Proceedings of the Fifth IIEEJ International Workshop
on Image Electronics and Visual Computing 2017
Da Nang, Vietnam, February 28- March 3, 2017

PREGNANT UTERINE ULTRASOUND IMAGE SEGMENTATION BY ENCODING-DECODING CONVOLUTIONAL NEURAL NETWORK

Yan Li† Rong Xu†† Jun Ohya†, †† Hiroyasu Iwata†
†Faculty of Science and Engineering, Waseda University
††Global Information and Telecommunication Institute, Waseda University

ABSTRACT

This paper explores the effectiveness of applying convolutional neural network (CNN) to segmenting pregnant uterus in ultra sound (US) images. More specifically, this paper treats this segmentation problem as a pixel-wise classification task. A data-driven method based on end-to-end designed CNN encodes the entire input image into multiple down scaled feature maps and decodes them to a confidence map as the uterine segmentation result. By conducting experiments that use pregnant uterine US images, the feasibility of the CNN is discussed. The result shows that we could achieve highly accurate segmentation results.

1. INTRODUCTION

Segmenting pregnant uteruses from echographic (ultrasound: US) images is very important for achieving an automatic fetal biometry measurement system based on fetal reconstruction. The segmentation is challenging, as shown in Fig. 1, because of irregular shapes of the uterine walls deformed by the fetal bodies that could change their postures, noisy reflected waves (black shadow-like areas in Fig. 1) that are generated by tissues and bones, and blurry edges.

Energy function based boundary searching method is one of the common ways to learn the deformable shape information on US images, for example, active contour model [2]. Another mainstream is morphology based methods such as growing regions [3]. However, these methods either heavily rely on strong and clear image patterns or require extra precisely initialization; therefore, it is difficult for these methods to be applied to the pregnant uterine US image segmentation.

Apart from these model-driven or pre-defined rules based approaches, deep learning based machine learning mechanism achieved great success in nature image processing. As a data-driven method, convolution network structures could approximate any objective function by tons of neurons and effectively choose the most discriminative features by local receptive fields filtering on input image signals. Recent research works have shown that CNN (Convolutional Neural Network) could handle segmentation on complex scenarios such as road images and indoor object images [4] [5] [6], by treating the problem as pixel-wise classification task.

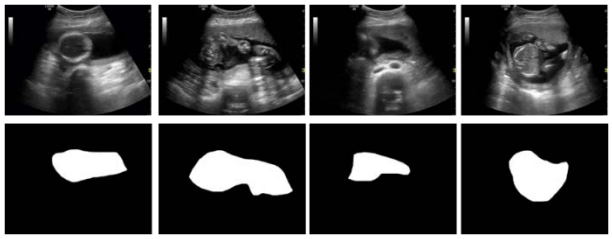

Figure 1.Examples of pregnant uterine US images
1st row: original images; 2nd row: segmented uterine (in white) containing the fetus and amniotic fluid

The methods achieved precise results in nature image processing.

This paper explores the effectiveness of applying the deep learning technology to segmenting pregnant uteruses from US images. More specifically, this paper aims at automatically determining whether each pixel belongs to the uterus or not in any input US images. The definition of the "uterus" area is the pixels inside the uterine walls, where the uterus area could probably include the amniotic fluid and fetal body. The rest of the pixels are defined as background, as shown in Fig.1.

This paper adopts a fully convolution network structure which could output the confidence that each pixel corresponds to the uterine area in the input image. In the training phase, the weights of the network are updated using manually labeled training data. During the test phase we directly feed the entire image to the trained model. By conducting experiments that use pregnant uterine US images, the feasibility of the CNN is explored.

2. RELATED WORK

US image segmentation There are two approaches: morphology based methods and shape model based methods. Morphology based methods are commonly used for segmentation of tissues in US images such as [3]. They look for specific areas or continuous edges by predefined rules. These kinds of simple approaches could be easily confused by similar appearances and blurry borders. On the other hand, contour extraction based methods [2] are also frequently used to segment US images. It could achieve quiet acceptable results by learning shape prior from the statistical model. However, these model-driven methods are hard to deal with complicated shapes because of lack of ability of generalizing samples that newly appear, and they heavily rely on the initial locations.

