Segmentation des organes à risque en imagerie scanner par architecture SharpMask et CRF

Roger TRULLO[1,2], Caroline PETITJEAN[1], Bernard DUBRAY[1], Dong NIE[2], Dinggang SHEN[1], Su RUAN[2]

[1]Normandie Univ, UNIROUEN, UNIHAVRE, INSA Rouen, LITIS, 76000 Rouen, France

[2]Department of Radiology and BRIC, UNC-Chapel Hill, USA

rogertrullo@hotmail.com

Résumé – La radiothérapie est un traitement de choix pour le cancer. La première étape consiste à identifier, sur les images scanner, le volume cible et les organes à risque (OAR) sains à protéger des irradations. Contrairement aux approches existantes pour la segmentation automatique des OAR qui utilisent de l'information locale et segmentent individuellement chaque OAR, nous proposons un cadre basé sur de l'apprentissage profond pour la segmentation conjointe des OAR dans des images scanner du thorax, dont le cœur, l'œsophage, la trachée et l'aorte. En nous appuyant sur les Fully Convolutional Networks (FCN), nous présentons plusieurs extensions qui améliorent les performances, dont une nouvelle architecture qui permet d'utiliser les caractéristiques profondes et superficielles, en combinant de manière efficace l'information locale et globale pour améliorer la précision de la localisation. L'utilisation de champs aléatoires conditionnels (en particulier le modèle CRF as Recurrent Neural Network) permet de prendre en compte les relations entre les organes pour raffiner encore les résultats de segmentation. Les expériences démontrent une performance compétitive sur une base d'images scanner de 30 examens.

Abstract – Radiotherapy is a standard treatment for cancer and the first step of the radiotherapy process is to identify the target volumes to be targeted and the healthy organs at risk (OAR) to be protected. Unlike previous methods for automatic segmentation of OAR that typically use local information and individually segment each OAR, in this paper, we propose a deep learning framework for the joint segmentation of OAR in CT images of the thorax, specifically the heart, esophagus, trachea and the aorta. Making use of Fully Convolutional Networks (FCN), we present several extensions that improve the performance, including a new architecture that allows to use low level features with high level information, effectively combining local and global information for improving the localization accuracy. Finally, by using Conditional Random Fields (specifically the CRF as Recurrent Neural Network model), we are able to account for relationships between the organs to further improve the segmentation results. Experiments demonstrate competitive performance on a dataset of 30 CT scans.

1 Introduction

La radiothérapie est un traitement de choix pour le cancer, notamment dans le cas des cancers du poumon et de l'œsophage. La planification de l'irradiation commence par segmenter la tumeur cible et les organes sains localisés à proximité, appelés Organes à Risque (OAR). En routine clinique, la délineation est largement manuelle, ce qui est fastidieux et chronophage. Dans le cas de l'œsophage, sa forme et sa position varient fortement d'un patient à l'autre et ses frontières, dans les images scanner, manquent singulièrement de contraste, et sont parfois même invisibles (Fig. 1).

La segmentation automatique des OAR a déjà été abordée dans la litérature, notamment dans le domaine de l'imagerie pelvienne [1, 2]. Dans [1], pour segmenter 17 organes à risque dans tout le corps, les auteurs utilisent 3 techniques différentes en fonction des organes : seuillage, transformée de Hough généralisée (GHT) et recalage d'atlas. Les auteurs de [3] proposent de combiner du recalage déformable multi-atlas avec une recherche locale par ensembles de niveaux pour segmenter plusieurs organes, dont l'aorte, l'œsophage, la trachée et le cœur. Ces travaux reposent sur des étapes de prétraitement

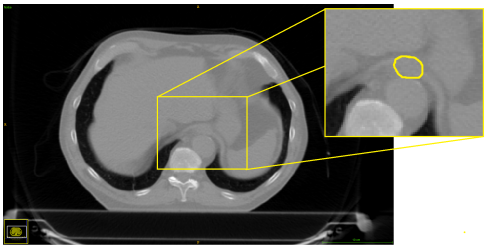


FIGURE 1: Image scanner où l'œsophage est segmenté manuellement en jaune. On note qu'il est difficile à distinguer.

assez lourdes pouvant représenter une forte charge de calculs. Surtout, ces travaux appliquent une segmentation de chaque organe de manière individuelle, qui ignore les informations de relations spatiales entre les organes.

