A Fusion Method Based on Deep Learning and Case-Based Reasoning which Improves the Resulting Medical Image Segmentations

Lisa Corbat[a,*], Mohammad Nauval[a], Julien Henriet[a], Jean-Christophe Lapayre[a]

[a] *FEMTO-ST Institute, DISC, CNRS, Univ. Bourgogne-Franche-Comté, 16 route de Gray, 25030 Besançon, France*

Abstract

The fusion of multiple segmentations of different biological structures is inevitable in the case where each structure has been segmented individually for performance reasons. However, when aggregating these structures for a final segmentation, conflicting pixels may appear. These conflicts can be solved by artificial intelligence techniques. Our system, integrated into the SAIAD project, carries out the fusion of deformed kidneys and nephroblastoma segmentations using the combination of Deep Learning and Case-Based Reasoning. The performances of our method were evaluated on 9 patients affected by nephroblastoma, and compared with other AI and non-AI methods adapted from the literature. The results demonstrate its effectiveness in resolving the conflicting pixels and its ability to improve the resulting segmentations.

Keywords: Fusion, Conflict management, Segmentation, Cancer tumour, Deep Learning, Case-Based Reasoning

*Corresponding author, email address: lisa.corbat@univ-fcomte.fr

Email addresses: lisa.corbat@univ-fcomte.fr (Lisa Corbat), mohammad.nauval@edu.univ-fcomte.fr (Mohammad Nauval), julien.henriet@univ-fcomte.fr (Julien Henriet), jean-christophe.lapayre@univ-fcomte.fr (Jean-Christophe Lapayre)

1. Introduction

The Wilms tumour, also called Nephroblastoma, is one of the most frequent abdominal tumours observed in children (generally in 1 to 5 years of age), representing 5 to 14% of malignant paediatric tumours. This type of tumour is situated in the kidney. Most often, its initial diagnosis is based on imaging. Generally, ultrasound observations are first planned in order to confirm the tumour's existence and to approximate its position. A medical scan then locates it with greater accuracy, along with affected organs and healthy tissues. Radiologists and surgeons need 3-Dimensional (3D) representations of the tumour and the border organs in order to establish the diagnosis, plan the surgery (estimated quantity of blood, specialised equipment required, estimation of the duration of the surgery, etc.) and also guide the actions of the surgeon during the surgery. This 3D representation is currently done through manual segmentations, which is a long and time-consuming task.

The French-Swiss border project SAIAD (Automated Segmentation of Medical Images Using Distributed Artificial Intelligence) aims at obtaining automatic segmentations of the nephroblastoma and other structures of the abdomen, through Artificial Intelligence (AI) methods. In the SAIAD platform, each structure is segmented separately by a technique of its own. Currently, the tumour is segmented by Deep Learning (DL) and more specifically by a Convolutional Neural Network (CNN) (Marie et al. (2019)), and the pathological kidney is segmented by Cased-Based Reasoning (CBR) coupled with region growing technique (Marie et al. (2018)), or by the same approach used for the segmentation of the tumour. However, once the segmentation step is done for each structure, the different segmentations must be aggregated together to obtain a final consensus segmentation. This aggregation is not obvious to realise, because when the different segmentations are superposed, some areas of conflicting segmentations can appear on the labelled pixels belonging to the different structures.

In this paper, we propose an original approach based on the use of AI tech-

niques to achieve the fusion of complementary segmentations of scanner images of deformed kidneys and tumours. The initial hypothesis we assumed in this study and the particular application domain of SAIAD is that similar segmentations lead to similar fused segmented images. Consequently, we evaluated the possibility of using a reasoning technique based on analogy (CBR) in order to enhance the fusion by deep learning, and introduce a hierarchical scheme on the set of cases to be used during the CNN's learning stage.

The following sections of the paper are organised as follows. Section 2 discusses related works in the field of segmentations fusion by mathematical and AI techniques. Section 3 presents our method of fusion of complementary segmentations by DL coupled with CBR approach and Section 4 explains the experiments and performances of our method. Finally, Sections 5 and 6 presents the discussions and the conclusions of the paper.

2. Related works

The combination of multiple segmentations can be done in different ways. It is, most of the time, used to achieve a better segmentation through the combination of multiple coarse segmentations, instead of using a more expensive and complex segmentation method. Intuitive methods can be realised for the combination of these segmentations, such as the use of majority vote (Rohlfing et al. (2004)) or the intersection and the union (Cabria & Gondra (2017)). However, these methods are limited. Many methods have emerged with the use of different metrics, via an iterative algorithm, for merging segmentations, such as using the Variation of Information (VoI) criterion (Mignotte (2014); Nguyen et al. (2018)), the F-Measure (or precision-recall criterion)(Mignotte & Hélou (2014)), the Global Consistency Error (GCE) (Khelifi & Mignotte (2016)) or the Probabilistic Rand Index (PRI) measure (Mignotte (2010)). More recently, other fusion approaches use the combination of multiple metrics like the VoI and the F-Measure criteria and the GCE and the F-Measure criteria (Khelifi & Mignotte (2017a,b)). Another approach is the use of spatial and intensity

information (like the pixel's grey level and the neighbour's labels) of an image and its segmentations for the fusion of MR-T2 brain images segmentation map (Feng et al. (2017)).

Segmentations fusion is also often employed in another field of image processing tasks, such as saliency detection. Saliency detection aims to highlight and segment the most important or visually distinctive objects or regions in an image by extracting its discriminative features and then compute their importance in the image. The neural networks, deep or not, and more particularly the convolutional neural networks (whose notions are presented in the article of LeCun et al. (2015)) are increasingly used in saliency detection due to their ability to extract high-level features from an image. They, therefore, could theoretically outperform conventional saliency detection methods that are based on low-level saliency features. The result of saliency detection is a saliency map containing the segmentation of salient objects from an image. Thus, saliency detection can also be classified as an image segmentation task.

