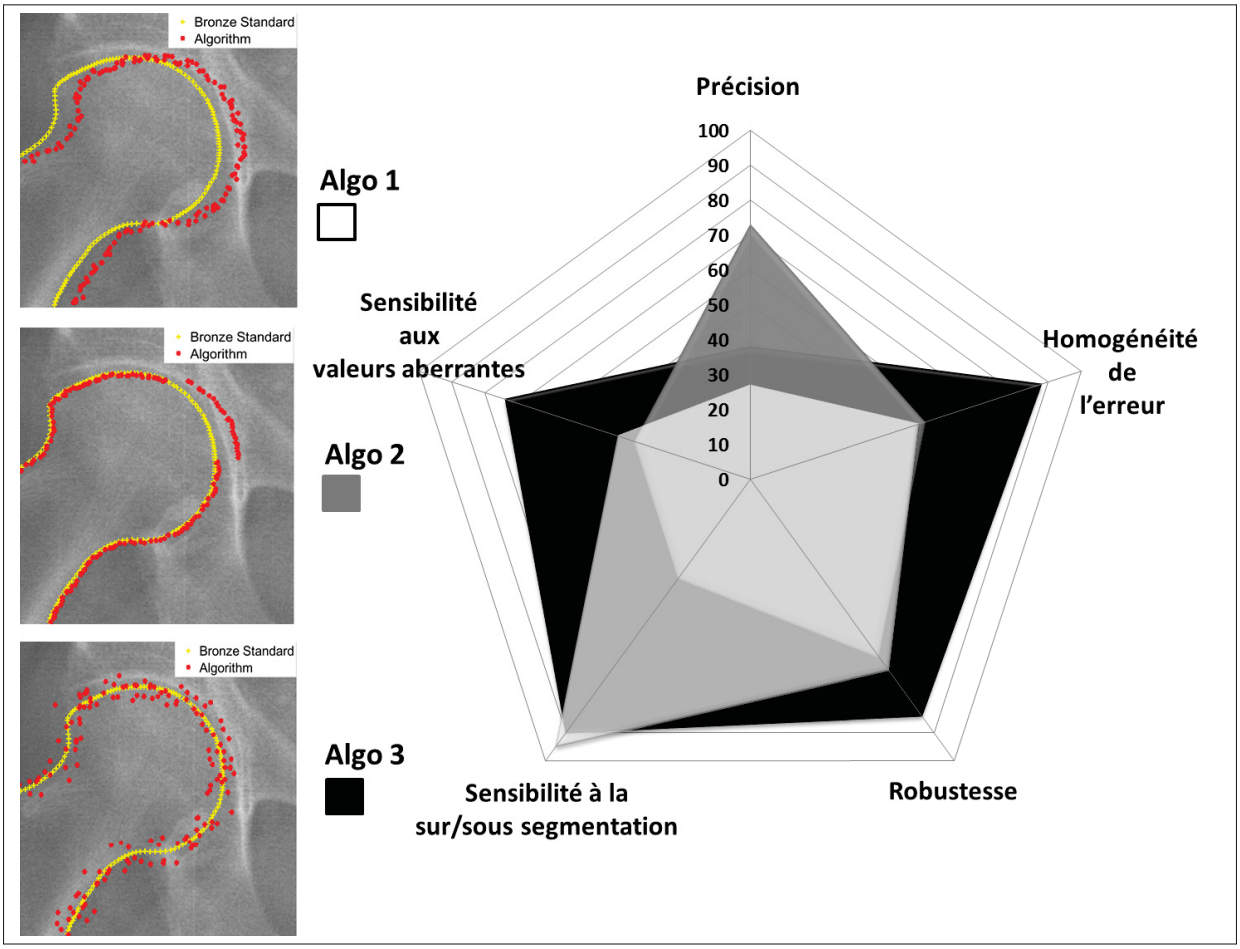

Figure 3.3 Trois lots de données synthétiques (rouge) par rapport à la référence (jaune) Les 3 algorithmes ont le même score final par rapport au tableau 1.7 (combinaison ASD, RMSD, indice de Jaccard, distance de Haussdorff et différence relative en surface) par rapport au tableau 1.7, pourtant les critères d'évaluation mettent en évidence un comportement bien différent

3.3 Programmation de l'interface 2D / 3D

3.3.1 Programmation logicielle pour l'étalon-bronze

Nous avons mis en place un programme développé en C++ avec la librairie open-source itk permettant de générer la référence. Sa caractérisation est faite à l'aide de MATLAB. Ce programme se veut modulaire et prend en entrée plusieurs types de données :

- des masques de segmentations binaires ;
- des contours continus ;
- des nuages de points discrets qui seront prolongés par interpolation linéaire.

Ce programme ne fonctionne pour l'instant que pour des générations de référence en 2D. Pour faire des références en 3D, il est toutefois possible de faire du tranche-par-tranche comme précisé en section 5.1.2.

Les diagrammes 3.4 et 3.5 présentent l'architecture logicielle utilisée pour mettre en place le programme de génération d'étalon-bronze.

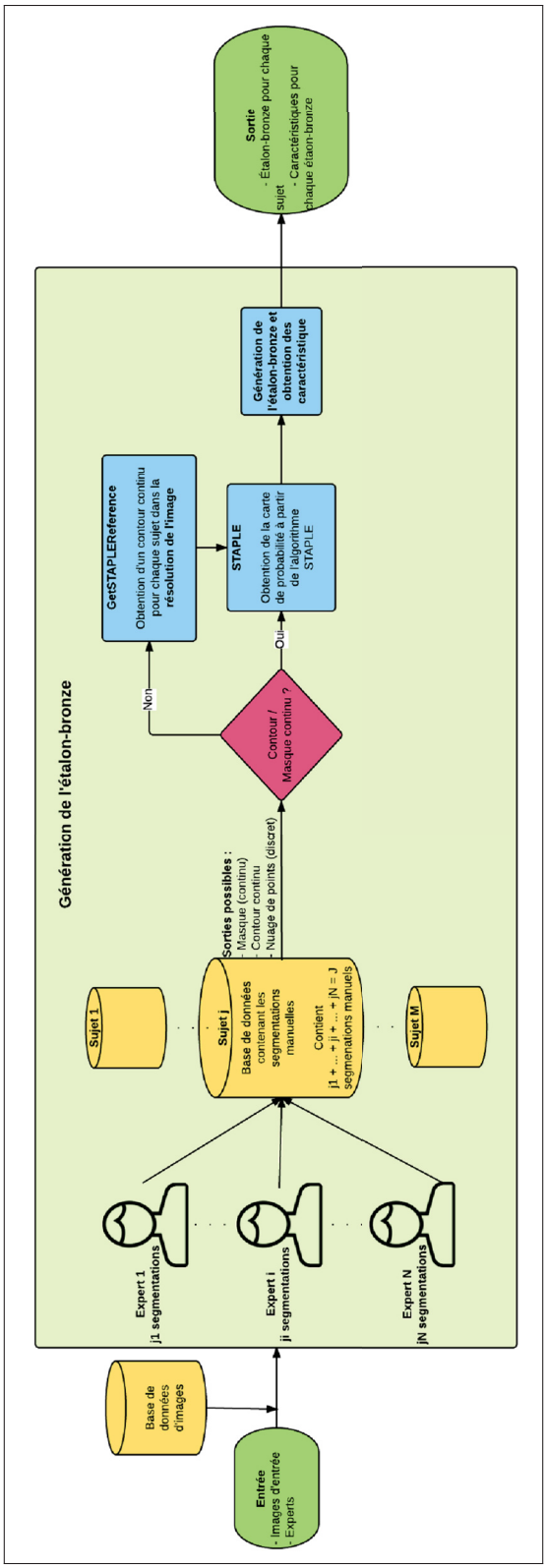

Figure 3.4 Diagramme explicatif du programme réalisé, permettant d'obtenir l'étalon-bronze

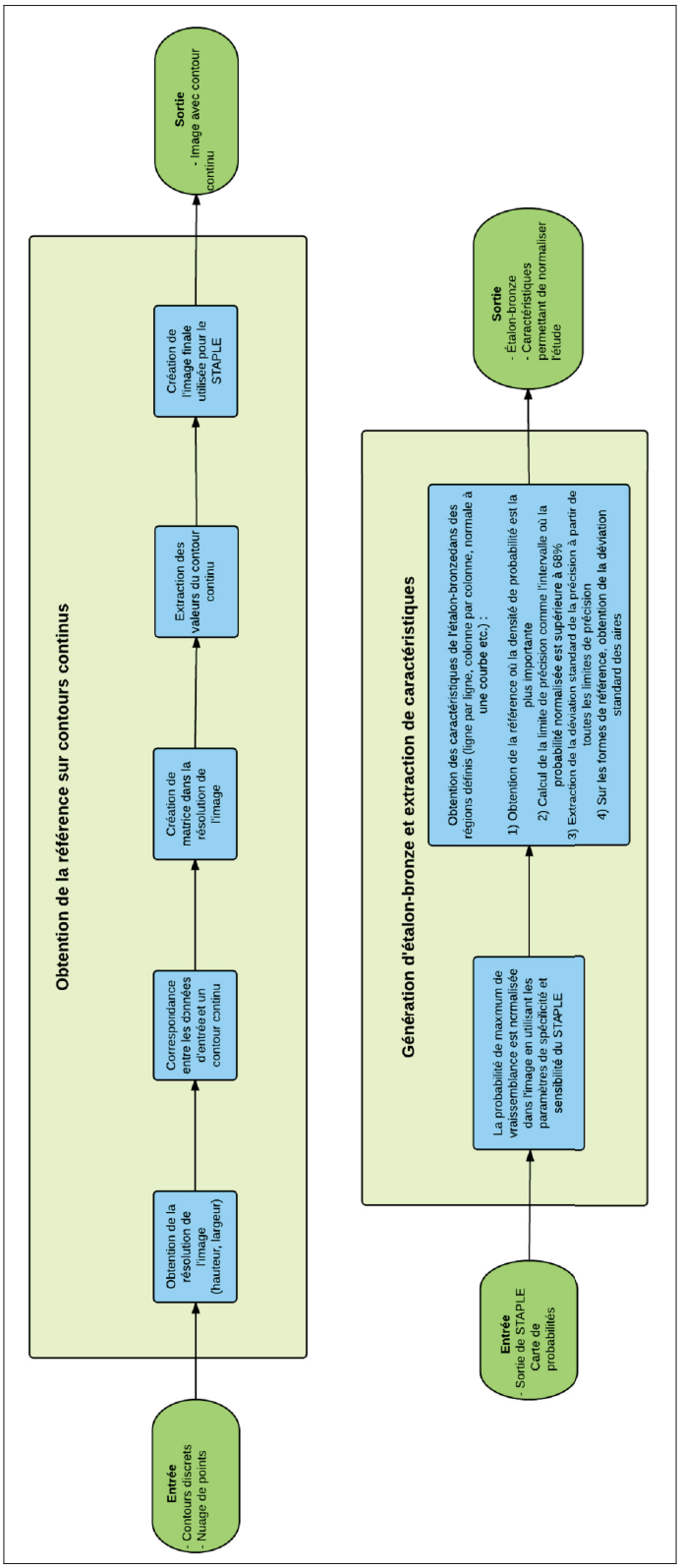

Figure 3.5 Détails de l'obtention de référence sur contours continus et de l'extraction de caractéristiques

3.3.2 Programmation logicielle pour la plateforme 2D / 3D

La plateforme 2D / 3D a été développée à l'aide de Matlab pour comparer un modèle 2D ou 3D à une référence, qui soit ou bien étalon-or ou bien étalon-bronze.
La plateforme 3D se veut le plus modulaire possible de manière à s'adapter au plus grand nombre de données d'entrée. En fonction des données d'entrée (resp. maillage ou volume 3D), celle-ci permet automatiquement de générer le modèle complémentaire correspondant (resp. volume de voxels ou maillage). Le diagramme 3.6 présente la manière d'obtenir les modèles d'entrée de la plateforme en fonction de la forme initiale. Dans le but d'automatiser les analyses, il est très facilement possible d'ajouter des scripts ou des plugins à la plateforme pour rapidement pouvoir avoir une idée du comportement global d'un algorithme.
Des scripts ont, par exemple, été mis en place pour homogénéiser les données lors de l'étude d'algorithmes qui ne renvoient pas les mêmes types de format de fichiers en sortie.

3.3.3 Analogie des métriques 2D / 3D utilisées

Dans le but d'adapter la plateforme à l'évaluation des algorithmes d'imagerie en 3D, nous utilisons les analogies présentées dans le tableau 3.2. Pour ce faire, nous utilisons les métriques présentées en chapitre 1 non plus sur des pixels mais sur des voxels.