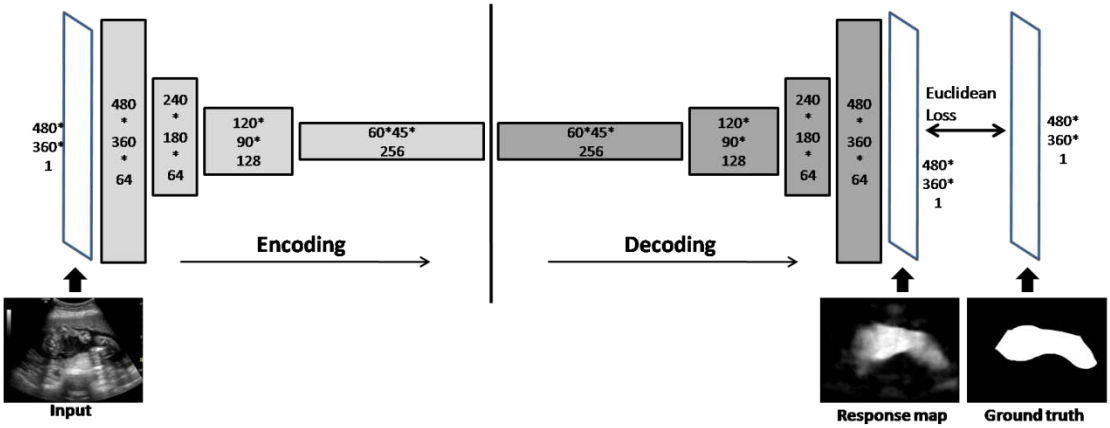

Figure 2.Network based on Segnet-Basic [5]
Light gray: Conv-RELU max pooling in the encoding part; Dark gray: up-sample conv in the decoding part. Kernel size of all of the conv layers are 7 as a result of the experimental comparison in Sec. 4.3.

Deep learning based dense label prediction The common CNN structure was been introduced in biological image segmentation as early as 2005 [1]. Recently deep learning based approaches show the ability of handling the pixel-wise semantic segmentation task. It is achieved by discarding the last fully connecting layers in a common network and replacing them by up-scale skills, such as bilinear interpolation [4], up-sample layers [5] or de-convolutions [6]. They could generate an end-to-end segmentation directly in input images. The convolution structure has ability to learn at pixel level and to re-define their values in the output side by passing the image data through a set of encoding-decoding process. However, to our knowledge, deep learning has not yet been applied to the US image segmentation for extracting the pregnant uterine areas.

3. APPROACH

This paper exploits deep learning for segmenting pregnant uterine areas from US images. More specifically, we use learn-able convolution kernels to supervise a mapping function from input gray level image to a confidence map.

For each single US image we pass the entire image as the input to a fully convolutional network to encode so as to obtain multiple down-scaled response maps. Then, in order to obtain an end-to-end prediction without losing details owned by the original US images, the model uses corresponding up-sample layers followed by learn-able convolution kernels to decode the response maps with the same size as the input image. The approach is detailed in the following sections.

3.1. Network structure for pixel-wise labeling

The main difference from common CNN with full connected layers structure is that the output layer of the CNN this paper uses is a dense map, which is a set of multiple labels. For this, this paper uses a modified structure based on [5], which is designed for object segmentation for nature images. Compared with their work, we use different convolution kernel numbers and different output function, as shown in Fig. 2.