Récemment, les architectures d'apprentissage profond sont apparues comme des alternatives performantes aux méthodes traditionnelles en vision par ordinateur ; en segmentation sémantique, l'approche pionnière dans ce domaine est celle des Fully Convolutional Networks (FCN) [4]. Plusieurs extensions des FCN [5, 6] ont souligné que cette architecture tend à igno-

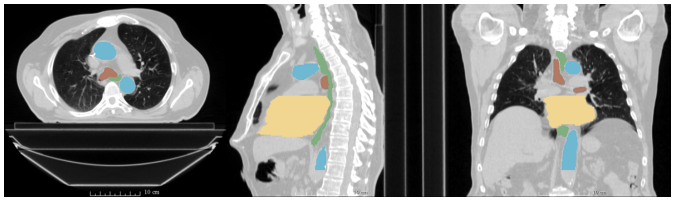


FIGURE 2: Image scanner avec segmentation manuelle de l'œsophage (vert), du cœur (jaune), de la trachée (marron) et de l'aorte (bleu).

rer la cohérence spatiale, ce qui produit des résultats assez grossiers en sortie.

Pour pallier les difficultés mentionnées, nous proposons d'utiliser un cadre d'apprentissage profond pour réaliser la segmentation conjointe des 4 organes à risque thoraciques (œsophage, cœur, trachée, aorte, cf Fig. 2) avec l'hypothèse que, ces organes étant voisins, ils peuvent fournir des informations complémentaires qui peuvent aider le réseau à apprendre les relations spatiales entre organes, augmentant ainsi les performances. Nous pensons que ces relations spatiales peuvent être apprises en combinant des caractéristiques bas niveau (des premières couches) avec des caractéristiques haut niveau (des couches plus profondes), comme ce qui est fait dans le réseau SharpMask. [7]. Nous montrons expérimentalement la supériorité de cette architecture, pour notre cadre applicatif, en comparaison avec l'architecture linéaire FCN. De plus, avec l'objectif d'améliorer le manque de cohérence spatiale implicitement présent en sortie du FCN et de renforcer les relations spatiales entre les organes, nous proposons d'utiliser les champs aléatoires conditionnels (CRF) comme étape de raffinement. Contrairement à d'autres travaux [8], nous utilisons l'architecture CRFasRNN [6] car il permet à l'opération de faire partie du réseau, rendant ainsi le système entrainable de bout en bout [9]. Nos expériences montrent que notre approche est supérieure à l'architecture FCN standard, à un FCN associé à un CRF, et à une méthode multi-atlas basée sur des patches. Nous pensons que ce travail représente une avancée vers la segmentation automatique conjointe des organes à risque pour le cancer du poumon, une application particulièrement délicate.

2 Méthode

2.1 Architecture FCN et SharpMask

Dans l'architecture FCN que nous allons utiliser, comme architecture de référence, les 5 premières couches sont des convolutions, des opérations de correction ReLU et de max pooling. Les 5 dernières sont composées de convolutions transposées (ou déconvolutions) avec des opérations ReLU. La dernière couche contient 5 canaux représentant les cartes de probabilité pour chacun des 4 organes et le fond. La fonction de perte est de type entropie croisée calculée à chaque voxel. Cette architecture est similaire à celle présentée dans [10], mais nous n'utilisons pas d'opérations de unpooling.