Tableau 3.2 Analogie permettant d'adapter la plateforme à l'algorithmie 3D

Critères	**2D**	**3D**
Erreur entre modèles	Distance point à point	Distance point plan
Similarité	Jaccard 2D	Jaccard 3D
Chevauchement	RSD[1]	RVD[2]
Calcul de forme	Surface 2D	Volume 3D

1. RSD : Différence Relative de Surface (voir sec. 1.5)
2. RVD : Différence Relative de Volume (voir sec. 1.5)

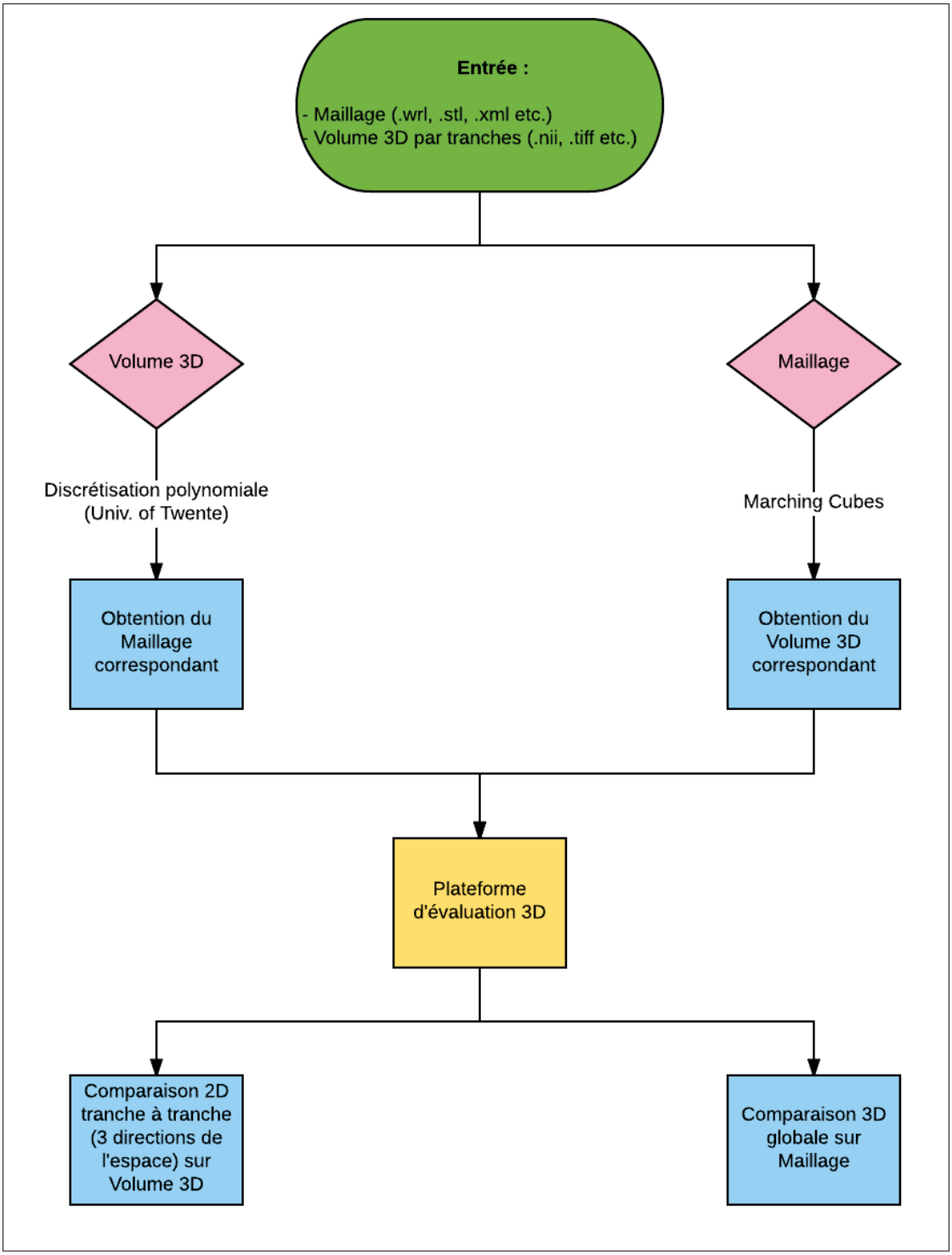

Figure 3.6 Principe de la plateforme 2D / 3D pour l'analyse d'un algorithme d'imagerie médicale

3.3.4 Interface graphique de la plateforme 3D

La figure 3.7 présente l'interface graphique mise en place pour le développement de la plateforme 3D. Dans cette analyse, nous présentons une évaluation d'algorithmes de reconstruction du fémur en fonction d'une référence. Plusieurs fonctionnalités ont été mises en place dans le but de rendre l'évaluation la plus complète possible :

1. Tout type de données peut être passé en entrée : que ce soit des maillages ou bien des volumes de voxels (pour plus d'informations, voir section 3.3.2) ;
2. La visualisation peut se faire tranche-par-tranche ;
3. L'analyse peut se faire suivant trois plans de l'espace (XoY, XoZ, YoZ) ;
4. La visualisation est également possible sur la figure 3D complète fournissant une carte d'erreurs et comparant le modèle obtenu par algorithme à la référence ;
5. L'extraction des métriques peut se faire sur les tranches 2D ou sur la figure 3D. Celle-ci s'adapte au 2D / 3D en fonction de ce qui est demandé, en suivant les analogies présentées au tableau 3.2 ;
6. Tous les outils graphiques Matlab sont également disponibles dans le but de rendre l'interprétation des figures la plus complète possible.

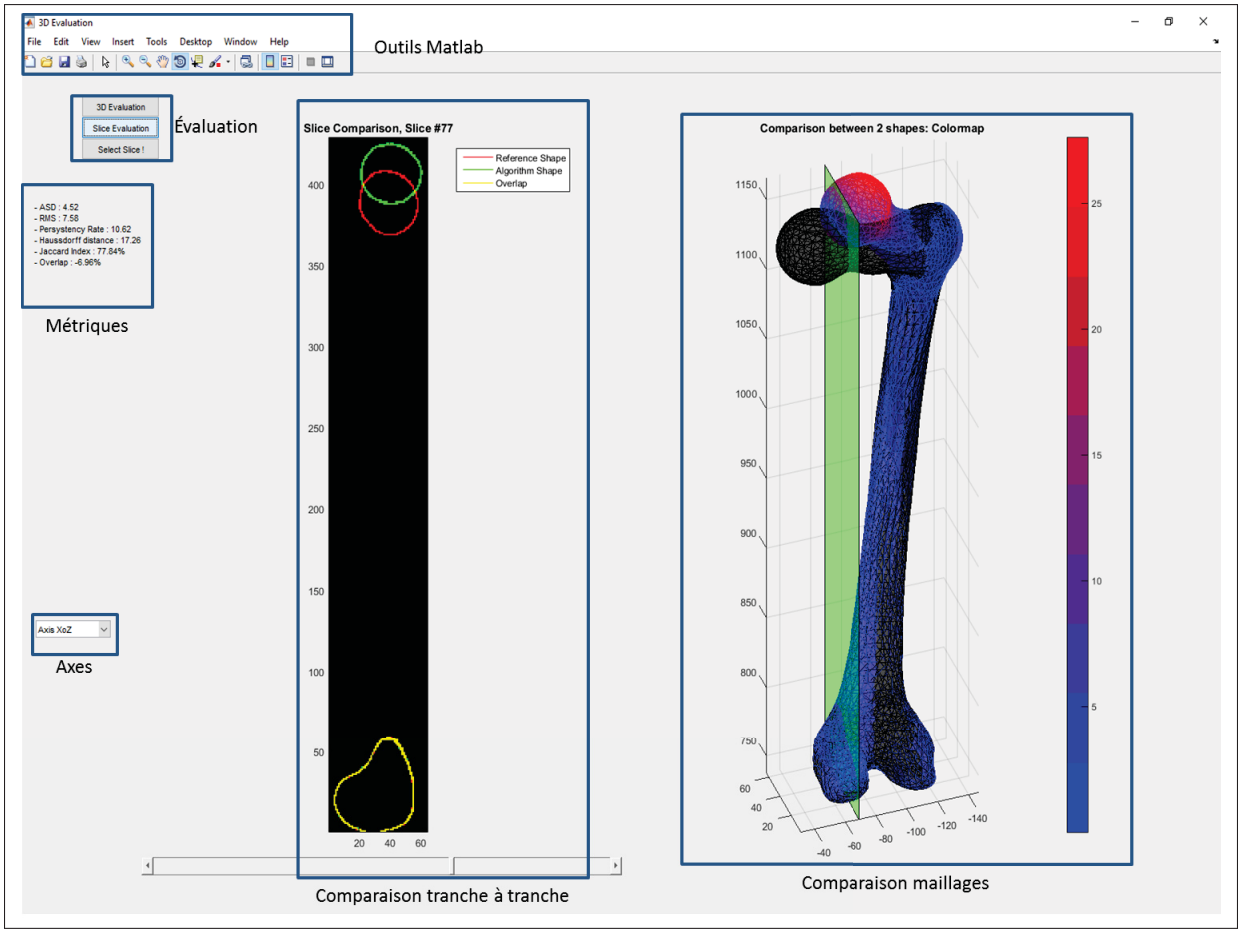

Figure 3.7 Présentation de la plateforme 3D, développée en Matlab

CHAPITRE 4

APPLICATION DE LA PLATEFORME AUX IMAGES 2D

4.1 Méthodologie spécifique au 2D

Dans ce chapitre, nous présentons l'utilisation de la plateforme d'évaluation 2D pour évaluer des algorithmes de détection de la tête fémorale dans des images EOS^{TM} en vue de face.

4.1.1 Ensemble d'images

Deux algorithmes de segmentation 2D ont été testés sur 69 images EOS ^{TM} présentant la hanche ainsi que la tête fémorale. Un protocole éthique a été mis en place et approuvé par les comités éthiques institutionnels de l'ÉTS et du CRCHUM dans le but de traiter ces images. Il s'agit par ailleurs d'images prises dans le plan frontal. La figure 4.4 présente une des images de la base de données.

4.1.2 Génération de l'étalon-bronze

S'agissant de "scènes réelles" (Udupa *et al.* (2006)), nous sommes dans le cas où il faut générer un étalon-bronze en demandant à des experts de segmenter manuellement les images. Pour générer l'étalon-bronze, 3 experts ont segmenté manuellement deux fois chaque image. Ces experts ont tous suivi la même formation sur le logiciel IdefX [1]. Un délai minimum de 2 semaines devait s'écouler entre chaque segmentation du même patient. Par ailleurs, nous avons demandé aux experts de ne pas discuter de leur travail entre eux.

4.1.3 Algorithmes d'identification de la tête fémorale

Deux algorithmes de segmentation 2D de la tête fémorale à partir d'images EOS^{TM} ont pu être comparés :

1. IdefX est un logiciel développé au laboratoire du LIO et utilisé pour les projets internes d'imagerie

- Chav *et al.* (2009), algorithme semi-automatique cherchant le chemin minimum d'un gradient d'intensité sur une forme a priori ;
- Ouertani *et al.* (2015), algorithme multi-structures et automatique inspiré de Chav *et al.* (2009) mais qui segmente simultanément la tête fémorale et le cotyle (contour de la hanche). Il s'agit d'une optimisation de l'algorithme de Chav *et al.* (2009) adapté au cas spécifique de la segmentation de la tête fémorale.

Pour cette étude, nous avons utilisé les critères d'évaluation 2D présentés au chapitre 3 dans le but de pouvoir tracer un graphique radar permettant d'évaluer les contributions de chacun des deux algorithmes.

4.2 Résultats sur des images EOS^{TM} de la tête fémorale

4.2.1 Étalon-bronze

Les experts ont été en accord sur toute la détection de la tête fémorale. Le taux moyen pour la limite de précision σ_{Bs} est d'environ 0.9mm, suivant les cas. L'écart type de la précision v_{Bs} est d'environ 1mm. La figure 4.2 illustre l'obtention de la limite de précision sur une région de l'étalon-bronze. La méthode actuellement développée pour la caractérisation de l'étalon-bronze est semi-automatique. Elle nécessite de cliquer manuellement deux points de l'image, comme nous pouvons le voir dans la figure 4.1. Les points sélectionnés permettent de définir 2 régions :

1. Les régions où la forme d'étude est définie horizontalement (colonne par colonne) ;
2. Les régions où la forme d'étude est définie verticalement (ligne par ligne).

À partir de chacune des régions définies, nous définissons comme limite de précision v_{Bs} la taille de l'intervalle dans lequel la distribution de probabilité normalisée est supérieure à 68% (voir figure 4.2). L'écart type de l'ensemble des limites de précision donne le paramètre σ_{Bs}. Sur les segmentations des experts, nous calculons l'écart type des aires, $SD(|Bs|)$.

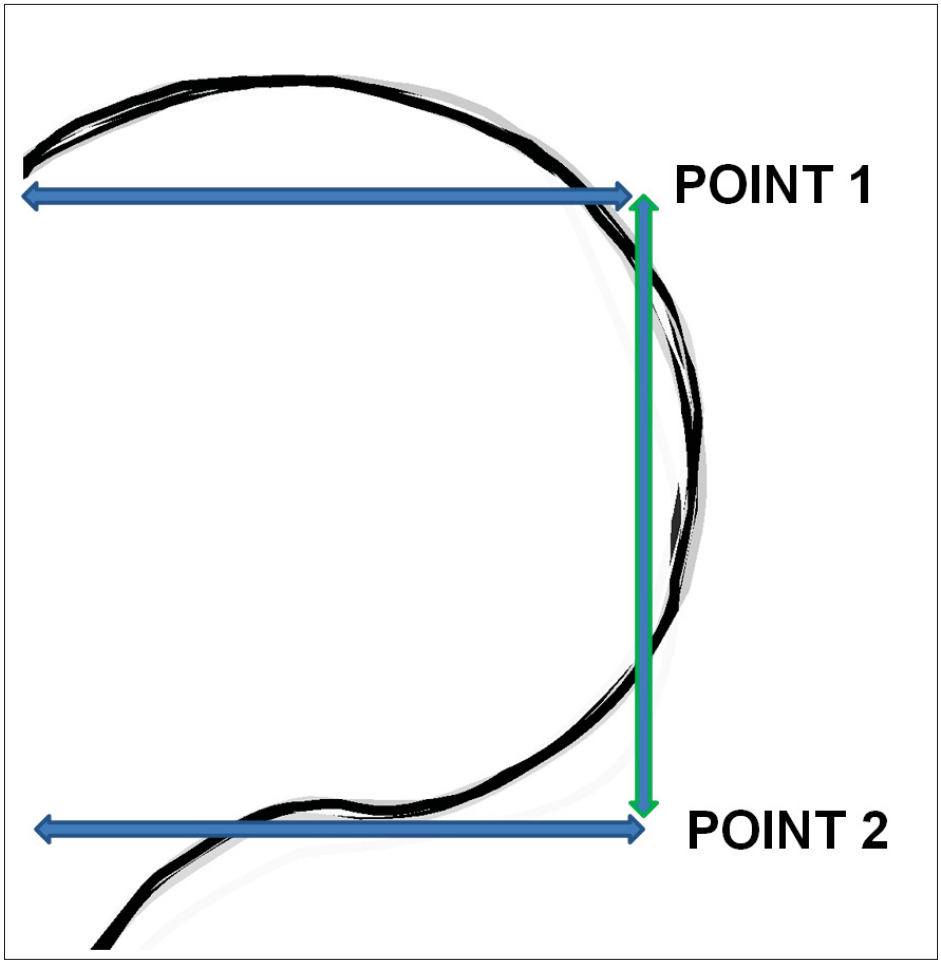

Figure 4.1 Deux points sont sélectionnés dans le but de définir les trois régions d'étude de l'étalon-bronze

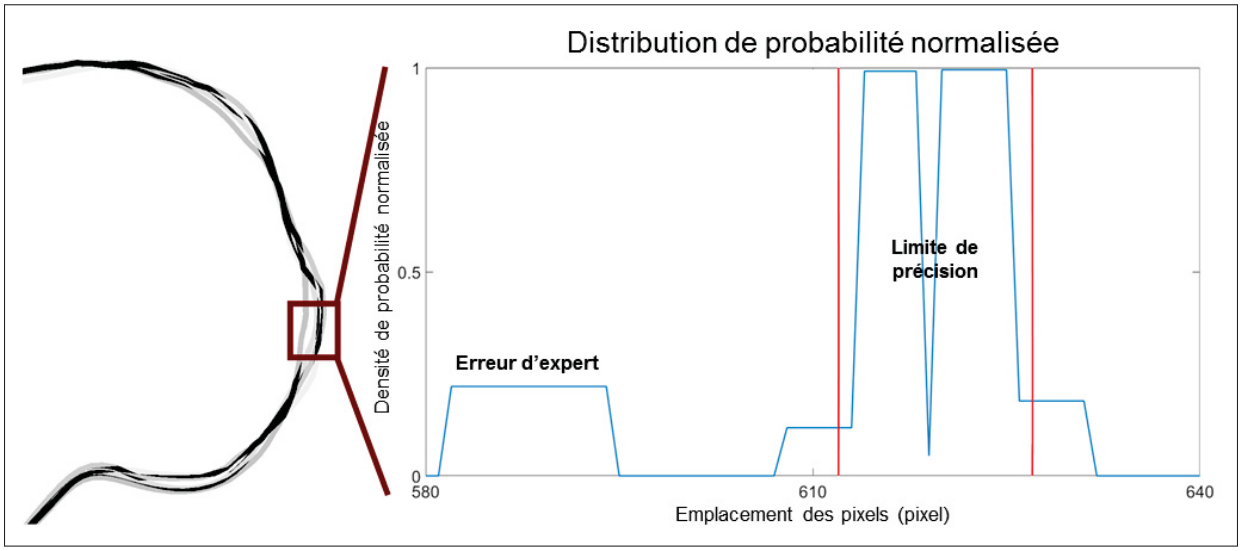

Figure 4.2 Distribution de probabilité normalisée (droite) pour une ligne de la carte de probabilités (gauche) générée par l'algorithme STAPLE
Nous prenons le maximum de probabilité comme étant l'étalon-bronze
La limite de précision est sélectionnée comme étant la région où la distribution de probabilité est supérieure à 68%

4.2.2 Évaluation des algorithmes

Le tableau 4.1 présente les métriques issues de l'analyse entre les algorithmes. La figure 4.3 présente le graphique radar de l'analyse entre les 2 algorithmes. Les critères d'évaluation sont calculés à partir des métriques présentées en chapitre 3.