The network consists of two parts: encoding and decoding parts. Specifically, in the encoding part, all of the kernel size for the convolution layer is fixed to a value of 7 according to the experimental results in Sec. 4.3. We pad each response map by zeros on the border of the matrix (kernel) in order to keep the size fixed after being filtered by the convolution kernel. Every convolution layer is followed by a rectified linear unit (RELU) activation function for the models. Then, the size of the response maps is determined as the one obtained by decreasing the size of the pooling layers. This could yield scale invariance and represent information from surrounding (larger) areas. For the pooling strategy we use max pooling which shrinks the input map by only keeping the maximum value of the specific pixel neighbors. Both of the size and stride we use for each pooling layer are 2. Note that we need to record the max indices, in order to transfer back the response map by the up-sample layers in the decoding part of the network. We repeat this kind of layer blobs (conv-RELU-maxpooling) four times totally in encoding network.

Concerning the decoding part, to decode the encoded the response maps without losing dense (the same resolution as the input) information, we need to up-scale the layer to the same resolution as the input image. This could be achieved by several optional up-scale schemes, for example by using bi-linear transform [4], or learn-able de-convolution kernels [6]. This paper uses simple up-sample to do this, which is same as [5]. The max pooling indices are recorded in the encoding layer as mentioned earlier, so that for each up-scaled image, we could restore the pixel values by finding the corresponding values and put them in the previously recorded position. The other pixels are all filled by zeros. Then, each up-sample layer is followed by the convolution layers with activations as usual as described before. We use exactly same numbers of layers in the reversed order as the encoding layers.

For each input image we resize it to a fixed size of 480*360, and convert to a single gray level channel. Since we use a fully convolution network structure, there are no full connected layers, which allows us to

handle any arbitrary input resolution in the test phase, and the output is always an end-to-end segmentation map. The specific resolution of each layer's output is shown in Fig.2. Different kernel numbers are set for each layer; for the first two convolution layers there are 64 shared kernels for each and for the following two layers the kernels are increased by factor of square of 2. For the decoding part we use corresponding number of kernels as shown in Fig. 2. The last layer has only one layer output, which indicates the confidence of uterus or background at each pixel.

The training procedure calculates the loss of the whole network by summing each sample's pixel-wise Euclidean distance between the response map and ground truth label map. The cost function to be minimized is written as:

$$f = \frac{1}{2}\sum_{x\in \mathrm{X}} \left\| v_{\mathrm{g}}(x) - v_{\mathrm{p}}(x) \right\|_2^2 \qquad (1)$$

where x indicates a pixel position in image X, $v_{\mathrm{g}}(x)$ is the ground truth label and $v_{\mathrm{p}}(x)$ indicates the outputted pixel value at pixel x. During the back propagations, for the decoding part of the network, the derivatives are down sampled by recorded max indices. This process is performed in the opposite direction of the decoding process.

To summarize, we use an encoding-decoding CNN network structure to perform an end-to-end, pixel-level supervised learning mechanism. Compared with only using down-scaled image output and then up-scaling by interpolations, the usage of up-sample layers followed by convolution layers brings more accurate predictions for each of the pixels [5].

3.2. Result thresholding

The value of each pixel of the response map indicates a confidence value of whether it belongs to "uterus" or not. In order to obtain a reasonable threshold for binary mask for the uterus class despite imbalanced numbers of pixels in the two classes, before setting this threshold to the test set, we first run testing on all of our training data. Therefore, by testing several, different thresholds within a specific range, a best threshold could be determined. We assume the testing data has a similar distribution with the training set; therefore, we could treat this threshold value as the most suitable threshold for the testing.

4. EXPRIMENT

4.1. Dataset

All of the US images, which are used for the experiments, are captured by a GEVoluson E8, where we have received approval from the Ethics Review Committee on Research with Human Subjects, Waseda University (2014-165). The gestational weeks of the fetal video we used for our experiment are around 19 and 23. We randomly separated the two as the training and testing set. As a total, 226 frames for the training set (week 19) and 188 for the test set (week 23) are sampled from the videos. All of the images are resized to a size of 480*360 pixels in order to directly input to the CNN.

Each pixel's label is manually given: specifically, if a person specifies the (enclosed) area of the uterus in each US image using a graphic tool, then, the "uterus" label is given to all the pixels inside the specified uterus area, while the "non-uterus" (background) label is given to the other pixels. Note that, for each US image, it is ensured that in almost all of the samples we could observe the fetal body inside the uterus.