Les opérations de pooling sont intéressantes pour des tâches

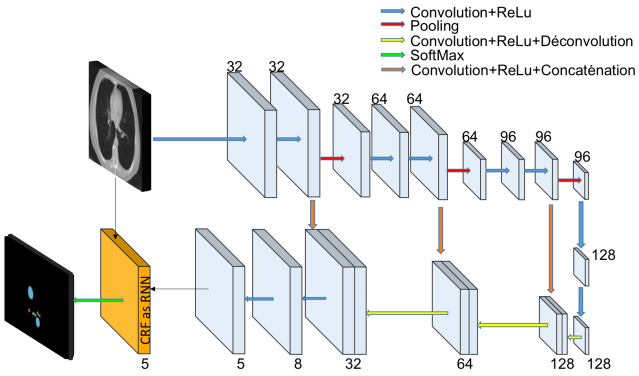


FIGURE 3: Architecture proposée pour la segmentation multi-organes. Les nombres indiquent le nombre de canaux dans chaque couche.

de classification mais pas en segmentation sémantique, de par la baisse de résolution induite. On peut contrer cet effet en combinant les caractéristiques des premières couches avec des caractéristiques des couches profondes, ce qui permet d'améliorer la précision de localisation en segmentation et d'alléger le problème de dissipation du gradient vu que les erreurs des couches profondes seront insérées dans les premières couches. Cet aspect a été exploité dans le U-Net [11], et plus récemment dans le SharpMask (SM) [7]. Dans le présent travail nous avons utilisé une architecture SM (Fig. 3) – mais au lieu d'utiliser du suréchantillonnage bilinéaire dans le module de raffinage, nous avons utilisé les convolutions transposées ce qui donne plus de paramètres et donc plus de capacité au réseau ; de plus, nous avons ajouté un module CRF après le réseau, qui est détaillé dans la section suivante. Au lieu de recopier les carte de caractéristiques comme dans [11], les opérations de convolution+ReLU sont utilisées pour avoir le même nombre de canaux que ceux du niveau profond, ce qui évite au réseau d'être biaisé dû à une trop grande différence entre le nombre de canaux des couches superficielles et profondes.

2.2 Champs aléatoires conditionnels "as RNN"

Les champs aléatoires conditionnels (CRF) sont utilisés en complément des classifieurs pour améliorer la cohérence des étiquettes, servis par une implémentation efficace [5, 12]. Pour l'inférence, on peut utiliser une approximation en champ moyen [12] qui implique un algorithme itératif.

Soit $\boldsymbol{x}$ l'étiquette, l'énergie de Gibbs est définie comme dans [5] par :

$$E(\boldsymbol{x}) = \sum_i \psi_u(x_i) + \sum_{ij} \psi_p(x_i, x_j) \qquad (1)$$

où $\psi_u(x_i)$ est le terme unaire, typiquement défini par $-logP(x_i)$, i.e. la log-vraisemblance au voxel i. Le terme de couple $\psi_p(x_i, x_j) = \mu(x_i, x_j)\kappa(\boldsymbol{f}_i, \boldsymbol{f}_j)$ mesure le coût d'affecter les étiquettes x_i, x_j simultanément aux voxels i et j, et $\kappa(\boldsymbol{f}_i, \boldsymbol{f}_j) = \sum_{m=1}^{M} \omega^m \kappa^m(\boldsymbol{f}_i, \boldsymbol{f}_j)$. La fonction μ est une fonction de compatibilité et chaque $\kappa^m(\boldsymbol{f}_i, \boldsymbol{f}_j)$ est un noyau gaussien entre 2 vecteurs de caractéristiques. Nous adoptons les potentiels standards suivants, définis en termes

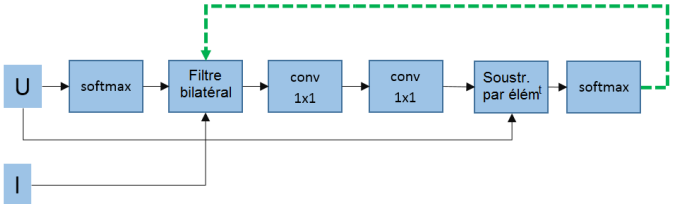


FIGURE 4: Architecture du CRFasRNN. U : terme unaire, sortie non normalisée du FCN. I : image d'entrée.