Tableau 4.1 Tableau de métriques avec écarts type associés pour Ouertani *et al.* (2015) et Chav *et al.* (2009)

Métriques	**Ouertani *et al.* (2015)**	**Chav *et al.* (2009)**
ASD (mm)	1.3 ± 1.0	2.0 ± 1.4
RMSD (mm)	1.6 ± 1.2	2.4 ± 1.6
Hauss (mm)	3.6 ± 2.5	5.2 ± 3.6
JI (%)	90 ± 9	81 ± 8
RSD (%)	7 ± 12	11 ± 11

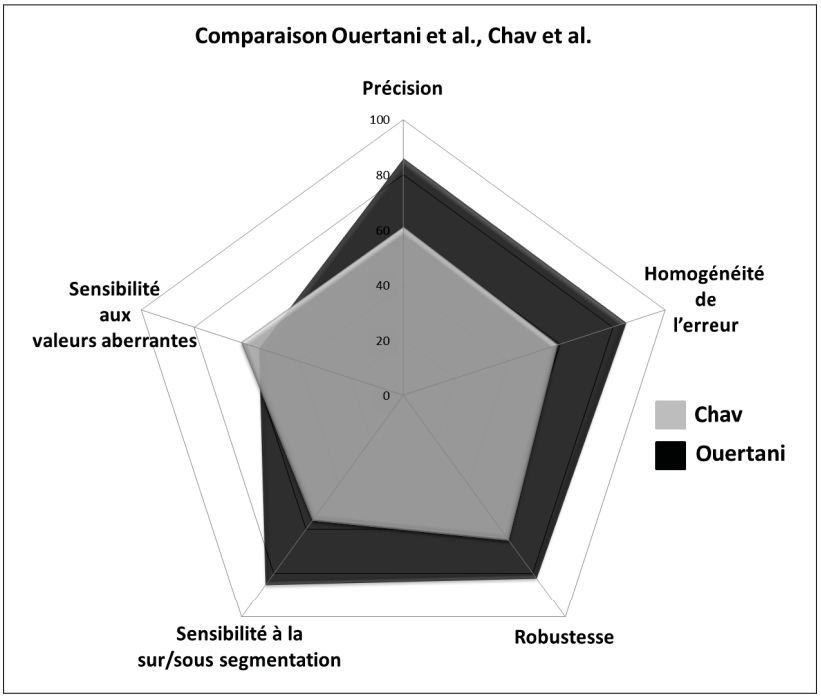

Figure 4.3 Comparaison entre l'algorithme de Ouertani *et al.* (2015) (noir) et celui de Chav *et al.* (2009) (gris)

4.2.3 Interprétation des résultats

D'après les résultats, le tableau de métriques 4.1 permet de constater qu'un des deux algorithme semble fournir des meilleurs performances. Il est toutefois compliqué d'identifier les pistes d'amélioration des deux méthodes et les causes d'erreur. Le graphique radar 4.3 présente les performances de comparaison par rapport à des critères d'évaluation : instantanément, on voit que l'algorithme de Ouertani *et al.* (2015) a des meilleures performances. L'algorithme de Chav *et al.* (2009) présente un score faible (autour de 50%) pour la sensibilité à la sur/sous segmentation. En analysant le comportement de l'algorithme, on constate qu'une fois sur deux il y a une mauvaise interprétation de l'image puisque l'algorithme de Chav *et al.* (2009) détecte le cotyle au lieu de la tête fémorale. La figure 4.5 illustre ce phénomène de sur segmentation. Ce score est largement amélioré (environ 85%) par l'algorithme de Ouertani *et al.* (2015), justement inspiré de Chav *et al.* (2009) mais adapté à la détection de la tête fémorale puisqu'il détecte les deux structures simultanément. Toutefois, on constate que pour un des cinq critères (sensibilité aux valeurs aberrantes), l'algorithme de Ouertani *et al.* (2015) réagit moins bien. Grâce à cette information, il est donc facile d'identifier les zones de lacunes et d'améliorations de l'algorithme. On constate par ailleurs que 15% des cas sont sur ou sous-segmentés. En isolant ces cas, on peut voir que l'algorithme se trompe et va détecter des lignes de densité osseuse (la figure 4.4 présente les différents artefacts qui peuvent être dans les images) ; ce qui explique que l'algorithme est amélioré mais présente quelques pistes de mauvaise interprétation. Un exemple de mauvaise segmentation est illustré en figure 4.5 (figure de droite). Grâce à ce genre d'informations, il devient plus facile de trouver des solutions pour améliorer encore l'algorithme.

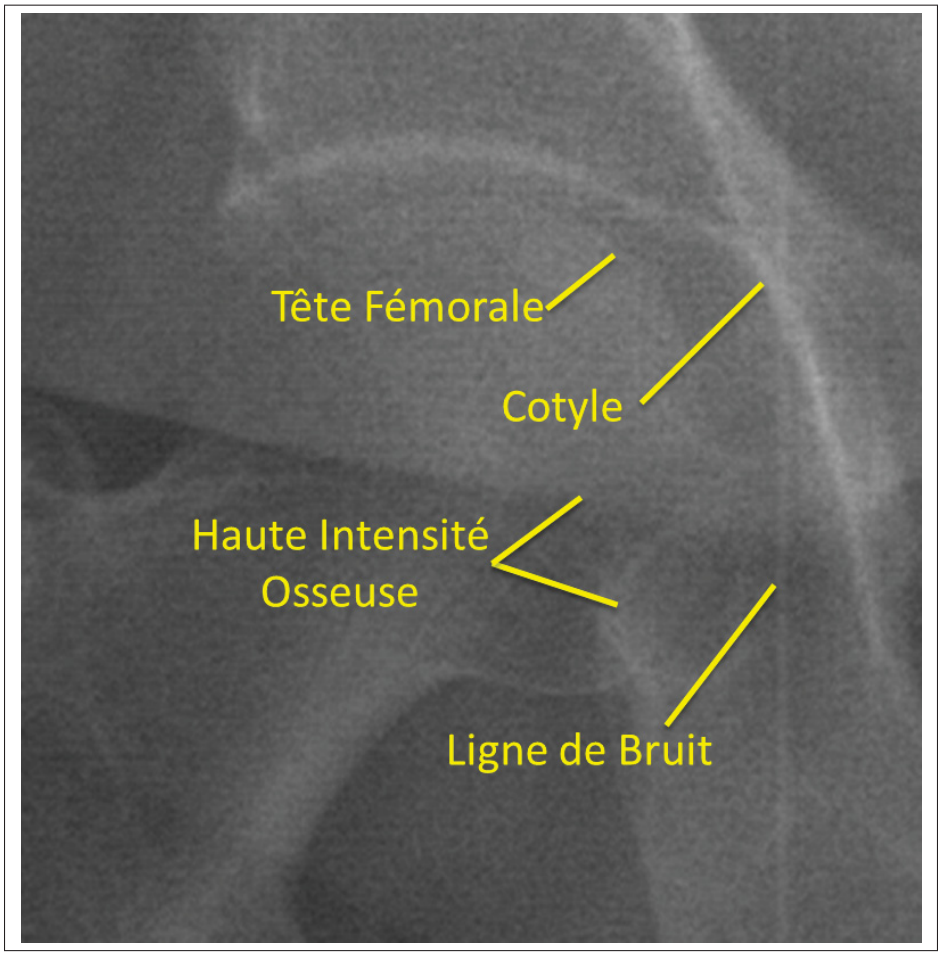

Figure 4.4 Image EOS^{TM} en vue de face présentant la tête fémorale et le cotyle Sur cette image, nous constatons des lignes de bruits ainsi que des zones de haute intensité osseuse pouvant biaiser l'interprétation de l'algorithme

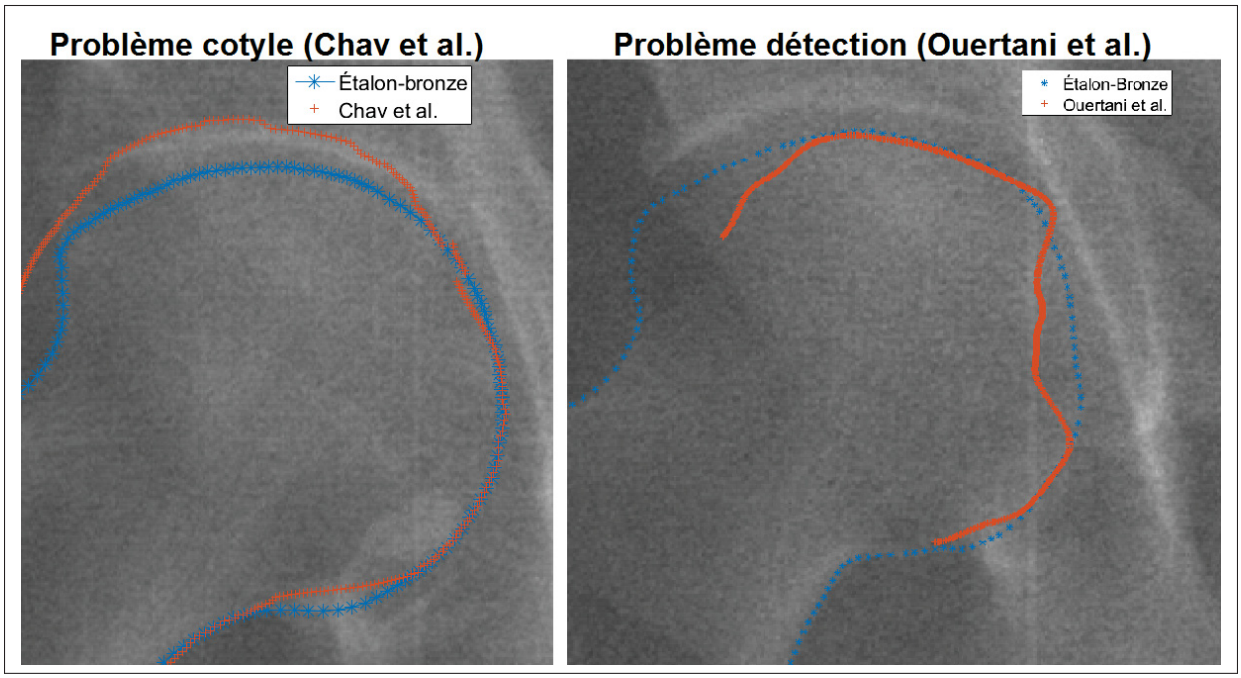

Figure 4.5 Exemple de mauvaises détection lors de segmentation 2D :
Gauche, Chav *et al.* (2009) : Sur segmentation avec détection du cotyle au lieu de la tête fémorale
Droite, Ouertani *et al.* (2015) : Mauvaise segmentation dûe à de l'intensité osseuse trop forte et une ligne de bruit

CHAPITRE 5

APPLICATION DE LA PLATEFORME AUX VOLUMES 3D

5.1 Méthodologie spécifique au 3D

La mise en place d'une plateforme spécifique au 3D nécessite des ajouts par rapport à ce qui a été présenté au chapitre 4. Dans ce chapitre, nous verrons les adaptations apportées à la plateforme de manière à optimiser l'utilisation de la plateforme. Pour présenter la plateforme dans ce contexte 3D, nous montrerons un exemple d'évaluation d'algorithme de détection de tumeurs cérébrales à partir d'images IRM.

5.1.1 Ensemble d'images

Dans cette étude menée en collaboration avec le CRAN de Nancy pour l'évaluation des tumeurs cérébrales (cf Ben Abdallah *et al.* (2016) pour les résultats préliminaires de l'étude), nous avons comparé le comportement d'un algorithme de segmentation 3D développé au laboratoire du LIO au comportement d'experts cliniques utilisant un logiciel de segmentation manuelle. Les images IRM du cerveau ont été acquises sur 12 patients et ont été segmentées dans le plan transverse. Les images de chacun des patients ont été segmentées manuellement tranche-à-tranche par 13 experts cliniques (plus de détail dans Ben Abdallah *et al.* (2016)) en utilisant le logiciel de segmentation manuelle OsiriX [1]. Les images acquises auprès du CRAN de Nancy ont été anonymisées. Ces images ont ensuite été transférées au laboratoire du LIO. La figure 5.1 présente une coupe dans le plan transverse contenant une tumeur.

5.1.2 Étalon-bronze 3D

La génération de l'étalon-bronze est basée sur l'utilisation de l'algorithme STAPLE. Actuellement, aucune génération d'étalon-bronze et caractérisation n'a été mise en place spécifiquement pour les modèles 3D. Nous avons par contre utilisé l'algorithme STAPLE pour faire une

1. OsiriX - http://www.osirix-viewer.com/

génération d'étalon-bronze sur chacune des tranches 2D. Pour chacune de celles-ci, le chemin de probabilité maximal est retenu. La mise en commun de toutes ces tranches permet de générer un étalon-bronze 3D. Il faut ensuite déterminer une manière d'obtenir les paramètres de normalisation ν_{BS}, σ_{BS} ainsi que $SD(|Bs|)$.

