On the Use of Fully Convolutional Networks on Evaluation of Infrared Breast Image Segmentations

Rafael H. C. de Melo, Aura Conci, Cristina Nader Vasconcelos

Computer Institute

Federal Fluminense University (UFF) - Niterói, RJ, Brazil

{rmelo,aconci,crisnv}@ic.uff.br

Abstract. *Medical images usually must have their region of interest (ROI) segmented as a first step in a pattern recognition procedure. Automatic segmentation of these images is an open issue. This paper presents an automated technique to define the ROI for infrared breast exams, based on the use of Fully Convolutional Networks (FCN). Adequate comparison among new approaches by using available databases is very important, here some comparisons with other techniques are made. Moreover, concerning on line diagnosis, the comparison among possible techniques must be efficient enough to be done in real time. With our approach the time to segment the ROI was 100 milliseconds and the average accuracy obtained was 95%.*

1. Introduction

Pattern recognition and image processing techniques have been applied in analysis of the most common types of medical diagnosis. In these techniques, the first step, after acquiring the image, is to separate the important element, that is, to obtain the region of interest (ROI). In computer aided systems, a complete automatic procedure of ROI segmentation is desirable. That is, this region must be found with no user interaction. Although due to complexities and importance of this step in the computer aided diagnosis the ROI identification algorithms must be considered to produce a correct result and for this must be evaluated comparing its results again the manually done results, or the named "ground truth". Moreover, in most of these computer aided diagnosis (CADx), computer aided detection (CADe) and in clinical decision support systems (CDSS) the ROI segmentation must be done in real time, that is, during the patient examination, to avoid time wasting for doctors in postprocessing information's.

A difficult part of evaluating ROI segmentation of medical images is to obtain the ground truth. This is because it is necessary a huge effort and time of specialists (doing a manual segmentation is very time consuming). In this work, we use a public database that has 285 images (original and ground truths) [Silva et al 2014] [Visual Lab 2016].

The use of Convolutional Networks (convnets) to solve many problems of image recognition are growing over the last years, mainly for semantic segmentation [Ciresan et al 2012], [Farabet et al 2013], [Pinheiro and Collobert 2014], [Hariharan et al 2014] and [Gupta et al 2014]. A new idea is presented here: to evaluate the power of a Fully

Convolutional Networks (FCN) [Shelhamer et al 2016] on ROI segmentation of infrared breast images.

The propose of use a Deep Network relies in the fact that it learns all the necessary filters and characteristics by itself in order to better describe, or, in our case, segment, the image given as input. Normally, to train deep nets like this, it is better to have in hand a huge dataset, but our universe of ground truths only has 285 inputs so, after separating the training, testing and validating sets a data augmentation approach was included on the training set. We use two approaches for it: horizontal mirror and displacement. Another strategy used to deal with the small amount of input data was the use of pre trained classifier weights (what is named fine-tuning). We better explain this on section 4.

The goal of this work is to present this new and fast way to extract ROIs for a diagnostic system. Thus, it explores the generalization capability of the Fully Convolutional Networks to the task of ROI identification (segmentation). Is important to keep in mind that once the Network is trained it generates the segmentation of new input images very fast even in a low cost hardware. The results were compared against other techniques in respect the quality of the response.

2. Validating the segmentation of region of interest (ROI)

The region of interest (ROI) segmentation in infrared images intends to separate the regions of the breast and it neighborhood from the input image. ROI must include all breast tissue, and, as much as possible, the entire related ganglion groups [Conci et al, 2015]. The correct definition of region of interest (ROI) and development of ground truth has a key role on the development of segmentations techniques and breast disease detection.

In the last two decades, the breast thermography has achieved an average sensitivity and specificity around 90% for breast tumors detection [Ng, 2009]. Studies shown that thermogram could identify precancerous or cancerous areas earlier than others exams [Arora et all 2008] [Amalu et all 2006]. Infrared images are not invasive or harmful to patients and is cheaper than traditional methods, such as mammography, ultrasound and magnetic resonance. It also has a potential use for diagnosis of young women since their breasts present a density that makes difficult early visualization of problems by x-ray. The various exams are complementary, so, the thermogram should be used with other method instead of substituting the traditional ones. The concept of combined diagnostic makes possible the achievement of a high degree of specificity and sensibility in diagnosis [Conci et al. 2013].