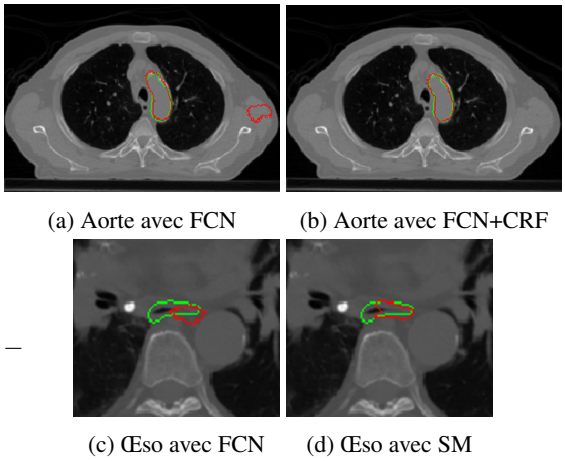
d'intensité I_i et I_j et de position $\boldsymbol{p}_i$ et $\boldsymbol{p}_j$: $\kappa(\boldsymbol{f}_i, \boldsymbol{f}_j) = \omega^1 exp(-\frac{|\boldsymbol{p}_i-\boldsymbol{p}_j|^2}{2\sigma_\alpha^2} - \frac{|I_i-I_j|^2}{2\sigma_\beta^2}) + \omega^2 exp(-\frac{|\boldsymbol{p}_i-\boldsymbol{p}_j|^2}{2\sigma_\theta^2})$. Une approximation permettant d'inférer le CRF par opérations rétro-propagables a récemment été présentée dans [6] ce qui permet de l'intégrer à l'architecture du réseau et d'entraîner celui-ci de bout en bout. Cette approche n'est plus une étape séparée mais une couche dont on peut apprendre les paramètres, comme le montre l'intégration après le réseau SM à la Fig. 3 (en orange). La rétropropagation est illustrée en pointillé vert dans la Fig. 4 implementé comme un réseau de neurones récurrent (RNN) appris avec une rétropropagation à travers le temps (BPTT). Cette architecture permet d'apprendre les paramètres ω^m et la fonction de compatibilité μ. Typiquement le modèle de Potts est utilisé [8] ; mais il est limité par le fait qu'il donne une pénalité fixe à des voxels similaires ayant des affectations d'étiquettes différentes. L'apprentissage de la fonction de compatibilité permet de pénaliser différement les affectations à des paires de voxels, d'après leur apparence et position [6].

3 Expériences

Dans notre réseau SM, nous utilisons des noyaux plutôt grands (7×7), conformément à d'autres travaux sur les images CT [8]. En ce qui concerne le raffinage du résultat avec le CRF, l'apprentissage est fait en 2 étapes ; d'abord nous entraînons un modèle sans le module CRF et ensuite nous entraînons un nouveau système avec le réglage des poids de l'ensemble, y compris le CRF. Cette étape est nécessaire parce que comme le montrent l'Eq. (1) et la Fig. 4, nous avons besoin d'un terme unaire qui représente une distribution de probabilité (non-normalisée) des étiquettes, pour effectuer l'inférence.

3.1 Caractéristiques des images et prétraitement

Nous évaluons notre méthode sur une base d'images composée de 30 volumes CT thoraciques de patients ayant un cancer pulmonaire ou un lymphome de Hodgkin, dans lesquels les contours des organes d'intérêt et du corps sont tracés manuellement. Ceux du corps sont utilisés dans le réseau pour éviter d'utiliser les informations de fond durant l'apprentissage. La résolution des images est de 0.98×0.98×2.5 pour une taille de 512×512×(150∼284) voxels. Chaque image est normalisée pour présenter une moyenne nulle et une variance unitaire. Le protocole consiste en une validation croisée à 6 partitions, soit 25 sujets pour l'apprentissage et 5 pour le test. La base d'images est augmentée par application d'une transformation affine aléatoire et d'une transformation obtenue par déformation aléatoire d'une grille de points de contrôle 2×2×2 et interpolation B-spline [13].