5.1.3 Exemple d'application avec un algorithme de segmentation 3D

L'objectif de l'étude du groupe de Nancy était de déterminer si le comportement des experts cliniques est conditionné par différents facteurs : formation, ancienneté, entraînement, etc. L'étude est également très intéressante pour tester la plateforme et principalement la génération d'étalon-bronze, puisque nous possédons des données manuelles de 13 experts.

L'algorithme semi-automatique évalué pour la segmentation 3D est celui développé par Zhou *et al.* (2016). Il s'agit d'un algorithme de segmentation 3D d'IRM prenant simultanément en compte les segmentations 2D dans les trois plans de l'espace (transverse, coronal et sagittal) et les reconstruisant ensemble dans un sous-espace commun pour générer le modèle final 3D. Deux études d'évaluation ont été menées sur ce jeu de données pour évaluer le comportement de l'algorithme :

1. Génération de l'étalon-bronze 3D et analyse du comportement des experts ;
2. Comparaison tranche-à-tranche (évaluations 2D).

5.2 Résultats sur des images IRM de tumeurs cérébrales

5.2.1 Génération d'étalon-bronze

L'étalon-bronze 3D est bâti à partir de contourages manuels réalisés par 13 experts cliniques ayant travaillé dans des conditions optimales (réglage de la luminosité, écran haute-définition etc.). L'objectif des experts était de segmenter les tumeurs cérébrales sur les coupes transverses.

Un exemple de détection manuelle de tumeur cérébrale pour une tranche de l'IRM est visible en figure 5.1.

Figure 5.1 Identification manuelle (contourage bleu) d'une tumeur cérébrale dans une coupe du plan transverse

L'algorithme STAPLE permet de fournir une carte de probabilité tranche-par-tranche de la référence. Il s'agit d'une ressource très conséquente et nous pouvons ainsi analyser la génération de référence à partir d'une carte de probabilités obtenue par le biais de 13 segmentations d'entrée. L'analyse de cette carte de probabilité donne des indications très intéressantes sur le comportement des experts. Contrairement à l'étude sur la détection de tête fémorale présentée en chapitre 4, ici la carte de probabilité n'est pas homogène. Cela signifie que dans certaines régions, les experts sont en accords sur la localisation de la tumeur alors que dans d'autres régions il y a un désaccord sur cette localisation. La géométrie des tumeurs étant très complexe ; contrairement à la tête fémorale du chapitre 4, il n'existe aucun patron simple de représentation des tumeurs. Il est alors laborieux de caractériser la référence avec la méthode semi-automatique proposée. La carte de probabilité amène toutefois de nombreuses informations sur le comportement des experts par rapport à la référence ; nous avons ainsi pu déterminer le pourcentage d'accord entre les experts. Pour ce faire, nous avons analysé la carte de probabilité

issue de l'algorithme STAPLE pour déterminer le nombre de pixels où l'accord final entre les experts est inférieur à 100% (avec une pertinence de 5% ; c'est-à-dire le ratio de pixels où la probabilité est inférieure à 95% sur la carte de probabilité finale). Le tableau 5.1 présente le ratio de désaccord supérieur à 5% par patient, c'est-à-dire le nombre de pixels non nul dont la probabilité issue du STAPLE est inférieure à 95%. Il présente également le nombre de tranches comportant la tumeur (l'épaisseur entre les tranches étant de 5.5mm).

Tableau 5.1 Pourcentage de désaccord supérieur à 5% et nombre de tranches pour les 12 sujets traités

	Désaccord $>$ 5%	**Nombre de tranches**
Pat 1	70.1%	10
Pat 2	34%	12
Pat 3	26.5%	15
Pat 4	33.7%	83
Pat 7	0.1%	7
Pat 8	51.8%	7
Pat 9	1.5%	8
Pat 10	12.4%	7
Pat 11	16.4%	60
Pat 12	25.5%	12
Pat 13	6.8%	12
Pat 14	73.2%	130

Nous avons ensuite sélectionné les tranches pour lesquelles il y a un désaccord entre les experts et nous avons établi que ces tranches se situaient principalement :

1. Ou bien en "extrémité" de tumeur, c'est-à-dire dans les premières et dernières tranches de "grosses tumeurs" (contenant plus de 10 tranches). Dans ces cas, les experts ont généralement du mal à s'entendre pour déterminer à quel niveau (numéro de tranche) la tumeur démarre et s'arrête ;
2. Ou bien sur la totalité de petites tumeurs (tumeurs inférieures à 10 tranches et donc peu développées).

Au niveau de l'intérieur des grosses tumeurs, il y a par contre un accord proche de 100% entre les experts au niveau de la détection de tumeurs. Cette ressource va nous être très utile pour l'évaluation de l'algorithme de Zhou *et al.* (2016) que nous allons présenter en section 5.2.2.

5.2.2 Évaluation d'algorithme de segmentation 3D

Nous avons évalué l'algorithme de Zhou *et al.* (2016) par rapport à l'étalon-bronze défini précédemment. Nous avons généré un étalon-bronze obtenu à partir du maximum de probabilité obtenu sur chacune des tranches 2D définissant le volume. À l'aide de cet étalon-bronze, nous nous avons comparé des métriques d'évaluation tranche-par-tranche. Les métriques choisies sont les suivantes :

1. ASD, distance moyenne par tranche (équation 1.1) ;
2. RMSD, distance moyenne quadratique par tranche (équation 1.2) ;
3. Distance de Hausdorff, erreur maximale par tranche (équation 1.3) ;
4. Indice de Jaccard, indice de similarité par tranche (équation 1.6) ;
5. RSD, différence relative d'aire par tranche (équation 1.5).

Les résultats globaux pour l'évaluation de l'algorithme sont disponibles dans le tableau 5.2. Étant donné que la caractérisation de l'étalon-bronze 3D doit être perfectionnée pour cette étude (voir section 6.2), nous avons extraits les paramètres de performance de la référence pour quelque cas particuliers seulement.

5.2.3 Interprétation des résultats

Le tableau 5.2 semble donner peu de renseignements sur l'analyse de l'algorithme. On constate néanmoins que plusieurs cas semblent se dégager :

Tableau 5.2 Résultat des métriques d'évaluation tranche-à-tranche pour les sujets traités

	ASD **(mm)**	*RMSD* **(mm)**	*Hauss* **(mm)**	*Jacc* **(%)**	*abs*(*RSD*) **(%)**
Pat 1	4.7	6.2	15.1	61	50
Pat 2	5.5	6.7	14.7	40	53.1
Pat 3	7.2	8.6	17.1	33	330
Pat 4	1.6	2.1	4.6	66	24
Pat 7	1.5	1.8	4.4	79	23
Pat 8	7.4	8.8	16.4	33.2	273
Pat 9	2.6	3.3	7.5	67	71
Pat 10	3.0	3.8	8.1	44	22
Pat 11	4.3	5.9	13.6	54	105
Pat 12	3.0	3.6	7.5	55	8
Pat 13	1.8	2.4	6.3	79	32
Pat 14	4.4	5.0	10.3	60	27

1. Des cas tels que les patients 3, 8 et 11 où le RSD est très important (supérieur à 100 %) Dans ces cas, il semble que la segmentation de la tumeur ait été erronée par l'algorithme. Les métriques de distance pour ce cas de figure présentent les valeurs les plus mauvaises ;
2. Des cas tels que les patients 1, 2, 4, 7, 13 et 14 où la somme de la valeur absolue du RSD et de l'indice de Jaccard est proche de 100%. D'après le tableau 3.1 ainsi que les visualisations de résultats au niveau de la plateforme, cela laisse penser à une sur ou sous-segmentation récurrente de la part de l'algorithme ;
3. D'autres cas (patients 9 et 12), où la segmentation fournit les meilleurs résultats.

Il est donc difficile de déterminer un comportement de l'algorithme à partir d'une analyse globale de l'algorithme. Ceci est rendu d'autant plus difficile que, contrairement au cas de la détection de la tête fémorale en chapitre 4, les experts présentent des régions d'accord et de désaccord. Grâce à la plateforme pour le 3D (dont l'interface graphique est visible en figure 3.7), il est possible de visualiser les sorties de l'algorithme par rapport à la référence et d'analyser la carte d'erreur. Nous constatons, dans ce cas que l'algorithme présente également deux comportements différents en fonction des régions d'étude :

1. Il segmente dans un intervalle de confiance proche de l'expert l'intérieur des tumeurs ;

2. Il sous-segmente systématiquement les extrémités des tumeurs ou les petites tumeurs.

Ce comportement est similaire à celui des experts présenté précédemment.

Nous avons testé la plateforme sur des tranches 2D prises dans les 2 régions d'intérêt (5 tranches pour les extrémités et 9 tranches pour les intérieurs des tumeurs). L'étalon-bronze est obtenu à partir de 13 segmentations d'experts. Utilisant la méthode semi-automatique 2D, nous obtenons un intervalle de confiance précis sur les tranches sélectionnées : σ_{Bs} de 1mm, v_{Bs} de 1mm pour les intérieurs de tumeurs, de 3mm pour les extrémités.

Nous remarquons que pour l'intérieur des tumeurs, dont les résultats sont reportés en figure 5.2 l'algorithme se comporte proche de l'intervalle de confiance des 13 experts. Il est à noter que la plateforme distingue une légère sous-segmentation de la part de l'algorithme. Ceci se constate sur le graphe radar associé.

Dans le cas des extrémités de tumeurs, tel que c'est visible en figure 5.3 et dans le graphe radar associé, l'algorithme génère une grosse sous-segmentation correspondant à 60% de l'aire de la référence.

Il est toutefois intéressant de constater que la sous-segmentation générée par l'algorithme est très répétable : dans tous les cas, l'algorithme sous-segmente de manière importante (autour de 60% de différence d'aire avec la référence) par rapport à l'étalon-bronze. Cela fournit des idées de pistes d'amélioration pour l'algorithme.

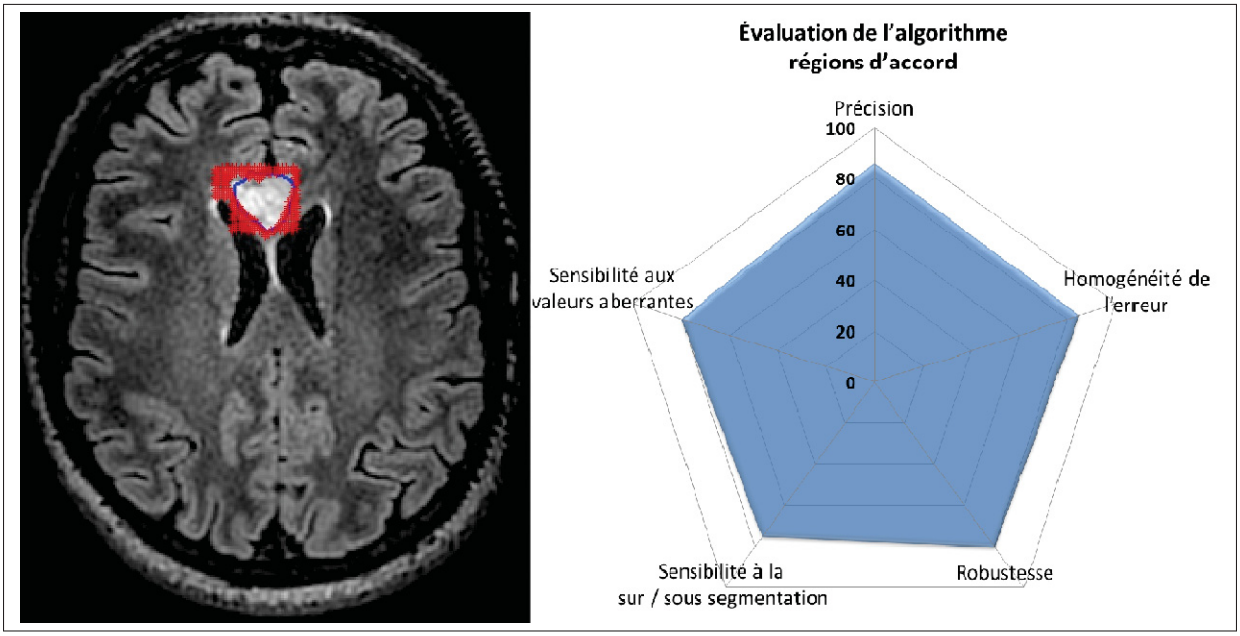

Figure 5.2 Comportement de l'algorithme dans les régions d'accord des experts
Gauche : exemple d'une tranche présentant l'étalon-bronze (croix rouge) et la segmentation de l'algorithme (points bleu)
Droite : Graphe radar associé

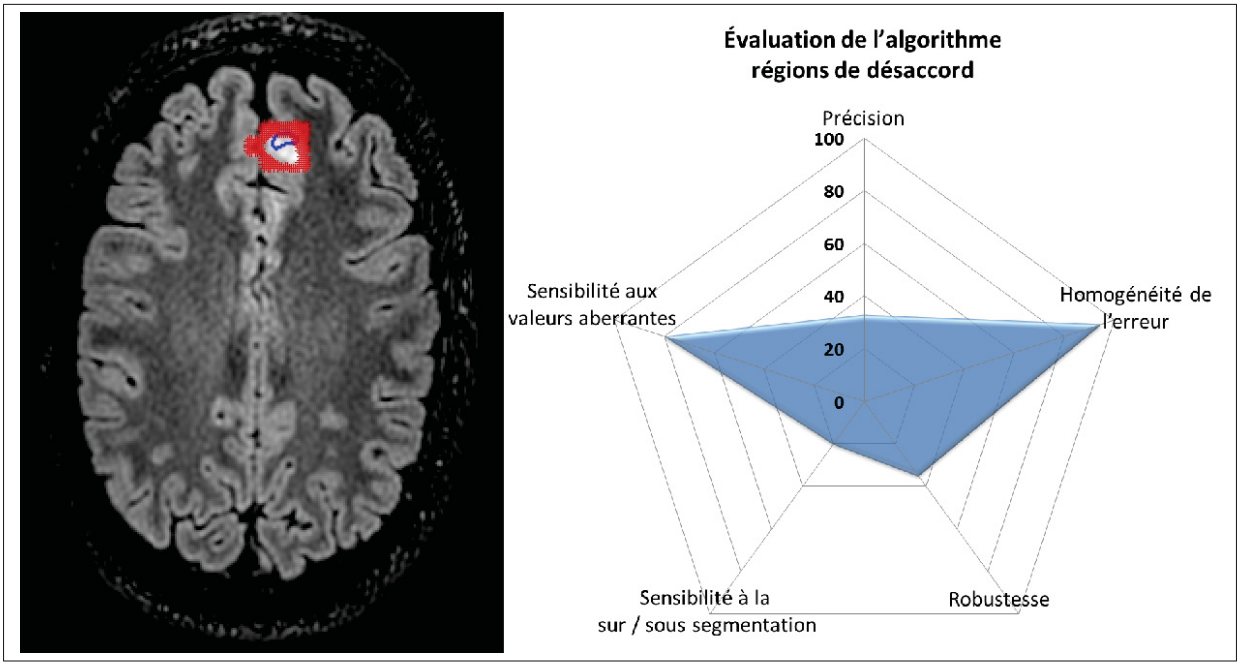

Figure 5.3 Comportement de l'algorithme dans les régions de désaccord des experts
Gauche : exemple d'une tranche présentant l'étalon-bronze(croix rouge) et la segmentation de l'algorithme (points bleu)
Droite : Graphe radar associé

CHAPITRE 6

DISCUSSIONS, CONCLUSIONS ET RECOMMANDATIONS

6.1 Discussions et conclusions

L'objectif de ce mémoire était de mettre en place une plateforme pour évaluer des algorithmes de traitement d'images médicale et de l'appliquer spécifiquement à des algorithmes de segmentation d'images. Cette plateforme se base sur un protocole objectif permettant de distinguer facilement les étapes de l'évaluation. Pour ce faire, nous l'avons développée sur 2 niveaux : dans un premier temps, la mise en place d'un étalon-bronze avec des paramètres de caractérisation puis l'évaluation des algorithmes avec synthèse des résultats dans des graphes de performances.