Many saliency detection methods have been proposed by combining two or more saliency maps in order to capture more diverse types of information and thereupon improve the segmentation in the final saliency map. Zhao et al. (2015) proposed a saliency detection method in which complimentary segmentations in two saliency maps are fused by a fully connected layer. This method falls within the category of region-level saliency detection in which an input image is first segmented into regions and its features are then extracted from these regions. On the other hand, pixel-level saliency detection approaches extract features directly from the pixels of an image. Li & Yu (2016) proposed a saliency detection approach that combines the segmentations from a pixel-level saliency map and a region-level saliency map using a single convolutional layer with a 1×1 kernel. The same fusion system is also proposed by Tang et al. (2016) to fuse segmentations from five saliency maps.

Increasingly deep networks are also proposed. A network is considered deep when it has several hidden layers. Thus, in the saliency detection method proposed by Tang & Wu (2016), the fusion of two segmentations, each from a pixel-

level saliency map and a region-level saliency map, is also used. The authors also incorporate the original image in the fusion process in view of providing more information and correcting the segmentation in the final saliency map. The fusion system's architecture consists of one concatenation layer and three convolutional layers.

Xiao et al. (2018) developed another saliency detection method that uses Recurrent Convolutional Neural Network to extract four saliency maps from an image that are later concatenated and fused by three convolutional layers where each one of them is followed by a Rectified Linear Unit (ReLU) layer. Another saliency method proposed by Qu et al. (2017) fuses multiple saliency feature maps by a more complex convolutional neural network and the resulting segmentation is passed through a Laplacian propagation to enforce a better spatial consistency in the final saliency map.

Some image segmentation methods also incorporate the fusion of two or more feature maps to achieve better segmentation results. A deep convolutional neural network consisting of seven convolutional layers, five ReLU layers, and two pooling layers is used to combine two feature maps in an image segmentation method proposed by Hu et al. (2019).

The merge processes shown above are either included in segmentation networks or are independent but used in the saliency maps fusion. Our approach is to create a deep neural network allowing to aggregate different structures between them and to indirectly manage the possible conflicts that could have happened by simply superimposing the segmentations. Also, our network is coupled to a simple CBR technique, allowing it to recover different segmentations from the case base for its learning process.

3. Proposed method

This part of the paper firstly presents the general architecture of the SAIAD project, and then the DL coupled with CBR system designed for the fusion of multiple complementary segmentations.

3.1. Overview of the general architecture of SAIAD project

Figure 1 shows the general architecture of the system designed in the SAIAD project. It is composed of three layers. The first one is the data layer, which contains a database for each segmentation system, and which is attached to a specific and secure server, called the COVOTEM collaborative platform (Maincare (2019)). When a new patient performs an examination, the slices of her/his abdomen are stored on the platform by the hospital that performed these examinations (Atlas part). These are scanned images in Dicom format. Experts then add the ground truths on the platform (Expert knowledge part). These ground truths allow us to measure the accuracies of the segmentations and fusions performed by our methods. Each database has access to all the CT scan images and expert knowledge. The second layer is the segmentation layer, where CT scan images are segmented by artificial intelligence systems, specifically by a CBR system coupled with region growing technique for the pathological kidney (Marie et al. (2018)) and by a Deep Learning segmentation, and more particularly by the FCN-8s for the nephroblastoma (Marie et al. (2019)). At the end of the segmentation processes, the system gives two complementary segmentations in two different images and the fusion layer combines them with a neural network coupled with a CBR technique planned for the fusion of the two segmentations.

3.2. DL coupled with CBR system

As summarised in Figure 2, the DL coupled with CBR system has several parts. CBR itself is a problem-solving method based on analogy relying on its past experiences to solve a new problem. Our fusion system has a case base allowing the storage of known cases, a retrieval phase allowing the recall of similar cases, a specific pre-processing for a case before being passed to the neural network input and finally the neural network for the fusion process. The details of each part are presented below. The principle of the system is as follows. When a new case (the two segmentations to fuse and the corresponding CT scan) is added, the system calculates its similarity value to each stored case

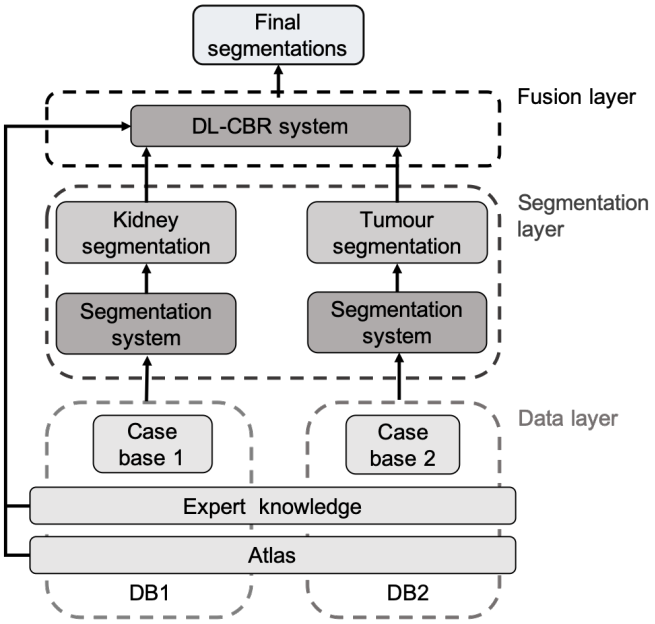

Figure 1: General schema of the SAIAD project.

and, during the retrieval phase, extracts a certain number of source cases with the highest similarities. Then, these cases are prepared for the input of the neural network and used for its training. Once the network is trained, the new case (previously prepared) is given at the input of the network for the calculation of the final result. The final result is then the final segmentation where tumour segmentation and pathological kidney segmentation are concatenated. This new case, with the solution, can be stored in the case base if the resulting segmentation is deemed relevant.