(a) Aorte avec FCN (b) Aorte avec FCN+CRF

(c) Œso avec FCN (d) Œso avec SM

FIGURE 5: Résultats de segmentation pour l'aorte (a,b) et l'œsophage (c,d). Vert : vérité terrain, rouge : sortie du système.

3.2 Apprentissage

Les classes sont déséquilibrées, i.e., le nombre de voxels est très différent d'un organe à l'autre. Nous avons donc opté pour une fonction de perte d'entropie croisée pondérée par les effectifs des classes. Mais les résultats étant décevants, nous avons eu l'idée de raffiner l'apprentissage des poids en utilisant l'entropie croisée standard sans pondération, dans un deuxième temps, et nous avons constaté que cela permettait d'améliorer les résultats. L'optimisation est ensuite basée sur une descente de gradient stochastique (SGD) avec un taux d'apprentissage de 0.1 divisé par 10 tous les 20 époques. Les poids sont initialisés par l'algorithme Xavier [14].

3.3 Résultats

Nous comparons les résultats de l'architecture proposée avec une architecture standard FCN, et son extension avec un CRF (FCN+CRF), et une méthode multi-atlas 3D à base de patchs, à l'état de l'art, appelée OPAL [15]. La Fig. 5 montre un exemple de résultat de segmentation pour l'aorte et l'œsophage. Nous pouvons voir que le FCN+CRF a éliminé la région des faux positifs à droite de l'image, et aussi qu'il permet de surpasser le simple FCN, notamment pour la segmentation de l'œsophage, qui est l'organe le plus difficile à segmenter. Nous pouvons observer que la localisation est plus précise que celle offerte par le FCN standard. Cela est dû à la fusion des caractéristiques haut et bas niveau pour alléger la perte de précision due aux opérations de pooling, qui sont encore nécessaires pour créer des représentations sémantiques. Finalement, la Fig. 6 présente des résultats de segmentation pour les 4 organes et le rendu volumique pour un des sujets de test.

Dans le Tableau 1 sont présentés les scores de Dice moyens obtenus. Pour la trachée, entourée par des voxels clairs et caractérisée par une intensité sombre, la détection par un algorithme

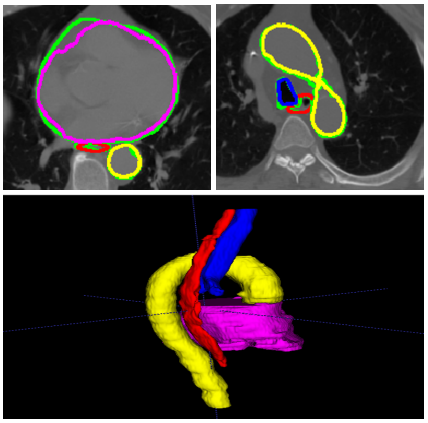


FIGURE 6: Résultats de segmentation pour les 4 organes. Vert : vérité terrain, sortie du système en rouge pour l'œsophage, magenta pour le cœur, bleu pour la trachée et jaune pour l'aorte. En bas : rendu volumique des segmentations.

TABLE 1: Scores de Dice moyens ± écart-type par différentes méthodes. 1ère ligne de p-valeur : comparaison à FCN ; 2nde ligne de p-valeur : comparaison à SM.

	OPAL	FCN	FCN+CRF	SM	SM+CRF
Œso	0.39±0.05	0.60±0.04	0.57±0.06	0.66±0.08	**0.67±0.04**
				$p = 0.13$	$p = 0.01$
					$p = 0.79$
Cœur	0.62±0.07	0.86±0.03	0.87±0.02	0.89±0.02	**0.90±0.01**
				$p = 0.07$	$p = 0.01$
					$p = 0.30$
Trachée	0.80±0.03	0.72±0.03	0.74±0.02	**0.83±0.06**	0.82±0.06
				$p = 0.01$	$p = 0.01$
					$p = 0.78$
Aorte	0.49±0.10	0.83±0.06	0.81±0.08	0.85±0.06	**0.86±0.05**
				$p = 0.58$	$p = 0.37$
					$p = 0.76$