4.2. Setup

Evaluation criteria We use the following three different criteria for quantitative evaluations of the experimental results: 1. Global accuracy (Accu_G): the properly classified pixel counts divided by the total pixel counts in the dataset. 2. Mean accuracy over all categories (mAccu_C): averaging the pixel-wise classification accuracy of all of the classes (in case of this paper, two classes). This index could reflect the imbalanced factor of each class. 3. Mean intersections over union (mIoU): the IoU of each class i is calculated by the following equation:

$$\mathrm{IoU}_i = \frac{\mathrm{N}_{i\mathrm{correct}}}{\mathrm{N}_{i\mathrm{correct}} + \mathrm{N}_{i\mathrm{incorrect}} + \mathrm{N}_{i\mathrm{missed}}} \qquad (2)$$

where $\mathrm{N}_{i\mathrm{correct}}$ is the number of correctly classified pixels; $\mathrm{N}_{i\mathrm{incorrect}}$ and $\mathrm{N}_{i\mathrm{missed}}$ indicate the false positive and false negative pixels, respectively. Equation (2) indicates that the IoU reflects both of the false positives and the false negatives; therefore, it can be said that this criterion is more strict than the other two criteria.

Training details To deal with small scale of the training data, we augment the data by randomly cropping in the original image and flipping, together with their corresponding label maps. For each sample we augment 10 samples totally. As for the hyper parameters of the model, we use a fixed learning rate at $1*10^{-6}$. The learning rate is quite small because the loss is calculated by summing all the pixels of the image, which will generate a large value for the loss. The loss tends to decrease if it reaches approximately 16,000. We stop the training at around 500, which corresponds to 20,000 iterations or more. Under this network structure and these settings, the training program costs about 6,600 Mb GPU memories with batch size set to 4. It takes for the training about 5 hours for 20,000 iterations on a NVIDIA GTX1080.

4.1. Results

First, whether the assigned label is correct or not is checked at each pixel in all of the test data. The 188 test data mentioned in Sec. 4.1 are used for the segmentation experiments. Table 1 lists the three criteria for the Euclidean loss (EuclidieanL) and Softmax classification (MultinomialL), where Euclidean_k3, Euclidean_k5 and Euclidean_k7 are the results of using kernel size 3, 5 and 7, respectively. By comparing Euclidean_k7 with MultinomialL, we could clearly see that all of the scores

Table 1.Overall segmentation accuracy

	Accu_G	mAccu_C	mIoU
EuclidieanL_k3	0.9045	0.8499	0.7314
EuclidieanL_k5	0.9428	0.9005	0.8212
MultinomialL	0.9300	0.9046	0.7957
EuclidieanL_k7	**0.9520**	**0.9073**	**0.8442**

Table 2.Class separated segmentation accuracy

	Accu_ BG	Accu_ U	IoU_B G	IoU_U
EuclidieanL_k3	0.9317	0.7680	0.8905	0.5723
EuclidieanL_k5	0.9639	0.8371	0.9335	0.7089
MultinomialL	0.9427	**0.8666**	0.9182	0.6732
EuclidieanL_k7	**0.9742**	0.8405	**0.9441**	**0.7443**

of the regression are higher than the classification. As can be seen in Table 1, among the three different kernel sizes, the largest kernel size (7) gives the best segmentation accuracy.

Second, the segmentation accuracy is evaluated in each of the two classes. To evaluate this accuracy, Accu_BG, Accu_U, IoU_BG and IoU_U are used, which indicate the accuracy and IoU (intersection of union, which indicates the measurement of the ratio of overlapping and union region) of the background (_BG) and uterine area (_U). From Table 2, it can be said that the background (_BG) area gives higher accuracy than the uterus (_U). Since the number of pixels in the non-uterus is larger than the uterus, their pixel value variation is larger; thus, many pixel values tend to be treated as the background. Another reason is that the features of the fetal body in the uterus tend to be quite similar to those of the background area.