3. Database

A public database with 285 medical images (in gray scale) as well as its ground truths (in black and white) have been used, its image has 320x240 pixels [Silva et al 2014]. Database could be accessed in [Visual Lab 2016]. A sample of the data can be seen on figure 1. Original images of the ground truths only have the contour of the ROI (in red) and the rest of the image in white on the original database. In order to enter as the label

of the convnet we transform the ROI in the white region and the background in black (on figure 1 the ground truth is already converted to black and white).

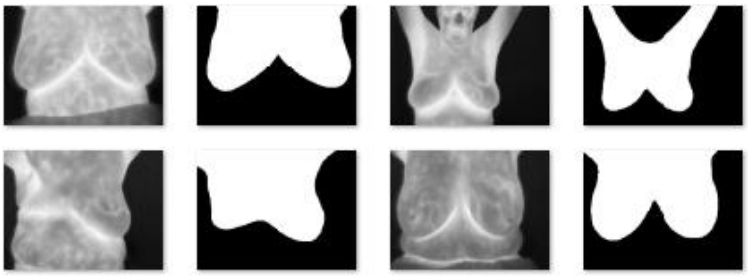

Figure 1. Representing a sample of the database with 4 thermal images in grayscale and the ground truths with the ROI in white and the rest in black.

4. Segmentation Architecture

For training, we use the VGG16 architecture used on FCN-VGG16 called FCN-8s on [Shelhamer et all 2016]. We use the concept of fine-tuning that is instead of starting the Net with random weights (no knowledge) we start our Net with weights from previous training that already looked many images. We can draw a parallel here using as sample a problem to child education that we would like to teach (train) how to spell, we could start with a kid that didn't have any reading skills (random start/weights) or with a kid that already know how to read (pre trained weights). For the fine-tuning we use the pre trained classifier weights obtained in [Shelhamer et all 2015].

The use of pre trained weights of segmentation problems that differs from the problem in had worth because the visual clues on digital images are always present. The Convolutional Networks, when presented with a huge amount of digital images starts to understand in the lower layers the basic clues like straight lines, corners, things that will help the Net to understand any kind of image.

The weights selected were the ones of the database PASCAL VOC 2011 segmentation challenge [Everingham et all 2011] because they presented the better results [Shelhameret all 2015]. The database used to generate this weight contains natural images (planes, birds, cats, sofas and so on) not related with the medical images used on this work but they contribute a lot to the training step because the low level visual patterns are present in any kind of image. So, the use of these weights helped a lot the training.

FCN-8s was the chosen architecture because was the one with best results. The strategy of taking in consideration the connection with lower layers of the net seeking spatial information about the image is the main advantage of this architecture in relation with the others presented on [Shelhameret all 2015].

5. Training approaches

The input passed to the convnet we used are the gray scale images of figure 1 and the output (or label) are also images, the black and white images of figure 1. As the database of the infrared breast image is too small to serve as input of deep networks (only 285 medical images), we adopted two distinct approaches in order to increase the amount of usable data seeking for the best result. The Horizontal Mirror on the data

augmentation was used because the breasts on image are supposed to be symmetric so it seems to be a good approach. The displacement of the 320x240 images on the 500x500 black background was used to produce a translation invariance on the method. Data augmentation approaches like that are very common on Deep Learning solutions like [Ciresan et al 2012], [Gupta et al 2014] and [Shelhamer et al 2016]. So, the two approaches used were: one duplicating the training set with the horizontal mirror and the other with the horizontal mirror plus the displacement of the images in nine different locations (top-left, top-center, top-right, center-left, center-center, center-right, bottom-left, bottom-center and bottom-right).