The input of the system (the input case) is the information of a transversal slice of a patient. Several transversal slices must be treated in order to obtain all the information about the patient's nephroblastoma. Thus, in order to achieve fusion of the kidney and nephroblastoma segmentations in their entirety, our process must be repeated on each segmented slice of the patient. Indeed, the segmentation techniques used segment the structures slice by slice. This allows

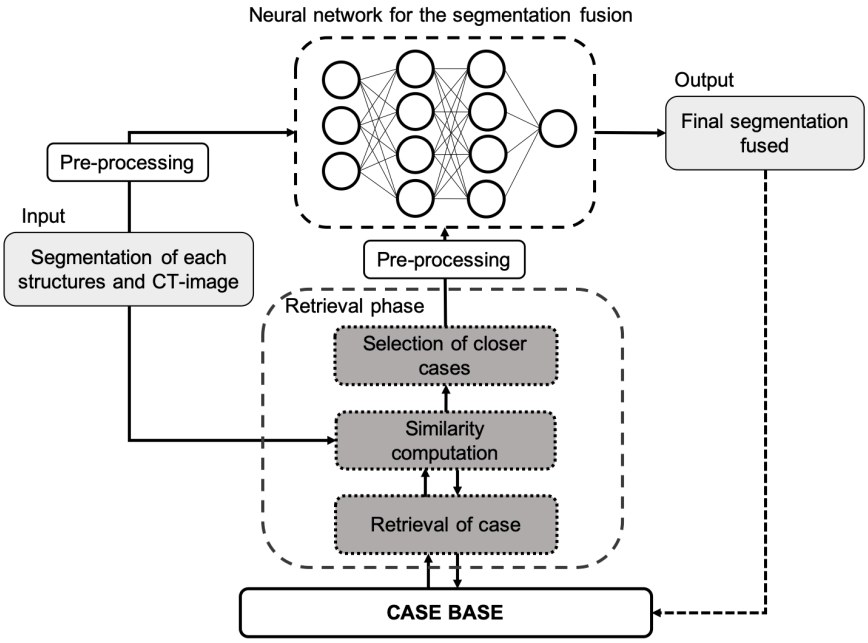

Figure 2: Overview of the DL-CBR system.

us, by superimposing these segmentations, to obtain the 3D segmentation containing all these structures.

3.2.1. Case model approach

A case is composed of a Problem part and a Solution part. The Problem part describes the characteristics of the problem to be solved and the Solution part describes the way to solve it. In our study, the Problem part contains the segmentations of each structure, where the superposition of these structures leads to conflicting pixel problems. The Solution part has the original CT scan image of the segmentations, and the ground truth of the segmentations (i.e. the true segmentations carried out by experts of the field). All these images will be used to train the neural network.

3.2.2. Retrieval phase

This section describes how the retrieval phase is performed using a similarity calculation between a new case and the stored cases. The case base is composed of two types of data: CT scan images and segmentation images. The similarity formula used, presented in Equation 1, is a mean between the Dice, which calculates the similarity between two segmentations S and S', described in Equation 2, and the MSSIM (Mean Structural SIMilarity) (Wang et al. (2004)) calculating the similarity between two CT scan images I and I', presented in Equation 3. The Dice metric is commonly used in the medical field and the MSSIM criteria is an improvement of the SSIM (Structural Similarity) criteria that are commonly used in image compression. MSSIM uses an iterative windowing to increase the capacity of structural comparison on images (i.e. each window describes a Region Of Interest which is compared independently).

$$s(S, S', I, I') = \frac{1}{2}(Dice(S, S') + MSSIM(I, I')) \tag{1}$$

The Dice is defined as:

$$Dice(S, S') = \frac{2 * TP_{S,S'}}{2 * TP_{S,S'} + FP_{S,S'} + FN_{S,S'}} \tag{2}$$

where $TP_{S,S'}$ is the number of true positive pixels between S and S', $FP_{S,S'}$ is the number of false positive pixels and $FN_{S,S'}$ the number of false negative pixels. The Dice's score ranges between 0 and 1. A score of 0 denotes that the two segmentations are completely different, whereas a score of 1 indicates that they are identical.

The MSSIM is defined as:

$$MSSIM(I, I') = \frac{1}{M}\sum_{i=1}^{M} SSIM(I_i, I'_i) \tag{3}$$

where M is the number of windows and SSIM is:

$$SSIM(I, I') = \frac{(2\mu_I\mu_{I'} + C_1) + (2\sigma_{II'} + C_2)}{(\mu_I^2 + \mu_{I'}^2 + C_1)(\sigma_I^2 + \sigma_{I'}^2 + C_2)} \tag{4}$$

where μ is the mean, σ is the standard deviation and $C_1 = 0.01 * L$ and $C_2 = 0.03 * L$ with $L = 255$. The values of C_1 and C_2 are the values used in the paper of Wang et al. (2004) and L is the range of possible pixel values.

As a case represents the segmentation of the tumour, the segmentation of the kidney and the CT scan image of the same slice, it is not possible to use Equation 1 directly, because it takes into account only one calculated segmentation (which will be compared to its ground truth). The two segmentations are then combined into a single segmentation with the addition of a new label for conflicting pixels. Likewise, their ground truths are also combined to obtain one ground truth that includes both structures. The Dice is then calculated globally on all the structures.

For the fusion of a new case, three known cases will be retrieved from the case base and will be used for the training of the neural network. This number has been defined by experimentation for the good learning of the network.

3.2.3. Pre-processing of the learning set

For the neural network to be able to process all three images (tumour segmentation, kidney segmentation, and CT scan image), they are concatenated into a three-channel image, where each channel corresponds to one image. This provides the network with all the information it needs in a single image.

3.2.4. The DL network

Architecture. For the specific purpose of fusing multiple complementary segmentations, a Deep Learning based system has been designed. The neural network, whose general architecture is presented in Figure 3, is inspired by the CNN used for image segmentation and more particularly by the networks used in Long et al. (2015). The network is composed of 8 convolutional layers and 2 pooling layers. It consists of the application of two 3x3 convolutions, each followed by a rectified linear unit (ReLU) and a 2x2 max pooling with a stride of 2 for downsampling, the whole repeated twice. Then, two other 3x3 convolutions without padding with ReLU are applied and followed by a 1x1 convolution in

order to obtain the desired number of classes. Finally, a 20x20 up-convolution (or deconvolutional layer) is used to get a result with the same size as the input of the network.