de type OPAL en est ainsi facilitée, car celui-ci va essayer de trouver des patchs similaires dans des zones proches. Comme on le voit, OPAL réussit bien pour cet organe. La situation est différente pour des organes avec une forte variation d'intensité, pour les organes à faible contraste comme l'œsophage, elle s'inverse, comme le montre le Tab. 1. Les meilleures performances pour chaque organe sont obtenues par l'architecture SM et la différence avec FCN est significative au sens d'un test de Student : $p < 0.05$ pour tous les organes – sauf l'aorte. Il est intéressant de noter que le module de raffinement CRF n'apporte pas d'amélioration significative, que ce soit avec l'architecture FCN ou SM (comment on peut le voir d'après les p-valeurs dans le dernier cas).

4 Conclusions

Nous avons présenté une approche de segmentation conjointe pour l'œsophage, le cœur, la trachée, et l'aorte dans les images scanner. La méthode utilise l'augmentation artificielle de données et une architecture SharpMask pour permettre de combiner efficacement des caractéristiques bas-niveau avec d'autres plus riches sémantiquement, et d'effectuer ainsi une fusion des caractéristiques multiéchelle. Le CRFasRNN permet de renforcer les relations spatiales entre les organes, avec un résultat mitigé dans le cadre de notre application. Nous explorons actuellement de nouveaux moyens d'incorporer du contexte spatial, inspirés notamment par le modèle d'auto-contexte.

Remerciements Ce travail est co-financé par l'Union Européenne via le Fond de Développement Régional Européen (ERDF, HN0002137) et par le Conseil Régional de Normandie via le projet M2NUM.

5 Références

[1] M Han et al., "Segmentation of organs at risk in ct volumes of head, thorax, abdomen, and pelvis," in *Proc. SPIE*, 2015, vol. 9413, pp. 94133J–94133J–6.

[2] M Guinin et al., "Segmentation of pelvic organs at risk using superpixels and graph diffusion in prostate radiotherapy," in *ISBI*, 2015, pp. 1564–1567.

[3] E Schreibmann et al., "Multiatlas segmentation of thoracic and abdominal anatomy with level set-based local search," *J Appl Clin Med Phys*, vol. 15, no. 4, 2014.

[4] J Long et al., "Fully convolutional networks for semantic segmentation," in *CVPR*, 2015.

[5] L.-C Chen et al., "Semantic image segmentation with deep convolutional nets and fully connected CRFs," in *ICLR*, 2015.

[6] S Zheng et al., "Conditional random fields as recurrent neural networks," in *ICCV*, 2015.

[7] P. H. O Pinheiro et al., "Learning to refine object segments," *CoRR*, vol. abs/1603.08695, 2016.

[8] Q Dou et al., "3D deeply supervised network for automatic liver segmentation from CT volumes," *CoRR*, vol. abs/1607.00582, 2016.

[9] R Trullo, C Petitjean, S Ruan, B Dubray, D Nie, and D Shen, "Segmentation of organs at risk in thoracic CT images using a sharpmask architecture and conditional random fields," in *IEEE ISBI*, 2017.

[10] H Noh et al., "Learning deconvolution network for semantic segmentation," in *ICCV*, 2015.

[11] O Ronneberger et al., "U-net : Convolutional networks for biomedical image segmentation," in *MICCAI*, 2015, vol. 9351, pp. 234–241.

[12] P Krähenbühl and V Koltun, "Efficient inference in fully connected CRFs with gaussian edge potentials," in *NIPS*, 2011.

[13] F Milletari et al., "V-net : Fully convolutional neural networks for volumetric medical image segmentation," *CoRR*, vol. abs/1606.04797, 2016.

[14] X Glorot and Y Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *AISTATS*, 2010.

[15] V.-T Ta et al., "Optimized patchmatch for near real time and accurate label fusion," in *MICCAI*, 2014, pp. 105–12.