Dans le cas de "scènes réelles", nous avons défini une méthode permettant non seulement de générer une référence d'étude que nous qualifions d'étalon-bronze, mais nous proposons également une manière de pouvoir caractériser cette référence. La caractérisation de l'étalon-bronze apporte des informations clés sur sa qualité. Elle permet notamment de définir un intervalle de confiance dans lequel les experts ayant produits les segmentations manuelles s'accordent. Cet intervalle de confiance permet de normaliser l'étude d'algorithmes qui s'en suit. Son intérêt est de proposer une référence non biaisée pour l'étude d'algorithmes. Une méthode de caractérisation semi-automatique a été proposée en définissant manuellement des régions représentant la forme d'étude. Toutefois, cette méthode est rendue difficile à utiliser lorsque la forme géométrique de l'objet d'étude se complexifie. L'autre inconvénient de la génération d'étalon-bronze est qu'il n'y a pas encore de méthode spécifiquement adaptée au 3D. Nous utilisons pour l'instant la méthode 2D sur chacune des tranches du modèle 3D.

Nous avons également défini des méthodes d'évaluation d'algorithmes de segmentation d'images. Contrairement à la littérature, nous ne proposons plus une validation uniquement à travers le calcul d'un certain nombre de métriques. Dans la plateforme, nous proposons de mettre en place des critères d'évaluation sous forme de scores normalisés par l'intervalle de confiance des experts et obtenus à partir de métriques de la littérature. Les résultats finaux peuvent en-

suite être reportés dans des graphes de performance radar. L'avantage de ces critères est qu'ils permettent d'obtenir une interprétation du comportement des algorithmes et il devient facile de les comparer entre eux. Cela permet donc non seulement de valider les algorithmes, mais également, d'en évaluer les lacunes et de pouvoir déterminer des pistes d'amélioration.

La mise en place de cette plateforme ainsi que les tests réalisés sur différents algorithmes de segmentation ont montré des résultats intéressants. Pour l'évaluation des algorithmes de détection de la tête fémorale à partir d'images EOS^{TM}, nous avons pu détailler les différences de comportement entre les deux algorithmes et rapidement identifier les points forts et les points faibles de chacune des méthodes. Ce travail d'évaluation permet donc de mettre en lumière les pistes d'amélioration dans le développement de l'algorithme de détection de la tête fémorale développé au laboratoire.

Par rapport à la littérature, nous proposons donc un cadre d'évaluation utilisant un étalon-bronze caractérisé permettant de donner une idée de la qualité de la référence. Par ailleurs, la mise en place de critères d'évaluation et d'un graphe radar de comparaison permet d'avoir des informations sur les voies d'amélioration des algorithmes développés. Enfin, cette méthode a aujourd'hui été implémentée et testée sur des cas d'étude concrets.

La section 6.2 présente les recommandations permettant d'optimiser la plateforme développée.

6.2 Recommandations

6.2.1 Optimisation de la génération de l'étalon-bronze

- **Automatisation**

  La génération de l'étalon-bronze à partir de la carte de probabilités est semi-automatique dans la plateforme. Sur des gros lots de données, cela demande un temps de traitement non négligeable ; par ailleurs, une légère variation en fonction des utilisateurs peut apparaître. Dans ce sens, il serait intéressant de pouvoir automatiser la génération de l'étalon-bronze. Une première proposition, dans le cas de contours de segmentation 2D, serait de faire passer

une spline au niveau des maxima de probabilité et de calculer les normales à la courbe. La figure 6.1 présente un exemple d'obtention automatique de l'étalon-bronze à partir des normales d'une spline. Cela permettrait de caractériser automatiquement et de manière plus précise l'étalon-bronze.

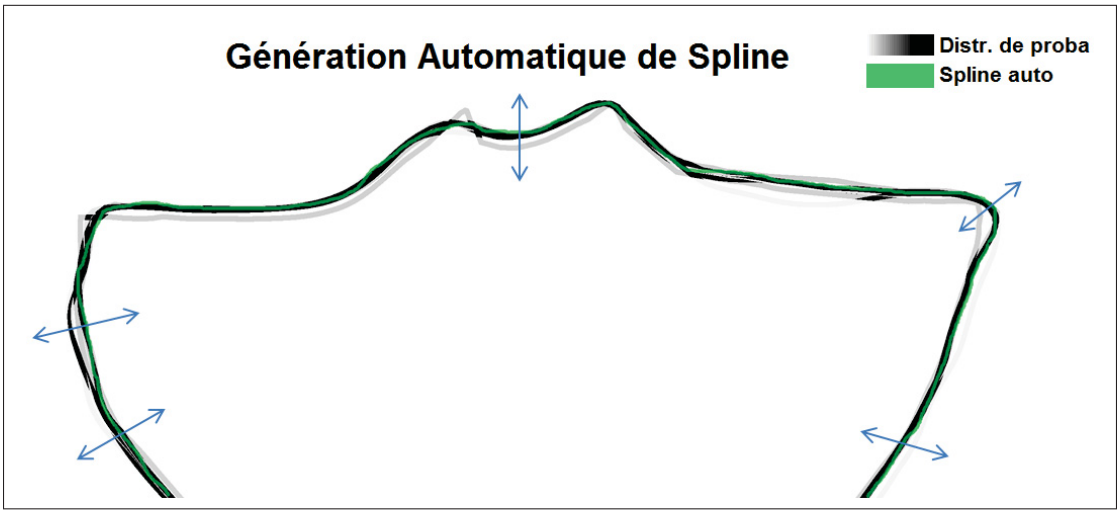

Figure 6.1 Passage de spline permettant d'optimiser et d'automatiser la caractérisation de l'étalon-bronze

- **Caractérisation de l'étalon-bronze**

  La caractérisation proposée a été jugée intéressante en fonction de nos propres choix. Des études supplémentaires pourraient être menées dans le but de caractériser de manière encore plus précise l'étalon-bronze : pour l'instant un seuillage est fait aux 2σ d'une loi normale pour définir l'intervalle de confiance ; il pourrait être intéressant d'étudier des processus plus continus tel qu'une pondération linéaire dépendante de la répartition de la densité de probabilité. Des améliorations du STAPLE tels que la possibilité de formuler des informations a posteriori sur les experts (Commowick & Warfield (2010b)) pourraient être étudiées.

6.2.2 Caractérisation de l'étalon-bronze sur des formes complexes

La caractérisation effectuée lors du chapitre 4 était semi-automatique et adaptée à des formes simples où il était simple de définir 2 points pour les analyses : nous avions des régions où la forme pouvait être traitée ligne par ligne et d'autres colonne par colonne. Dans le cadre de formes à géométrie plus complexe et non prévisible (cf figure 5.1), il faut prévoir un moyen de caractériser l'étalon-bronze à partir d'une forme quelconque ; notamment en faisant passer une spline qui suivrait la forme d'étude et en prenant les valeurs aux normales pour la densité de probabilité (cf figure 6.1).

6.2.3 Adaptation de l'étalon-bronze aux modèles 3D

La génération de l'étalon-bronze spécifiquement pour le 3D n'a pas été traitée dans le cadre de ce projet. Pour l'instant, nous utilisons la méthode 2D sur chacune des tranches du modèle. La mise en commun de l'information fournit un étalon-bronze 3D. Même si les résultats sont encourageants, il pourrait être intéressant, dans le but d'avoir une plateforme complète, de proposer une méthode pour la génération d'un étalon-bronze directement sur les modèles 3D.

6.2.4 Adaptation des critères aux besoins de cahier des charges

En fonction des besoins cliniques, industriels ou de recherche, il peut être intéressant d'adapter la plateforme pour qu'elle puisse répondre le plus possible aux besoins d'un cahier des charges et proposer des rapports personnalisés. Dans ce sens, il devient intéressant de collaborer avec les utilisateurs des algorithmes (cliniciens, industriels, développeurs d'algorithmes) pour leur permettre de profiter amplement des performances de la plateforme.

6.2.5 Amélioration de l'interface logicielle

L'interface logicielle est pour l'instant codée en partie en C++ et en partie en MATLAB, qui est un langage propriétaire. Il pourrait être intéressant de porter tout le code informatique en C++

et l'intégrer dans les logiciels développés par le laboratoire (LibLIO et sterLIO [1]) pour pouvoir diffuser encore plus l'utilisation de la plateforme auprès des chercheurs du laboratoire. Par ailleurs, il pourrait également être envisagé de porter le code sur une plateforme libre internet de manière à pouvoir proposer l'utilisation de la plateforme à quiconque fait du traitement d'images.

6.3 Communication scientifique des résultats

Ce travail a fait l'objet de quatre publications lors de conférences scientifiques :

1. Présentation orale lors du "13th International Symposium Computer Methods in Biomechanics and Biomedical Engineering"
   Le résumé de conférence présenté est disponible en annexe I ;
2. Présentation par affiche lors de la "36ème journée de la recherche du POES et de la division d'orthopédie de l'université de Montréal"
   Le résumé de conférence présenté est disponible en annexe II ;
3. Présentation par affiche lors du "22nd Congress of the European Society of Biomechanic" [2].
   Le résumé de conférence est disponible en annexe III ;
4. Présentation orale lors du "38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society" [3].
   L'article de conférence présenté est disponible en annexe IV.

1. Il s'agit de ressources informatiques développés par le LIO et utiles à tous les étudiants
2. https://esbiomech.org/conference/index.php/congress/lyon2016/
3. http://embc.embs.org/2016/

ANNEXE I

A FRAMEWORK TO EVALUATE AND VALIDATE 2D SEGMENTATION ALGORITHMS ON LOWER-LIMB X-RAYS

Pierre Laurent[1], Thierry Cresson[1], Joseph R. Dadour[1], Julien Clément[1], Nathalie J. Bureau[1], Nicola Hagemeister[1], Carlos Vazquez[1], Jacques A. De Guise[1]

[1] Laboratoire de Recherche en Imagerie et Orthopédie, École de Technologie Supérieure, Centre de Recherche du Centre Hospitalier de l'Université de Montréal, 900 Rue Saint-Denis, Montréal, Québec, Canada H2X 0A9

Résumé soumis et accepté en présentation orale au « 13th International Symposium Computer Methods in Biomechanics and Biomedical Engineering » du 01 au 05 Septembre 2015 à Montréal, Canada.

1. Introduction

Semi or fully-automatic segmentation algorithms allow detecting lower-limb (Femur+Tibia) structures in X-Rays with the advantage of reducing segmentation time and operator variability. However the validity of these algorithms is difficult to evaluate. The literature proposes various approaches to evaluate the accuracy, repeatability and robustness of these algorithms. However, no standard validation framework is currently in use. The purpose of this study is to provide such a framework to complete the full validation and evaluation of 2D Segmentation Algorithms for the lower-limb using an Expert Based Reference, defined as Bronze Standard (Jannin *et al.* (2002)).

2. Methods

The framework is divided in five steps

1. Creation of a Bronze Standard Reference for every lower-limb using the STAPLE algorithm (Warfield *et al.* (2004)) performed on manual segmentation generated by experts. Robustness of the Bronze Standard is obtained by the estimates of performance provided by the STAPLE algorithm ;

2. Quantitative assessment of the Bronze Standard is then performed by evaluating the intra and inter-operator repeatability using RMS, ASD and Hausdorff Distance (HD) metrics (Heimann *et al.* (2009)). At this point, a Bronze Standard Reference with quantitative values for robustness and repeatability is available to evaluate and validate Segmentation Algorithms ;
3. Quantitative metrics such as RMS, ASD and HD are then used to evaluate Segmentation Algorithms in comparison to the Bronze Standard previously generated, offering an analysis of accuracy of the Segmentation algorithm ;
4. Qualitative metrics such as Jaccard Index and Overlap Target (Heimann *et al.* (2009)) are computed in order to evaluate the potential bias in the Segmentation Algorithm and to highlight the areas for improvement of the algorithm ;
5. On the final step, statistical tests : ICC, Bland-Altman Analysis and SEM (Weir (2005)) determine the global accuracy of the algorithm. Step (1) and step (2) of this framework have been tested on a limited number (N = 8) of lower-limbs manually segmented twice by a trained operator to perform an intra-operator analysis.

3. Results

Intra-operator measurements showed excellent quality segmentations for the Femur/Tibia, with an average RMS of 0.5mm/0.5mm, ASD of 0.4mm/0.4mm and HD of 2.7mm/2.5mm. An analysis emerging from this study is shown in figure-A I-1 ; left figure shows that repeatability is below the millimeter ; whereas right figure underlines a bias in the segmentation.