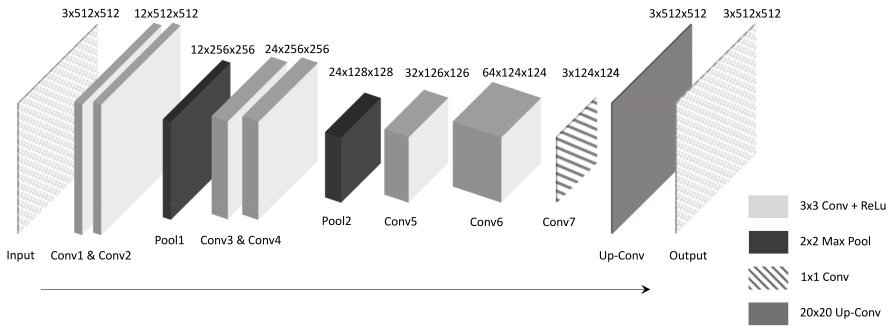

Figure 3: Overall architecture of the proposed network.

At the end of the predication phase, an argmax function is used in order to obtain a final segmentation fused with one channel.

Training stage. The training of the neural network is also specific to our DL-CBR system. For each case chosen during the retrieval phase (the two segmentations to fuse and their corresponding CT scan image), training is performed with a batch size of 1 and an automatic calculation of the number of iterations according to the similarity calculated between the new case and the chosen case (see Equation 1). The trainings are done in the increasing order of the similarities obtained. Thus, the chosen case with the lowest similarity will be used for the 1st training, with the network's weights initialised randomly, and the most similar case will be used at the end of the training process. We have chosen to use the most similar known case for the last training of our neural network, in order to specialise and focus our training on the case that is the most similar to our new case. Thus, with training for each known case, one iteration is equivalent to one epoch.

More precisely, the retrieval phase using the CBR technique gives a list of 3

known cases (C_1, C_2 and C_3) where C_1 is the most similar case, C_2 is the second most similar case, and C_3 is the third most similar case retrieved. These cases are given to the Deep Learning part of our method with their similarity values respectively equal to S_1, S_2 and S_3. These values are computed according to Equation 1. Then, the neural network's training is performed using the images of C_3 with the number of iterations fixed to $S_3 \times NBMaxIt$, where $NBMaxIt$ corresponds to a constant number of iterations fixed to 3 000. Then, this trained neural network is trained again using the images of C_2, with $S_2 \times NBMaxIt$ iterations. Finally, this two-timed trained network is trained for the third time using the images of C_1, with $S_1 \times NBMaxIt$ iterations.

This method of general training allows us to indirectly add the notion of weight on the whole learning set. The more a known case is similar to the new case to be fused, the higher the number of iterations of the neural network's training on that particular known case will be.

4. Experiments

4.1. Generation of segmentation set

We have tested the performance of our system over 9 patients with a mean of 109 slices per patient. All of the CT scan images and calculated segmentations have the same size: 512 x 512 pixels. The pathological kidney and the nephroblastoma were segmented on each slice of each patient by Deep Learning (the convolutional neural network FCN-8s first trained with the PASCAL VOC 2012 database in Everingham et al. (2015)) enhanced with the OVÅSSION protocol (Marie et al. (2019)) with 10000 training iterations. The learning protocol OVÅSSION is a protocol specifically designed for small learning sets, because the neural network is trained only on certain spaced slices of a patient and then it is tested on all the remaining slices.

A Conditional Random Field (CRF) post-processing (Krähenbühl & Koltun (2011)) is also applied after the Deep Learning segmentations. And before

the fusion process, the segmentation labels are modified to improve the neural network's efficiency.

The pixels belonging to the tumour in the tumour segmentation and the pixels belonging to the pathological kidney in the pathological kidney segmentation are initially labelled at a value of 1. They are then adjusted to 255 to compensate for the higher pixel values in the CT scan images.

We also have the ground truth segmentations of all the patients, carried out by experts (surgeons and radiologists) at the *Centre Hospitalier Régional Universitaire de Besançon* (University hospital of Besançon). These ground truths are used in order to verify the reliability of our processes and the resulting fused segmentations.

4.2. Training

For the fusion of all the segmentations of one patient, the case base is filled with slices from all the other patients, as well as some slices spaced from the current patient (with a gap of 4), which were used for training thanks to the OV^{2}ASSION protocol.

The deep neural network was trained on a maximum of 3000 iterations per case and a batch size of 1. It uses 3 cases for its training and a learning rate of 1e-9.

4.3. Results

Table 1 presents the percentage of conflicting pixels that have been correctly classified by the different tested approaches and the standard deviation for each approach. On average, the pixels in conflict on a slice correspond to 1.70% of its total number of pixels. Table 1 allows us to compare the performance of our DL-CBR method with other methods of fusion by neural networks (Our neural network with a classical training, three convolutional layers as proposed by Tang & Wu (2016), three convolutional layers each followed by a ReLU layer as proposed by Xiao et al. (2018) and a deep neural network as proposed by Qu et al. (2017)) and more classical methods by similarity (Feng et al. (2017))

and VoI-criterion (Mignotte (2014)). All other fusion methods based on neural networks have been subjected to classical trainings. They were trained on segmentation slices from all the other patients in order to fuse the segmentations of the current patient, with a learning rate of 1e-10 and the number of iterations is fixed at 20000.

Table 1: Percentage of correctly resolved conflicting pixels for each patient with different methods. The values in bold correspond to the best values obtained. From left to right: The different patients; The results of our DL-CBR system; The results of our DL system with a classical training; The results by the network of three convolutional layers; The results by the network of three convolutional layers followed by ReLU layers; The results by a deep neural network; The results by the use of similarity; The results by the use of VoI-criterion.