We visualize some examples of the result in Fig.3. Figure 3 shows that most of the pixels are correctly labeled, even in some of the complex or weak edge areas. From the third row we could see that the overall border is consistent with the ground truth shape of the uterus in different scan slide. The segmentation could be hardly affected by fetal bodies, which tend to have very similar pixel value variations to non-uterus pixels. It indicates that the CNN based method is robust to the easy-to-be-confused local patterns.

On the other hand, some problematic cases can be seen. The problems can be classified into the following three cases. First, the wrong segmentations occur in areas contaminated by noisy reflections, specifically, the misclassified pixels are easily happened at the blurry edge. In such an area, enough amounts of training data was not obtained, and a large size of the convolution kernel cannot cover this problem. Second, sometimes multiple enclosed areas were obtained, but obviously, this is against the reality. From now on, to suppress this issue, some post-process such as posing positional constraints should be developed. Third, errors often happen in case that part of the border (typically, the left and/or right border) of the amniotic fluid area in the uterus overlaps with the outer (out of the field of view of the US probe) area, where the gray-levels of amniotic fluid and outer area are dark and similar.

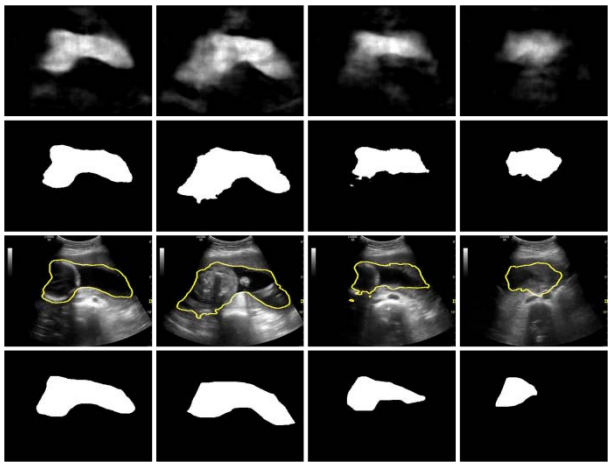

Figure 3.Visualized result samples. The four columns correspond to four different US images, and four rows show the original US image, ground truth for the uterus area (indicated by white), the obtained border between the uterus and non-uterus (yellow line) and the obtained uterus area (white), respectively.

5. CONCLUSION

This paper has explored the effectiveness of applying a deep learning based method to achieving a pixel-wise labeling to pregnant uteruses in US images by fully convolution network structure with down scaled and then up scaled scheme.

Experiments that use 188 pregnant uterine US images were conducted. The network is trained by using 226 sample US images whose uterine areas are manually masked. We studied the feasibility of the CNN, comparing different kernel sizes with Softmax classification. It turns out a large kernel size can achieve the best segmentation accuracy.

Remaining issues for improving the segmentation accuracy include how to deal with noisy reflections and split (multiple segmentations) of the uterus area, as well as how to deal with the case in which the border of the amniotic fluid in the uterus overlaps with the dark outer area.

6. REFERENCES

[1] Ning et al, “Toward Automatic Phenotyping of Developing Embryos from Videos”, IEEE Transactions on Image Processing, vol. 14, issue 9, pp. 1360-1371 (2005)

[2] Ni et al, “Segmentation of uterine fibroid ultrasound images using a dynamic statistical shape model in HIFU therapy” Comput Med Imaging Graph, vol. 46, Part 3, pp. 302 -314(2015)

[3] Usha et al, “Segmentation of Doppler Carotid Ultrasound Image using Morphological Method and Classification by Neural Network”, Int Journal of Engineering Research and General Science vol. 3, issue 3, pp.650-656 (2015)

[4] Long et al, “Fully Convolutional Networks for Semantic Segmentation”, CVPR (2015)

[5] Kendallet al, “SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation” TR, arXiv, (2015)

[6] Noh et al, “Learning Deconvolution Network for Semantic Segmentation” ICCV (2015)