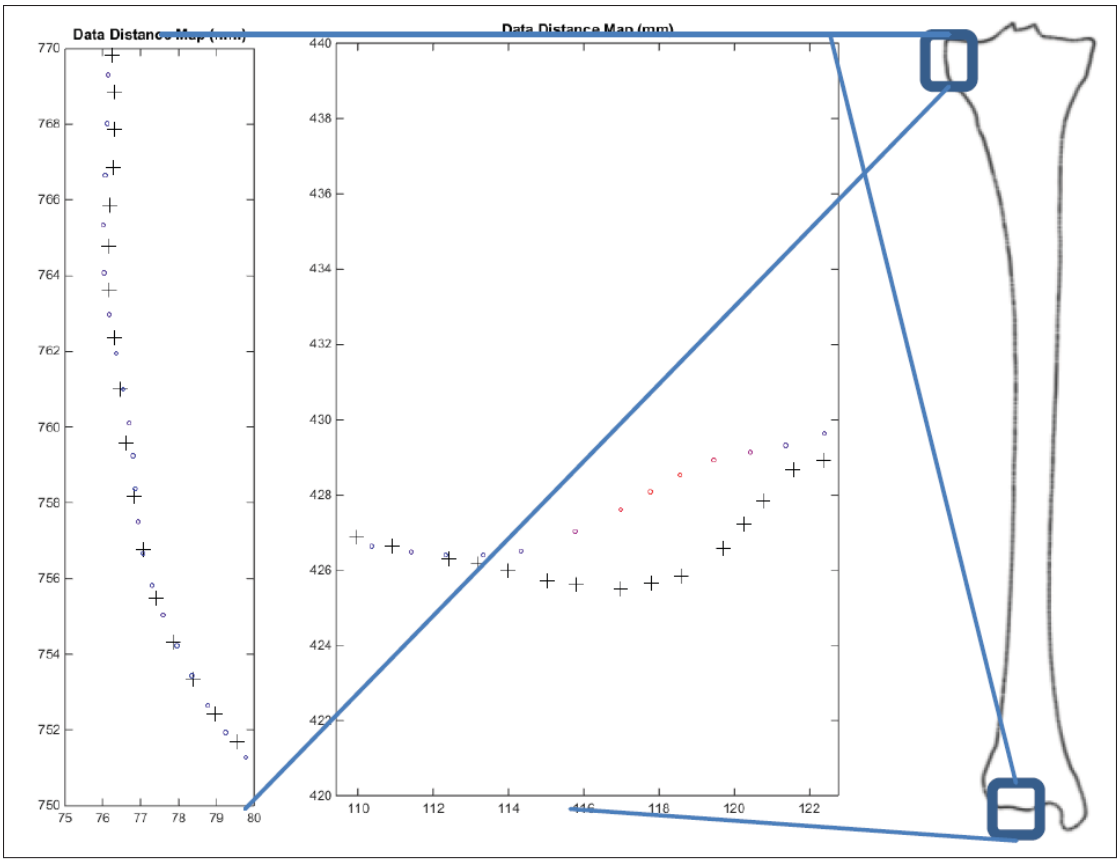

Figure-A I-1 Left : Intra-operator repeatability is below 1mm
Right : Quantitative analysis of the intra-operator showed a bias in the segmentation of a region (deviation > 2.5 mm)

4. Conclusions

These preliminary results are very encouraging for the intra-operator repeatability of the Bronze Standard, which is below the millimeter. The proposed framework will offer a quantitative evaluation for the quality of the Bronze Standard and will provide quantitative and qualitative metrics combined with statistical tests for the global evaluation of Segmentation Algorithms. Further work will consist in analyzing the inter-operator repeatability, evaluate the framework on more images to complete a full study and test the framework on Segmentation Algorithms.

ANNEXE II

PLATEFORME D'ÉVALUATION D'ALGORITHMES DE TRAITEMENT D'IMAGES MÉDICALES

Pierre Laurent[1], Thierry Cresson[1], Nicola Hagemeister[1], Carlos Vazquez[1], Jacques A. De Guise[1]

[1] Laboratoire de Recherche en Imagerie et Orthopédie, École de Technologie Supérieure, Centre de Recherche du Centre Hospitalier de l'Université de Montréal, 900 Rue Saint-Denis, Montréal, Québec, Canada H2X 0A9

Résumé soumis et accepté en présentation par affiche à la « 36ème journée de la recherche du POES et de la division d'orthopédie de l'université de Montréal » le 12 Mai 2016 à Montréal, Canada.

1. Objectifs

Les algorithmes de traitement d'images sont une composante essentielle de l'identification et de la modélisation de structures anatomiques à partir d'images médicales (EOS^{TM}, IRM, etc.). Toutefois, leur évaluation est difficile : il convient de comparer la segmentation produite à une référence manuelle d'un expert, en utilisant une métrique de similarité, souvent compliquée à interpréter. L'objectif de la plateforme est de proposer des outils permettant de :

1. Définir et caractériser une référence à partir de segmentations manuelles d'experts. Nous définissons cette référence comme étant un standard de bronze ;
2. Analyser les performances d'algorithmes en comparant la segmentation produite au standard de bronze ;
3. Fournir une représentation graphique synthétique permettant une évaluation globale des résultats.

2. Méthodes

Un standard de bronze est d'abord généré à partir de multiples segmentations en utilisant l'algorithme STAPLE. Un intervalle de confiance est ainsi obtenu pour un groupe d'experts. Ensuite,

les algorithmes de segmentation sont évalués à l'aide de métriques normalisées issues de la littérature : précision, robustesse, fiabilité, sensibilité à la sur/sous segmentation et distribution des données aberrantes. Les résultats sont enfin reportés dans un graphe radar pour faciliter l'interprétation multicritères. La plateforme a été testée sur deux algorithmes de segmentation de la tête fémorale dans des images radiographiques EOS^{TM}. La base de données d'évaluation comprend 69 images ; chacune ayant été segmentée deux fois par trois experts, permettant de générer le standard de bronze.

3. Résultats

La plateforme a permis de comparer deux algorithmes de segmentation et d'évaluer de façon systématique et quantifiée leur comportement, mettant en lumière leurs avantages et inconvénients.

4. Conclusions

La plateforme propose une démarche d'évaluation normalisée multicritères d'algorithmes de traitement d'images. Il est possible d'évaluer de façon objective et quantifiée les méthodes de traitement d'images et de s'assurer que celles-ci répondent aux besoins du cahier des charges.

ANNEXE III

AN EVALUATION PLATFORM FOR SEGMENTATION ALGORITHMS : AN APPLICATION TO FEMORAL HEAD X-RAY IMAGES

Pierre Laurent[1], Thierry Cresson[1], Joseph R. Dadour[1], Julien Clément[1], Nathalie J. Bureau[1], Nicola Hagemeister[1], Carlos Vazquez[1], Jacques A. De Guise[1]

[1] Laboratoire de Recherche en Imagerie et Orthopédie, École de Technologie Supérieure, Centre de Recherche du Centre Hospitalier de l'Université de Montréal, 900 Rue Saint-Denis, Montréal, Québec, Canada H2X 0A9

Résumé soumis et accepté en présentation par affiche au « 22nd Congress of the European Society of Biomechanic » du 10 au 13 Juillet 2016 à Lyon, France.

1. Introduction

Segmentation algorithms allow detecting bone structures in X-Rays with the advantage of reducing processing time and operator variability. Image interpretation being relative to human perception, the validation of these algorithms is a difficult task. The purpose of this study is to present an evaluation platform, based on Udupa *et al.* (2006) framework, to not only validate segmentation algorithms but also to evaluate their different behaviors using an expert-based reference, defined as bronze standard (Jannin *et al.* (2002)). As an example, we will compare the impact of two different algorithms on the femoral head detection.

2. Methods

The evaluation is divided into 4 stages, computed automatically by an in-house developed software :

1. Creation of a bronze standard for 2D segmentations given experts' inputs. Its reliability and robustness are assessed using the STAPLE algorithm (Warfield *et al.* (2004)) ;
2. Assessment of the quality of segmentations using specific metrics and applying specific rules :

- first, RMS (Heimann *et al.* (2009)) and ASD (Heimann *et al.* (2009)) are computed to determine the mean errors of both algorithms. The ratio between those metrics informs on the relative discrepancy of outliers and their global distribution ;
- then, the Haussdorff distance (Heimann *et al.* (2009)) together with a Persistency Rate (PR) that we defined allow underlining outliers' regions on the images. The Persistency Rate represents the error value above which 33% of outliers are located ;
- finally, the Jaccard Index (Heimann *et al.* (2009)) together with the Overlap Index (Heimann *et al.* (2009)) applied on those specific regions allow identifying and interpreting the behavior of algorithms : over/under or wrong segmentation.

3. Interpretation of the metrics by performing a Bland-Altman analysis defines the local and global behavior of algorithms ;
4. Global analysis allows comparing algorithms within the bronze standards reliability range, asserting their performances and finding ways of improvement. Those stages have been tested on 69 femoral head X-Rays comparing segmentation algorithms from Chav *et al.* (2009) with that from Ouertani *et al.* (2015). The bronze standard was generated by 3 trained operators who segmented each subject twice. The global behavior was asserted by a Bland-Altman analysis.

3. Results

The STAPLE algorithm revealed a reliability range of about 0.9mm for the bronze standard. It was used to perform an evaluation and assess the performances of two segmentation algorithms. Results are shown in table-A III-1.

Performing a Bland-Altman analysis (see figure-A III-1) on 69 cases allowed characterizing a bias of over-segmentation for Chav *et al.* (2009) in 40% of the cases while Ouertani *et al.* (2015) was in agreement with the bronze standard.

Tableau-A III-1 Results on 2D metrics on 69 X-Rays femoral head

2D	**Ouertani *et al.* (2015)**	**Chav *et al.* (2009)**
ASD (mm)	1.3 ± 1.0	2.0 ± 1.4
RMSD (mm)	1.6 ± 1.2	2.4 ± 1.6
Hauss (mm)	3.6 ± 2.5	5.2 ± 3.6
JI (%)	90 ± 9	81 ± 8
RSD (%)	7 ± 12	11 ± 11

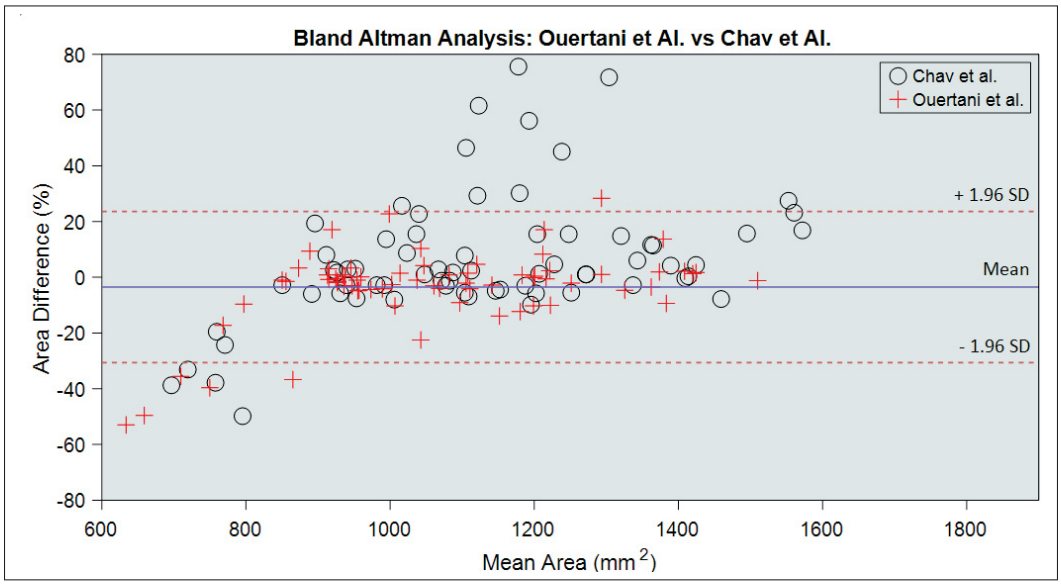

Figure-A III-1 Bland-Altman analysis for 69 cases

4. Conclusion

Using the platform, we could compare algorithms quantitatively and understand how they behave. It helped in comparing two segmentation algorithms : we noticed that Ouertani's algorithm is more robust but can be improved in 20% of the cases where it is under-segmenting, due to image interpretation mismatch.

5. Acknowledgements

The authors would like to acknowledge the financial support of NSERC, Chaire de recherche du Canada and MITACS.

ANNEXE IV

A MULTI-CRITERIA EVALUATION PLATFORM FOR SEGMENTATION ALGORITHMS

Pierre Laurent[1], Thierry Cresson[1], Nicola Hagemeister[1], Carlos Vazquez[1], Jacques A. De Guise[1]

[1] Laboratoire de Recherche en Imagerie et Orthopédie, École de Technologie Supérieure, Centre de Recherche du Centre Hospitalier de l'Université de Montréal, 900 Rue Saint-Denis, Montréal, Québec, Canada H2X 0A9

Article soumis et accepté en présentation orale au « 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society » du 16 au 20 Août 2016 à Orlando, États-Unis.

1. Abstract

The purpose of this paper is to present a platform for evaluating segmentation algorithms that detect anatomical structures in medical images. Structure detection being subject to human interpretation, we first describe a method to define a ground truth model, i.e. a generated bronze standard, that will be the reference for subsequent analysis. This bronze standard will be characterized in order to retrieve its confidence level that will later be used to normalize the algorithm evaluation. We then describe how the developed platform helps in evaluating algorithm performances described using five evaluation criteria : accuracy, reliability, robustness, under/over segmentation sensitivity and outlier sensitivity. First, we explain how to extract those evaluation criteria using specific normalized metrics commonly found in the literature, then we present how to combine all the information in order to get a global evaluation of segmentation algorithms. Lastly, a radar-style graph analysis is presented for easy multi-criteria interpretation.

2. Introduction

Segmentation algorithms allow detection of anatomical structures in medical images with the advantage of reducing processing time and operator variability. Ways of ranking algorithms can

be found (Landman & Warfield (2012)), however there is little information about comparing and evaluating the algorithm performances. To analyze and validate such algorithms, the results are generally compared to manual references produced by a single expert (Sun *et al.* (2012), Chav *et al.* (2009), Chen *et al.* (2014)). As detection of structures results from a subjective interpretation of the images, there is a major difficulty in providing a valuable reference with known limits (Warfield *et al.* (2004), Udupa *et al.* (2006), Khooshabi (2013)). To generate a robust reference, Warfield *et al.* (2004) suggested the STAPLE algorithm, updated by Commowick *et al.* (2012), that produces a “ground truth" on a large number of manual segmentations of the same image performed by different experts. As the standard is, thus generated by many different manual references, we label it “bronze standard" (Jannin *et al.* (2002)). To evaluate and compare algorithms to this bronze standard, the literature suggests various evaluation criteria :

Accuracy

represents the “difference between observed values and theoretical values " (Jannin *et al.* (2002)).

Reliability

is the “consistency of a test or measurement " (Weir (2005)), i.e. the capability to retrieve an homogeneous pattern in the algorithm behavior.

Robustness

is the “capability to segment the object of interests " (Diop & Burdin (2013)).

In addition, we present in this paper two new evaluation criteria in order to refine algorithm evaluation.

Under/over segmentation sensitivity

as a global criterion that aims at determining the tendency of the algorithm to over or under segment beyond a defined sensitivity parameter.

Outlier sensitivity
as the evaluation of the impact of outliers on the final identification. This criterion is helpful in identifying the presence of clusters of outliers that have a significant impact on the final segmentation.