Patient	DL-CBR method	DL method	Network by Tang & Wu (2016)	Network by Xiao et al. (2018)	Network by Qu et al. (2017)	Method by Feng et al. (2017)	Method by Mignotte (2014)
P1	**0.7119**	0.6635	0.6324	0.5989	0.6493	0.6062	0.6782
P2	0.7316	**0.7623**	0.6839	0.6338	0.7371	0.5834	0.6681
P3	0.8589	0.4232	**0.8963**	0.7510	0.8050	0.6556	0.5104
P4	0.7158	0.6062	0.7269	0.3503	**0.7855**	0.6549	0.6650
P5	0.6391	0.5523	0.5699	0.5767	0.6195	0.5907	**0.7089**
P6	0.5123	0.6153	0.5228	0.5867	0.5392	**0.6201**	0.6044
P7	**0.7089**	0.5737	0.5653	0.6145	0.4696	0.3868	0.2846
P8	0.5484	0.6214	0.5083	0.5193	**0.6371**	0.5746	0.4759
P9	0.6146	0.5751	**0.6310**	0.6177	0.3983	0.6114	0.5697
Mean	**0.6713**	0.5992	0.6374	0.5832	0.6267	0.5871	0.5739
Std Dev (%)	10.52	9.08	12.08	10.70	13.89	**8.03**	13.44

First of all, we can see that there is no clear supremacy of one method over the others. Three methods give the best arbitration of conflicting pixels twice (the DL-CBR method, the network proposed by Tang & Wu (2016) and the network proposed by Qu et al. (2017)). Three other methods give the best arbitration once (the DL method, the method proposed by Feng et al. (2017) and the method proposed by Mignotte (2014)). Nonetheless, our DL-CBR method is one of the three best methods for seven of the nine patients. The performance of the DL-CBR method is relatively poor for patients P6 and P8 which are the

patients for which the segmentation system had the most difficulties to segment the pathological kidney correctly. Currently, for these two patients, on some of the slices, the pathological kidney is importantly deformed and thus, makes an automatic segmentation very difficult. Also, even if for each patient, our method is not necessarily the best, on average, it is superior to the other methods with 67.13% of correct resolution. The standard deviation of the DL-CBR method is 10.52%, which corresponds to the lowest value of the three methods giving it the best arbitration.

Tables 2 and 3 show the Dice scores obtained (between calculated segmentation and ground truth) for tumour and kidney, before and after the fusion according to the different methods and for each patient. This Dice score per patient is obtained by calculating the average of the Dice scores of all of the patient's slices. In both of the structures, most patients achieve better Dice (bold values) with our DL-CBR method. Evidently, for five of the patients, the final fused segmentation obtained with our method is the best one (patients P1, P4, P5, P6 and P9 for the tumour and patients P1, P3, P4, P7 and P9 for the kidney). DL-CBR method improves the Dice score before fusion of about 0.8% for the tumour (from 0.9120 up to 0.9193) and 1.2% for the kidney (from 0.8745 up to 0.8850). For the kidney, considering the mean Dice, the DL method outperforms the DL-CBR one (global mean Dice respectively equal to 0.8860 and 0.8850), but the DL-CBR method gives the best results for five patients and the DL method for only three patients. This supremacy of the DL-CBR and the DL methods can be explained by the fact that these methods (and methods by neural networks in general) do not only affect the set of pixels in conflict (As is the case for the methods by Feng et al. and by Mignotte), but all the pixels of a considered segmentation slice. These two methods (DL-CBR and DL) have the behaviour of an effective post-treatment on the entire slice. For that reason, even if one of the other methods gives a better arbitration strictly on the set of pixels in conflict (Table 1), the DL and DL-CBR methods can finally offer a better fused and final segmentation of each structure (Tables 2 and 3).

Table 2: Dice of the tumour before the fusion and obtained after the fusion for each patient according to different methods.

Patient	Dice before fusion	DL-CBR method	DL method	Network by Tang & Wu (2016)	Network by Xiao et al. (2018)	Network by Qu et al. (2017)	Method by Feng et al. (2017)	Method by Mignotte (2014)
P1	0.9422	**0.9493**	0.9422	0.9419	0.9403	0.9441	0.9427	0.9430
P2	0.8876	0.8930	**0.8957**	0.8751	0.8837	0.8903	0.8856	0.8866
P3	0.8298	0.8435	0.8245	0.8300	0.8288	**0.8543**	0.8273	0.8261
P4	0.9554	**0.9606**	0.9489	0.9527	0.9518	0.9560	0.9553	0.9554
P5	0.9338	**0.9436**	0.9403	0.9151	0.9080	0.9349	0.9336	0.9338
P6	0.9317	0.9344	**0.9383**	0.9324	0.9327	0.9355	0.9330	0.9330
P7	0.9233	**0.9269**	0.9255	0.9243	0.9225	0.9257	0.9225	0.9223
P8	0.9456	0.9465	0.9475	0.9447	0.9455	**0.9476**	0.9457	0.9455
P9	0.8585	**0.8758**	0.8721	0.8605	0.8676	0.8691	0.8606	0.8603
Mean	0.9120	**0.9193**	0.9150	0.9085	0.9090	0.9175	0.9118	0.9118
Std Dev (%)	4.35	3.96	4.28	4.30	4.13	**3.68**	4.40	4.43

Table 3: Dice of the pathological kidney before the fusion and obtained after the fusion for each patient according to different methods.

Patient	Dice before fusion	DL-CBR method	DL method	Network by Tang & Wu (2016)	Network by Xiao et al. (2018)	Network by Qu et al. (2017)	Method by Feng et al. (2017)	Method by Mignotte (2014)
P1	0.8690	**0.8918**	0.8861	0.8692	0.8661	0.8781	0.8694	0.8707
P2	0.9226	0.9246	**0.9269**	0.9245	0.9246	0.9243	0.9241	0.9245
P3	0.9422	**0.9442**	0.9429	0.9436	0.9431	0.9431	0.9424	0.9424
P4	0.9240	**0.9352**	0.9207	0.9254	0.9152	0.9257	0.9258	0.9259
P5	0.8581	0.8762	**0.8772**	0.8701	0.8552	0.8763	0.8597	0.8605
P6	0.8144	0.8263	**0.8493**	0.8184	0.8057	0.8321	0.8147	0.8134
P7	0.9081	**0.9103**	0.9078	0.9099	0.9094	0.9101	0.9086	0.9084
P8	0.6993	0.7084	0.7181	0.7067	0.7105	**0.7185**	0.7004	0.6986
P9	0.9331	**0.9476**	0.9448	0.9313	0.9448	0.9429	0.9335	0.9332
Mean	0.8745	0.8850	**0.8860**	0.8777	0.8750	0.8857	0.8754	0.8753
Std Dev (%)	7.79	7.66	7.04	7.55	7.67	**6.61**	10.52	10.72

Finally, Table 4 shows the mean Dice for each patient across the two structures (Dice obtained between the calculated segmentation and the ground truth). Again, our system provides better Dice scores for five patients and the final mean Dice. On average, a Dice for a patient increases by 0.88% with our method.

Table 4: Result of mean Dice (tumour and pathological kidney) before the fusion and obtained after the fusion for each patient according to different methods.

Structure	Dice before fusion	DL-CBR method	DL method	Network by Tang & Wu (2016)	Network by Xiao et al. (2018)	Network by Qu et al. (2017)	Method by Feng et al. (2017)	Method by Mignotte (2014)
P1	0.9056	**0.9206**	0.9141	0.9055	0.9032	0.9111	0.9061	0.9069
P2	0.9051	0.9088	**0.9113**	0.8998	0.9042	0.9073	0.9049	0.9056
P3	0.8860	0.8938	0.8837	0.8868	0.8860	**0.8987**	0.8849	0.8843
P4	0.9397	**0.9479**	0.9348	0.9390	0.9335	0.9409	0.9401	0.9407
P5	0.8959	**0.9099**	0.9088	0.8926	0.8816	0.9056	0.8967	0.8972
P6	0.8731	0.8804	**0.8938**	0.8754	0.8692	0.8838	0.8739	0.8732
P7	0.9157	**0.9186**	0.9167	0.9171	0.9160	0.9179	0.9156	0.9154
P8	0.8225	0.8274	0.8328	0.8257	0.8280	**0.8430**	0.8231	0.8221
P9	0.8958	**0.9117**	0.9085	0.8959	0.9062	0.9060	0.8971	0.8968
Mean	0.8933	**0.9021**	0.9005	0.8931	0.8920	0.9016	0.8936	0.8936
Std Dev (%)	6.41	6.17	5.85	6.17	6.23	**5.44**	9.29	9.44

Figure 4 shows the results of the segmentation fusion with conflict management of the DL-CBR system. It shows, in particular, the results obtained on five slices belonging to different patients. All pixels in conflict are labelled as belonging either to the tumour, kidney or background. The segmentations are also slightly improved, smoothed, and filled.

5. Discussion

The presented DL-CBR fusion system achieves the highest percentage in terms of conflicts management and segmentations improvement. But, the percentage of conflicting pixels that are correctly resolved, for each patient, is not relatively high, because of the difficulty of the conflicts to be solved. Most conflict zones are located at the intersection of the different structures. These areas remain very ambiguous and there is not always a clear delimitation between two structures, even for a radiologist, who then realises the contours by experience. Similarly, segmentation and fusion performed by neural networks are also based on the experience that they gained during their training. Nevertheless, it becomes very difficult to obtain a relatively high average percentage of correct resolution of conflicting pixels because of these many difficult areas.