Based on these five evaluation criteria, we propose a platform that aims to :

1. Generate and characterize a unique reference from a set of expert-based segmentations using the STAPLE algorithm. We label this reference **bronze standard** ;
2. Evaluate the algorithm performances by comparing the resulting segmentation to the bronze standard by applying the five **evaluation criteria** obtained using normalized **metrics** ;
3. Provide a simple and easy to understand **graphical** representation of the global multi-criteria evaluation.

3. Methodology

The proposed platform is divided into two main steps : the first goal is to generate a ground truth segmentation retrieved from experts' manual segmentation that will become the bronze standard for future analyses. The bronze standard has to be characterized in order to determine its confidence level. Once a bronze standard is generated, segmentation algorithms can be evaluated in terms of the presented evaluation criteria. Figure-A IV-1 summarizes the proposed platform.

3.1 Bronze standard generation and confidence level

This paper is dedicated to what Udupa *et al.* (2006) considers being "real scenes", i.e. images of patients undergoing medical imaging protocols. In this case, it is impossible to establish an absolute true segmentation. As mentioned by Warfield *et al.* (2004) and Udupa *et al.* (2006), a reference segmentation used as ground truth, needs to be defined. This reference is created using medical images. Each image is segmented by a number of experts. This data is introduced to the STAPLE algorithm (Commowick *et al.* (2012)), which has been chosen to generate the

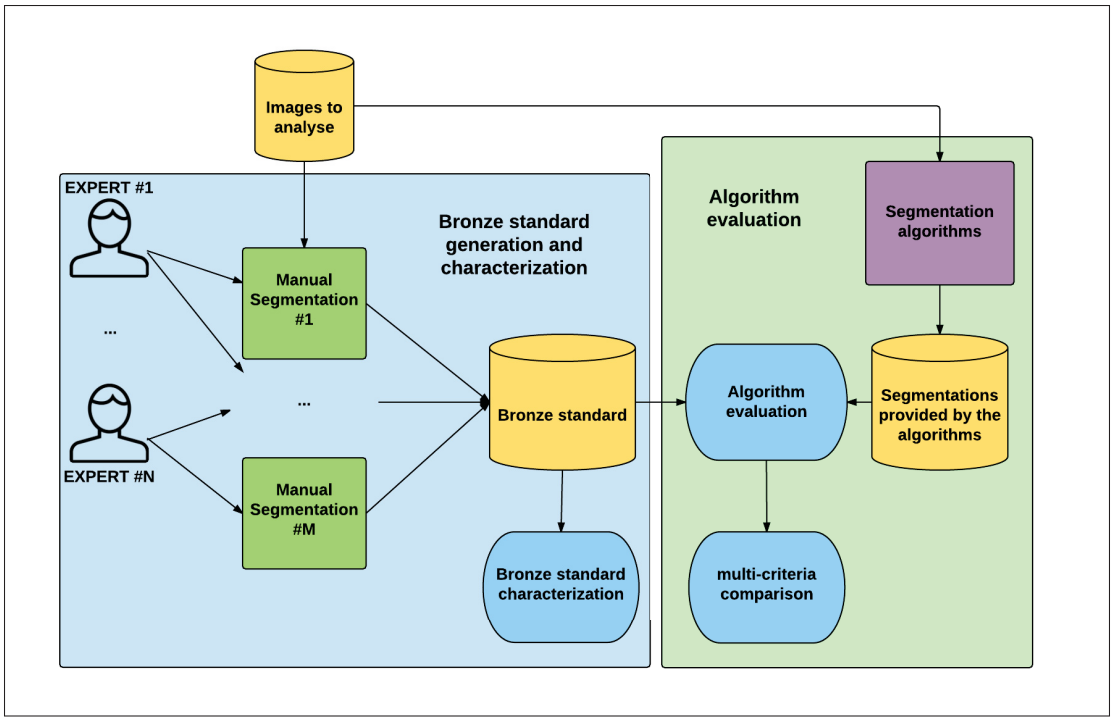

Figure-A IV-1 Summary diagram of the platform

reference, in order to get the ground truth. As the reference is generated, we label it the bronze standard. STAPLE's estimate of performance assesses **robustness** providing an understanding on how interpretation of the medical images can differ in the experts' minds depending on the regions. It determines the probability distribution of the final reference in accordance to the experts' inputs. It therefore outlines the experts' **reliability** on "agreement/disagreement" regions. For each region of the generated ground truth, the spatial distributions of structures is retrieved. The bronze-standard shape is generated were the maximum-likelihood probability is higher. An accuracy limit ν_{BS} with corresponding standard deviation σ_{BS} is calculated as the spatial width in which the normalized probability distribution is higher than 68,2 % of maximum-likelihood. On each generated closed shape, the **area standard deviation** $SD(|Bs|)$ is also calculated. These characteristics represent the bronze standard **confidence level**.

3.2 Evaluation criteria

To evaluate segmentation algorithms performances, we compare them to the bronze standard and retrieve five evaluation criteria based on a normalized set of metrics extracted from the literature. For each criterion, the score is normalized. A score close of 100 signifies that the algorithm is within the bronze standard's confidence level. For all this part, let $S(Al)$ (resp. $S(Bs)$) denote the set of 2D/3D points of the shape generated by the algorithm Al (resp. Bs for the bronze standard shape). These shapes are closed within the region of interest. We define the following expressions :

$$
\begin{aligned}
&|S(U)|, \text{ number of 2D/3D points in } S(U)\\
&|U|, \text{ area defined by the closed shape } S(U)\\
&d^n(p_i, S(U)) = (\min_{\omega \in S(U)} \|p_i - \omega\|)^n\\
&D^n_U(U,V) = \sum_{p_i \in S(U)} d^n(p_i, S(V))
\end{aligned}
$$

Accuracy A

is the relative distance of the shape generated by Al to the bronze standard, considered ground truth. It is obtained as the ratio of Average Symmetric Distance Heimann *et al.* (2009) (ASD) between Al and Bs and the accuracy limit ν_{Bs}.

$$
A = 100 \times \frac{\nu_{Bs}}{ASD(Al)} \tag{A IV-1}
$$

$$
\text{with, } ASD(Al) = \frac{D^1_{Al}(Al,Bs) + D^1_{Bs}(Bs,Al)}{|S(Al)| + |S(Bs)|}
$$

Reliability R

express how Al data are homogeneous compared to Bs. It uses the standard deviation σ_{Al} of Al data distance to Bs. It is inspired by the reliability coefficient defined by Weir (2005). The higher the reliability is, the more a global behavior can be determined. We defined it as :

$$R = 100 \times \frac{2 \times \sigma_{Bs}^2}{\sigma_{Bs}^2 + \sigma_{Al}^2} \quad \text{(A IV-2)}$$

$$\text{with, } \sigma_{Al} = SD(\{d(p_i, S(Bs)), p_i \in S(Al)\})$$

Robustness Φ

is the capacity of *Al* to detect the targeted object. We defined it multiplying the absolute value of Relative Surface Difference (Heimann *et al.* (2009)) (*RSD*) by the Jaccard Index (Heimann *et al.* (2009)) (*JI*) that measures the ratio of *Al* area in common to *Bs* area. To increase the performance of the robustness analysis, it is important to have a large database with many different cases representing all the various possible cases.

$$\Phi = JI(Al, Bs) \times (100 - |RSD(Al, Bs)|) \quad \text{(A IV-3)}$$

$$\text{with, } \begin{cases} RSD(Al, Bs) = 100 \times \dfrac{|Al| - |Bs|}{|Bs|} \\ JI(Al, Bs) = 100 \times \dfrac{|Al \cap Bs|}{|Al \cup Bs|} \end{cases}$$

Under/over segmentation sensitivity Θ

represents the sensitivity of *Al* to under/over segment. It is calculated using the relative surface difference over a defined sensitivity parameter α. By default it is $1.96\ SD(|Bs|)$, i.e. the range where 95% of experts' shapes lie, assuming they are normally distributed (Bland & Altman (1986)).

$$\Theta(\alpha) = 100 \times \frac{Card(\{S(Al) : |\ |Al| - |Bs|\ | < \alpha\})}{Card(\{S(Al)\})} \quad \text{(A IV-4)}$$

$$\text{with, } \alpha = 1.96\ SD(|Bs|) \text{ by default}$$

Combining $\Theta(\alpha)$ and Φ gives information on *Al* behavior :

$\Theta(\alpha)$ \ Φ	**Low**	**High**
Low	Tends to generate a wrong segmentation	Tends to under / over segment
High	Tends to shift the structure detection	Tends to segment the structure of interest

Outlier sensitivity Δ

calculates the relative weight of outlier clusters based on their statistical dispersion. Outliers are defined using the *RMSD* (Heimann *et al.* (2009)) value.

$$\Delta = 100 \times (1 - \frac{\delta(O)}{\delta(Al)}) \tag{A IV-5}$$

$$\text{with, } \begin{cases} \delta(U) = \sum\limits_{p_i \in U} \sum\limits_{\substack{p_j \in U \\ p_j \neq p_i}} \dfrac{d^2(p_i, S(Bs))}{d^2(p_i, p_j)} \\ O = \{p \in Al, d(p, S(Bs)) > RMSD(Al)\} \\ RMSD(Al) = \sqrt{\dfrac{D^2_{Al}(Al, Bs) + D^2_{Bs}(Bs, Al)}{|S(Al)| + |S(Bs)|}} \end{cases}$$

3.3 Multi-criteria graph interpretation

Once the evaluation criteria are retrieved, results are displayed in a radar-style graph for easy multi-criteria interpretation. Figure-A IV-2 presents a global radar-style graph with case study examples on three synthetic algorithms.

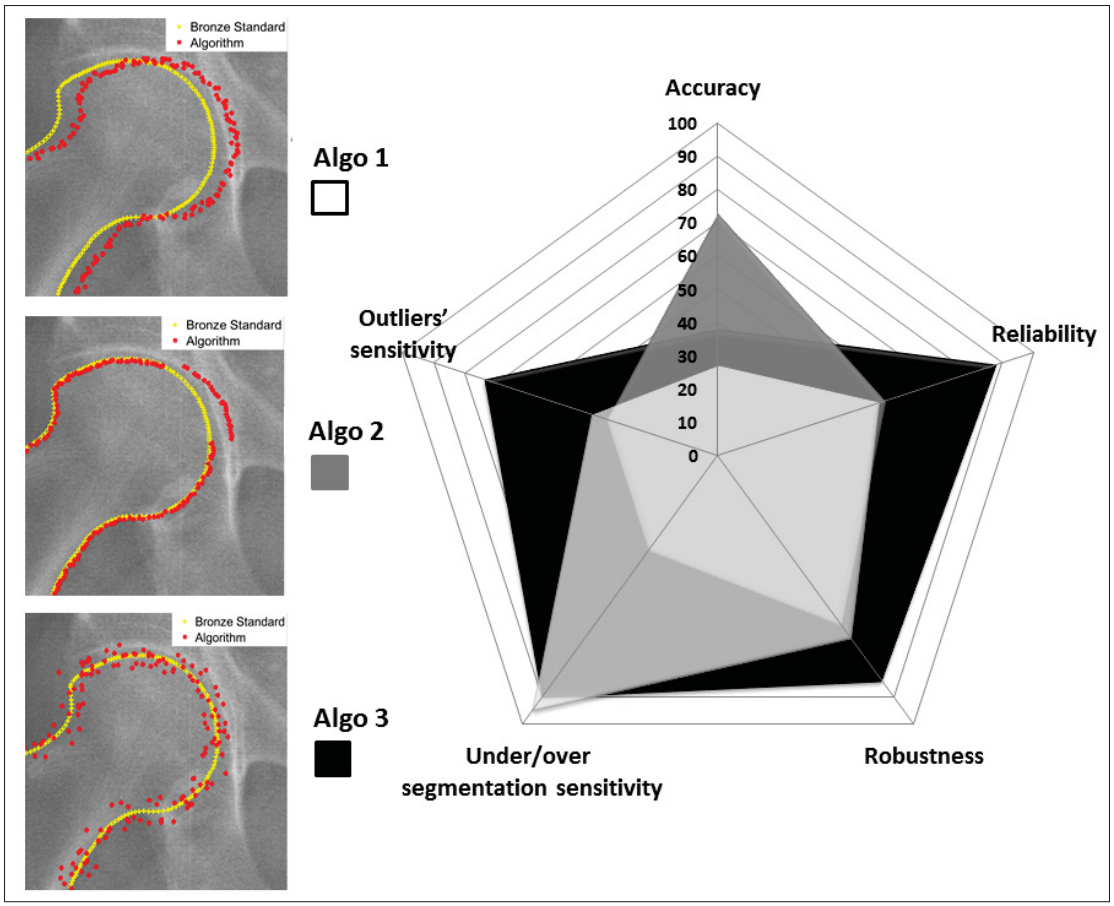

Figure-A IV-2 Radar-style graph obtained on three synthetic algorithms
Algo1 has a registration offset
Algo2 is partially over-segmenting
Algo3 has its data scattered

3.4 Experimental conditions

3.4.1 Segmentation algorithms

We compared two segmentation algorithms for femoral head detection :

1. Chav *et al.* (2009), semi-automatic minimal path segmentation algorithm based on a prior shape ;
2. Ouertani *et al.* (2015), multi-structure algorithm of the femoral head that detects both the femoral head and the cotyle at the same time.

3.4.2 Image data

The chosen evaluation dataset consists of 69 X-Rays images of the femur taken in antero-posterior view. For this study, the femoral head was segmented.

3.4.3 Manual segmentation for bronze standard generation

To segment the images used to generate the bronze standard, experts manually selected points on the outside shape of the femoral contour in X-Rays. Each image was segmented twice by three different experts. An in-house software generated a continuous shape using active contours. Before recording the final continuous shape, experts could perform manual retouching using a smoothing filter. Experts were not allowed to compare their work.

4. Results and discussions

4.1 Bronze standard generation

Experts agreed with a 0.9mm rate of accuracy and 1.0mm standard deviation ($\nu_{Bs} = 0.9mm$, $\sigma_{Bs} = 1.0mm$). As exposed in figure-A IV-3 the region at the extremity of the femoral head was most difficult to segment. Figure-A IV-3 shows an example of a STAPLE ground truth generation with associated probability distribution and accuracy limit taken on the case of the femoral head.

4.2 Algorithms evaluation

4.2.1 Analysis

The evaluation process was performed on the database of 69 cases using an in-house software. Global results are displayed in the radar-style graph of figure-A IV-4.