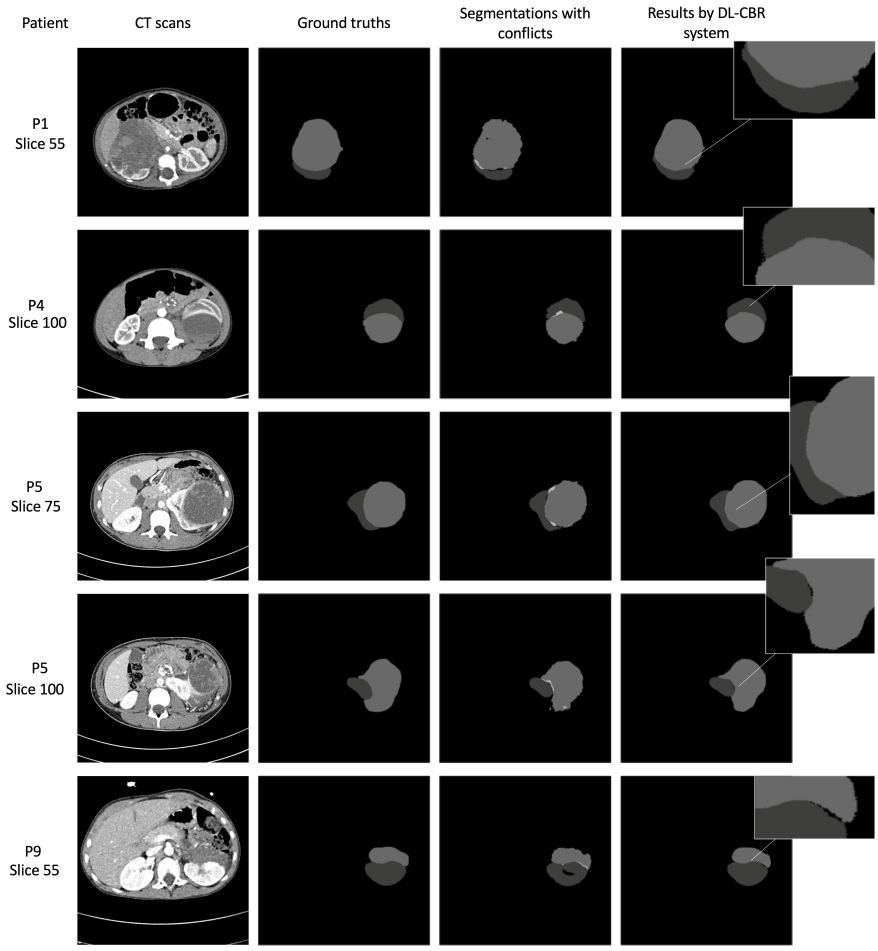

Figure 4: Results of the segmentations fusion. From left to right: CT scan images; Ground truths; Segmentations with conflicts; Result of the segmentations fusion by the DL-CBR system. For the three columns representing the segmentations, the pixels from darkest to brightest represent: the background; the pathological kidney; the tumour; and finally, the conflicting pixels in the third column.

The methods using neural networks achieve on average slightly higher results (cf. Table 1) than the conventional methods. This is because these methods can label a conflicting pixel as part of the background, which is not possible for the other two non-AI methods. Some pixels in conflict, find themselves labelled as background in the ground truth. These methods can also slightly modify the segmentations (smoothen and refine the edges of the segmented structures) as post-processing because they realise a new segmentation from the current segmentation. A deep neural network will then improve segmentations effectively. This also explains why in some cases, the percentage of correctly labelled pixels of our proposed method is low (and even lower than other methods) but that the Dice scores are better because the post-treatment has been very beneficial.

The depth of the neural network does not seem to have an impact on the conflicts management, as can be seen with the method of Tang (Tang & Wu (2016)) which gets a better percentage of correctly resolved pixels than the other neural networks based methods, with only three convolutional layers. However, deeper neural networks allow for better Dice scores, due to the general improvement of the segmentations.

The deeper neural network proposed by Qu et al. (2017) may seem more interesting compared to our neural network, especially for the Dice score obtained. However, for conflict management, the standard deviation is the highest compared to other methods, so it is the method with the most dispersed results. Moreover, if this network is coupled with our CBR technique, the system becomes much more time-consuming (about 24 hours for the fusion of a patient, against 7 hours for our current system). We, therefore, favoured a less complex network allowing a quicker fusion, while at the same time still obtaining significant results.

Neural network methods are more complex and time-consuming in terms of their execution times compared to more alternative approaches. Also, the deeper the networks, the more complex they are. The DL-CBR method is also a little longer due to the retrieval phase using the CBR technique that adds a few minutes of execution time. However, the calculation time of our method

is quite acceptable in our context. Even if the methods using neural networks process all the pixels of the images against only the conflicting pixels for the alternative mathematical methods, the differences in time and complexity are not proportional to the number of pixels to be processed in the segmentation. Thus, the neural networks that we tested perform a fusion in a few hours for the training and the validation on one patient (with images of 512x512 pixels), against ten minutes for the management of conflicts of one patient using alternative mathematical methods, only changing the conflicting pixels label (which represents less than 0.01 % of all pixels in the image).

For more precision and on the basis of He & Sun (2015), the complexity of our CNN can be summarized as the complexity of all the convolution layers, i. e:

$$O\left(\sum_{l=1}^{d} m_{l-1}.s_l^2.m_l.n_l^2\right) \tag{5}$$

Where l is the current convolutional layer, d the total number of convolutional layers, m_l is the number of output channels of the convolution l, and m_{l-1} the number of input channels of convolution l. s_l is the size of the filter, and n_l is the size of the output feature map. In this case, the complexity can be simplified in relation to the input image size n: $O\left(n^2.b\right)$, with b corresponding to the set of parameters which do not depend on the image size.

The execution times of these systems may seem long, but it must be taken into account that experts (surgeons and radiologists) are not monopolized by this calculation time. They can indeed devote themselves to other tasks and let the supercomputer carry out the segmentations and fusions. Moreover, we are not in a state of absolute emergency and experts believe that it is reasonable to obtain a numerical representation of one patient's abdomen in a few days or even a week. Currently, the presented patient segmentation process is done in 8 hours and this fusion process in 7 hours.

Regarding the problems in detecting the borders between different structures, our system could be improved by modifying the CT scan images. Specifically, it may be interesting to add filters on these images to bring out or highlight

the contours of their different structures. These modified CT scan images could then be used instead of the basic CT scan images to offer more information to the segmentation system.

As a general observation, our fusion method based on analogy is sometimes outperformed by other fusion methods. Consequently, it would be interesting to also take into account these other methods and find a way to select the most adapted one to each slice. The possibility of our system to automatically select the best fusion method from a few candidates will be considered and studied in further work.

6. Conclusion and further work

In this paper, we have presented a new DL-CBR fusion system for complementary segmentations of pathological kidneys and cancerous tumours. In average, for our set of patients, our method gives the best fused segmentations compared to the other fusion tested methods. The system is based on AI techniques with Deep Learning allowing it to carry out the fusion by experience thanks to the knowledge base provided by the CBR technique. In other words, reasoning by analogy (CBR) creates a hierarchy between the data of the CNN learning set.

For the fully automatic segmentation of tumorous kidneys, this method of fusion increases the robustness of the general system of the SAIAD project.

Anyway, since we observed that our method was outperformed by others in particular cases, we will study the possibility for the system to select automatically the best fusion method for each case.

Further work will also focus on improving the system by another pre-processing input images, improving the neural network and also adding functionalities in view of fusing new structures appearing on the scans, such as arteries, veins, and excretory cavities.

Acknowledgements

The authors wish to thank Professor F. Auber, Dr Y. Chaussy, and Dr M. Lenoir-Auber from the *BESANCON University Hospital* for their expertise in nephroblastoma, and Loredane Vieille and Elise Lacroix for performing the manual segmentations.

The authors would like to thank the European Union for financing this project as part of the SAIAD INTERREGV program and the SAIAD consortium partners.