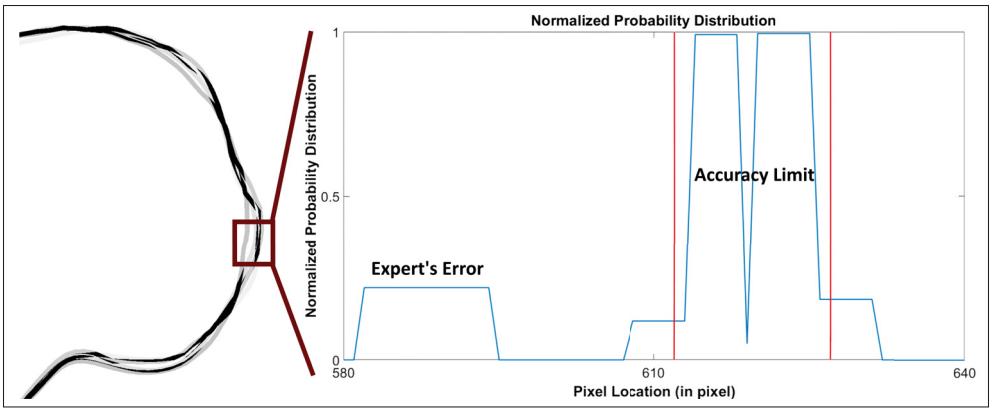

Figure-A IV-3 Example of normalized spatial distribution for the pixel location on the gound truth generation. The figure shows that some experts have a different interpretation (expert's error) for the ground truth detection and will therefore be pulled out of the analysis. Accuracy limit for this region is about 25 pixels (0.9 mm in our case)

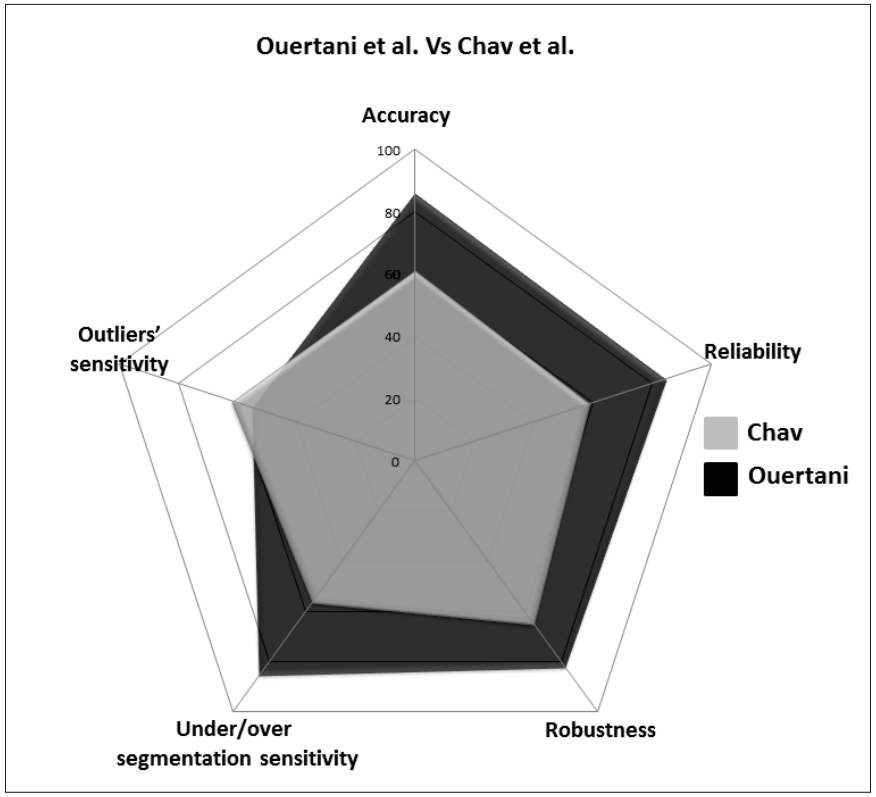

Figure-A IV-4 Radar-style graph representing the evaluation of two algorithms
Black : Ouertani *et al.* (2015)
Grey : Chav *et al.* (2009)

4.2.2 Discussion

The platform showed that Ouertani's algorithm is superior to Chav's algorithm for 4 out of 5 criteria. On figure-A IV-4, it can be seen that :

Chav's algorithm presents a lower score on under/over segmentation sensitivity (56) compared to Ouertani's algorithm (85). This is explainable by the fact that Chav's algorithm is based on a minimal-path algorithm applied on gradient intensity. In most images, it appears that the cotyle has a higher intensity variation than the femoral head. Therefore, Chav's algorithm has a tendency to detect the cotyle instead of the femoral head. This happens in 45% of the cases and induces an error of over segmentation. As for Ouertani's algorithm, thanks to a multi-structural approach detecting both the cotyle and the femoral head at the same time, this problem is managed. This is reflected in the accuracy, reliability and robustness score, where Ouertani's algorithm performs with score above 85. Only outlier sensitivity is lower for Ouertani's algorithm. In addition, a score of 85 in segmentation sensitivity expresses that there are cases of under/over segmentation (15%). Indeed, when the images present high bone density intensity artifacts (see figure-A IV-5), Ouertani's algorithm falsely detect these zones and produces an erroneous output.

In conclusion, the platform allows to highlight that Chav's algorithm uses a method that detects the wrong structure in 45% of the cases inducing a small over-segmentation, whereas Ouertani's algorithm uses a method that is globally highly reliable and accurate and acts in 85% of the cases within the bronze standard confidence level, but has some few cases of total mismatch.

5. Conclusions

We have proposed an evaluation platform that aims to evaluate segmentation algorithms. It generates a unique bronze standard, of which the confidence level is characterized. The confidence level is used to normalize the evaluation process. Using the characterized bronze standard, it is then possible to evaluate and compare segmentation algorithms. The evaluation is based on five evaluation criteria retrieved from a normalized combination of metrics found in

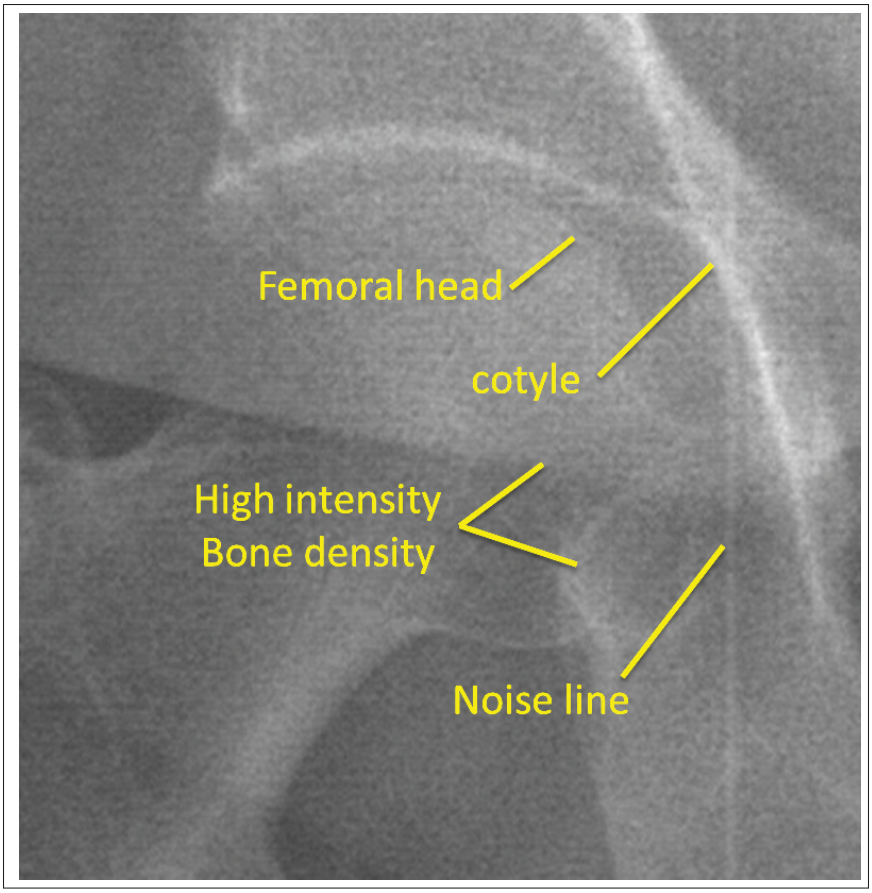

Figure-A IV-5 Problem in femoral head detection, misinterpretation of the structures

the literature. The result is provided in a radar-style graph for easy multi-criteria interpretation. Further work will consist in testing the platform on more segmentation algorithms and grow a collaborative database where it should be possible to evaluate new segmentation algorithms and to compare them to what already exists in the literature.

6. Acknowledgements

The authors would like to acknowledge the financial support of NSERC, MEDTEQ, Research Chairs and MITACS.

BIBLIOGRAPHIE

Akhondi-Asl, A. & Warfield, S. K. (2011). A Tutorial Introduction to STAPLE. *Crl.med.harvard.edu*.

Balestra, S., Schumann, S. & Heverhagen, J. (2014). Articulated Statistical Shape Model-Based 2D-3D Reconstruction of a Hip Joint. *... processing in computer ...*, 128–137. Repéré à http://link.springer.com/chapter/10.1007/978-3-319-07521-1{_}14.

Bayarri, M. J., Berger, J. O., Paulo, R., Sacks, J., Cafeo, J. a., Cavendish, J., Lin, C.-H. & Tu, J. (2007). A Framework for Validation of Computer Models. *Technometrics*, 49(2), 138–154. doi : 10.1198/004017007000000092.

Ben Abdallah, M., Blonski, M. & Wantz-m, S. (2016). Statistical evaluation of manual segmentation of a diffuse low-grade glioma MRI dataset. *38th conference of the ieee engineering in medicine and biology society*.

Bernard, O., Bosch, J. G. & D'hooge, J. (2014). Left Ventricle Segmentation in 3D Echocardiography Algorithm Evaluation Platform.

Bland, J. M. & Altman, D. G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet*, 1, 307 -310. doi : 10.1016/S0140-6736(86)90837-8.

Chaibi, Y. (2010). Adaptation des méthodes de reconstruction 3D rapides par stéréoradiographie : Modélisation du membre inférieur et calcul des indices cliniques en présence de déformation structurale. *Sciences-new york*, 220.

Chalana, V. & Kim, Y. (1997). A methodology for evaluation of boundary detection algorithms on medical images. *Ieee transactions on medical imaging*, 16(5), 642–652. doi : 10.1109/42.640755.

Chartrand, G. (2016). Segmentation 3D du Foie. *Thèse de l'École de technologie supérieure, montréal, canada*.

Chav, R., Cresson, T., Kauffmann, C. & de Guise, J. A. (2009). Method for fast and accurate segmentation processing from prior shape : application to femoral head segmentation on x-ray images. *Spie medical imaging 2009 : Image processing*, 7259(28720), 72594Y–72594Y–8. doi : 10.1117/12.812459.

Chav, R., Cresson, T., Chartrand, G., Kauffmann, C., Soulez, G. & de Guise, J. A. (2014). Kidney Segmentation from a Single Prior Shape in MRI. 818–821.

Chen, C., Xie, W., Franke, J., Grutzner, P. A., Nolte, L. P. & Zheng, G. (2014). Automatic X-ray landmark detection and shape segmentation via data-driven joint estimation of image displacements. *Medical image analysis*, 18(3), 487–499. doi : 10.1016/j.media.2014.01.002.

Commowick, O., Akhondi-asl, A. & Warfield, S. K. (2012). Estimating a reference standard segmentation with spatially varying performance parameters : Local MAP STAPLE. *Ieee transactions on medical imaging*, 31(8), 1593–1606.

Commowick, O. & Warfield, S. K. (2010a). Estimation of inferential uncertainty in assessing expert segmentation performance from STAPLE. *Ieee transactions on medical imaging*, 29(3), 771–780. doi : 10.1109/TMI.2009.2036011.

Commowick, O. & Warfield, S. K. (2010b). Incorporating priors on expert performance parameters for segmentation validation and label fusion : A maximum a posteriori STAPLE. *Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics)*, 6363 LNCS(PART 3), 25–32. doi : 10.1007/978-3-642-15711-0_4.

Diop, E. & Burdin, V. (2013). Bi-planar image segmentation based on variational geometrical active contours with shape priors. *Medical image analysis*, 17, 165–181. doi : 10.1016/j.media.2012.09.006.

Heimann, T. et al. (2009). Comparison and evaluation of methods for liver segmentation from CT datasets. *Ieee transactions on medical imaging*, 28(8), 1251–1265. doi : 10.1109/TMI.2009.2013851.

Huttenlocher, D., Klanderman, D. & Ruckligem, A. (1993). Comparing images using the hausdorff distance. *Ieee trans. pattern anal. mach. intell.*

Jannin, P., Fitzpatrick, J. M., Hawkes, D. J., Pennec, X., Shahidi, R. & Vannier, M. W. (2002). Validation of medical image processing in image-guided therapy. *Ieee transactions on medical imaging*, 21(12), 1445–1449. doi : 10.1109/TMI.2002.806568.

Khooshabi, G. S. (2013). *Segmentation Validation Framework*. (Mémoire de maîtrise, Linköping University). Repéré à http://liu.diva-portal.org/smash/get/diva2:631210/FULLTEXT02.pdf.

Landman, B. A. & Warfield, S. K. (2012). Workshop on multi-atlas labeling. *Miccai*.

Ouertani, F., Vazquez, C., Cresson, T. & de Guise, J. (2015). Simultaneous Extraction of Two Adjacent Bony Structures in X-Ray Images : Application to Hip Joint Segmentation. *Ieee conference on image processing*.

Prastawa, M., Bullitt, E. & Gerig, G. (2005). Synthetic ground truth for validation of brain tumor MRI segmentation. *Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics)*. doi : 10.1007/11566465_4.

Sun, S., Zhang, B., Meng, S., Liu, D. & Sun, J. (2012). An improved interactive segmentation method for extracting the edge features of femur digital radiographs. *International society for optics and photonics*, 8335, 421–424. doi : 10.1117/12.917605.

Udupa, J. K., LeBlanc, V. R., Zhuge, Y., Imielinska, C., Schmidt, H., Currie, L. M., Hirsch, B. E. & Woodburn, J. (2006). A framework for evaluating image segmentation algorithms. *Computerized medical imaging and graphics*, 30, 75–87. doi : 10.1016/j.compmedimag.2005.12.001.

Warfield, S. K., Zou, K. H. & Wells, W. M. (2004). Simultaneous truth and performance level estimation (STAPLE) : An algorithm for the validation of image segmentation. *Ieee transactions on medical imaging*, 23(7), 903–921. doi : 10.1109/TMI.2004.828354.

Weir, J. P. (2005). Quantifying test-retest reliability using the intraclass correlation coefficient and the sem. *Journal of strength and conditioning research*, 19(1), 231–240.

Zheng, G., Li, S. & Belavi, D. (2016). Automatic Intervertebral Disc Localization and Segmentation from 3D Multi-modality MR (M3) Images.

Zhou, L., Chav, R., Cresson, T., Chartrand, G. & de Guise, J. A. (2016). 3D Knee Segmentation Based on Three MRI Sequences from Different Planes. *38th conference of the ieee engineering in medicine and biology society*.