Finally, computations have been performed on the supercomputer facilities of the Mésocentre de calcul de Franche-Comté.

References

Cabria, I., & Gondra, I. (2017). Mri segmentation fusion for brain tumor detection. *Information Fusion*, *36*, 1–9. doi:10.1016/j.inffus.2016.10.003.

Everingham, M., Eslami, S. A., Van Gool, L., Williams, C. K., Winn, J., & Zisserman, A. (2015). The pascal visual object classes challenge: A retrospective. *International journal of computer vision*, *111*, 98–136. doi:10.1007/s11263-014-0733-5.

Feng, Y., Shen, X., Chen, H., & Zhang, X. (2017). Segmentation fusion based on neighboring information for mr brain images. *Multimedia Tools and Applications*, *76*, 23139–23161. doi:10.1007/s11042-016-4098-3.

He, K., & Sun, J. (2015). Convolutional neural networks at constrained time cost. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 5353–5360). doi:10.1109/CVPR.2015.7299173.

Hu, Y., Soltoggio, A., Lock, R., & Carter, S. (2019). A fully convolutional two-stream fusion network for interactive image segmentation. *Neural Networks*, *109*, 31–42. doi:10.1016/j.neunet.2018.10.009.

Khelifi, L., & Mignotte, M. (2016). A novel fusion approach based on the global consistency criterion to fusing multiple segmentations. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, *47*, 2489–2502. doi:10.1109/TSMC.2016.2531645.

Khelifi, L., & Mignotte, M. (2017a). Efa-bmfm: A multi-criteria framework for the fusion of colour image segmentation. *Information Fusion*, *38*, 104–121. doi:10.1016/j.inffus.2017.03.001.

Khelifi, L., & Mignotte, M. (2017b). A multi-objective decision making approach for solving the image segmentation fusion problem. *IEEE Transactions on Image Processing*, *26*, 3831–3845. doi:10.1109/TIP.2017.2699481.

Krähenbühl, P., & Koltun, V. (2011). Efficient inference in fully connected crfs with gaussian edge potentials. In *Advances in neural information processing systems* (pp. 109–117).

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *nature*, *521*, 436. doi:10.1038/nature14539.

Li, G., & Yu, Y. (2016). Deep contrast learning for salient object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 478–487). doi:10.1109/cvpr.2016.58.

Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3431–3440). doi:10.1109/CVPR.2015.7298965.

Maincare (2019). Plateforme de télémédecine. URL: https://www.maincare.com/solutions/telemedecine-et-teleradiologie/covalia-plateforme-de-telemedecine/covalia-plateforme-de-telemedecine-192-205.html.

Marie, F., Corbat, L., Chaussy, Y., Delavelle, T., Henriet, J., & Lapayre, J.-C. (2019). Segmentation of deformed kidneys and nephroblastoma using case-

based reasoning and convolutional neural network. *Expert Systems with Applications*, *127*, 282–294. doi:10.1016/j.eswa.2019.03.010.

Marie, F., Corbat, L., Delavelle, T., Chaussy, Y., Henriet, J., & Lapayre, J.-C. (2018). Segmentation of kidneys deformed by nephroblastoma using case-based reasoning. In *ICCBR 2018, 26th International conference on Case-based reasoning* (pp. 351–365). Stockholm, Sweden. doi:10.1007/978-3-030-01081-2_16.

Mignotte, M. (2010). A label field fusion bayesian model and its penalized maximum rand estimator for image segmentation. *IEEE Transactions on Image Processing*, *19*, 1610–1624. doi:10.1109/TIP.2010.2044965.

Mignotte, M. (2014). A label field fusion model with a variation of information estimator for image segmentation. *Information Fusion*, *20*, 7–20. doi:10.1016/j.inffus.2013.10.012.

Mignotte, M., & Hélou, C. (2014). A precision-recall criterion based consensus model for fusing multiple segmentations. *Int J Signal Process Image Process Pattern Recognit*, *7*, 61–82. doi:10.14257/ijsip.2014.7.3.07.

Nguyen, D. C. T., Benameur, S., Mignotte, M., & Lavoie, F. (2018). Superpixel and multi-atlas based fusion entropic model for the segmentation of x-ray images. *Medical image analysis*, *48*, 58–74. doi:10.1016/j.media.2018.05.006.

Qu, L., He, S., Zhang, J., Tian, J., Tang, Y., & Yang, Q. (2017). Rgbd salient object detection via deep fusion. *IEEE Transactions on Image Processing*, *26*, 2274–2285. doi:10.1109/TIP.2017.2682981.

Rohlfing, T., Brandt, R., Menzel, R., & Maurer Jr, C. R. (2004). Evaluation of atlas selection strategies for atlas-based image segmentation with application to confocal microscopy images of bee brains. *NeuroImage*, *21*, 1428–1442. doi:10.1016/j.neuroimage.2003.11.010.

Tang, Y., & Wu, X. (2016). Saliency detection via combining region-level and pixel-level predictions with cnns. In *European Conference on Computer Vision* (pp. 809–825). Springer. doi:10.1007/978-3-319-46484-8_49.

Tang, Y., Wu, X., & Bu, W. (2016). Deeply-supervised recurrent convolutional neural network for saliency detection. In *Proceedings of the 24th ACM international conference on Multimedia* (pp. 397–401). ACM. doi:10.1145/2964284.2967250.

Wang, Z., Bovik, A. C., Sheikh, H. R., Simoncelli, E. P. et al. (2004). Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, *13*, 600–612. doi:10.1109/tip.2003.819861.

Xiao, F., Deng, W., Peng, L., Cao, C., Hu, K., & Gao, X. (2018). Msdnn: Multi-scale deep neural network for salient object detection. *IET Image Processing*, *12*, 2036–2041. doi:10.1049/iet-ipr.2018.5631.

Zhao, R., Ouyang, W., Li, H., & Wang, X. (2015). Saliency detection by multi-context deep learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1265–1274). doi:10.1109/CVPR.2015.7298731.