In all approaches, we separate 20% of the images for the testing set and 20% to the validation. We have done this before the data augmentation. As input for the net image of size 500x500 is expected, then original images were positioned initially in top left of a black canvas and the augmentation with displacement moves the original image around the other eight possible positions cited.

On horizontal mirror, the number of training samples doubled and on horizontal mirror plus displacement, we positioned the 88 samples in eight new positions (examples of new images are in figure 2). We describe the input size of each approach in table 1.

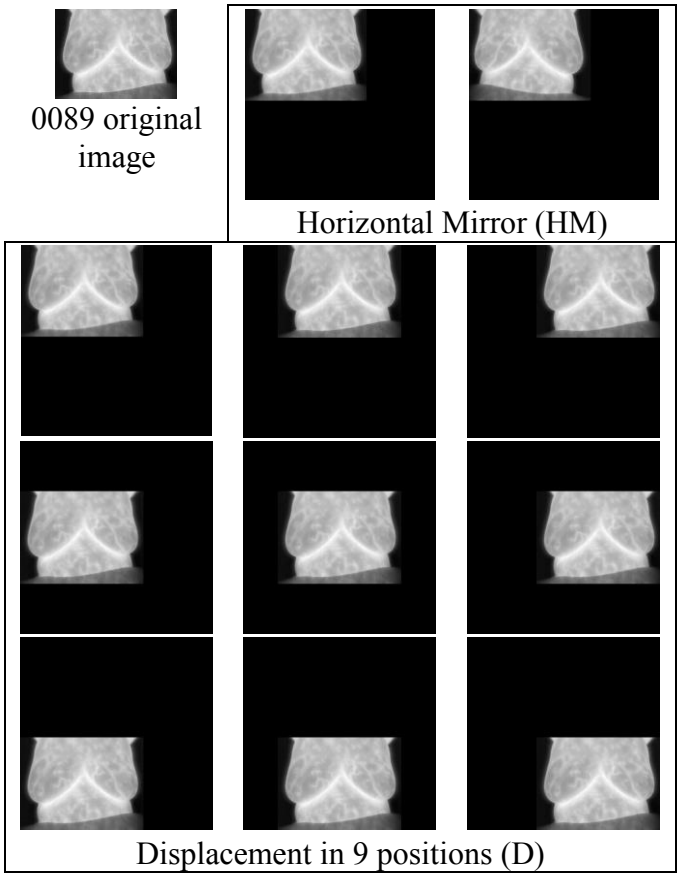

Figure 2. Illustration of the original image (320x240) and the two approaches of data augmentation used on images (500x500).

In spite of the input images are on gray scale they were considered as colored (on DIGITS input database) in order to be possible using the fine-tuning with the FCN-8s pre trained weights.

Table 1. Number of images in each augmentation approach

Augmentation approach		Mnemonic	Train	Validation	Test
1	Horizontal Mirror	HM	342	57	57
2	Horizontal Mirror plus Displacement	HM+D	1052	57	57

6. Experiments

The training step of the FCN need a good capacity of processing, so we use a GPU GTX TITAN X 12GB on Ubuntu 14.04 operating system. As only the training step is costly for the net execution once trained we used an Intel Core i7-4770 CPU with 3.4 GHz and 16Gb of RAM also in Ubuntu 14.04 operating system. For the batch size, we use one and two and variations of batch accumulation from one to four. Each model was trained using learning rates (alpha) of 1e-15, 1e-14, 1e-13 e 1e-12. Learning rate is the training parameter that controls how much a sample contributes on the update of network parameters (weight and bias) in one epoch (epoch is one complete pass through the whole training set). The other parameters used were the DIGITS default. The training converges in less than 30 epochs and after that the results shown that the model stops growing knowledge. All trainings take around 2 hours to converge.

We use the Interactive Deep Learning GPU Training System, DIGITS [NVIDIA DIGITS 2016], version 4 as the tool of loading and training the network. This version do not has native models to the segmentation task, so we have to make some adaptations commented on [Shelhamer et al 2016] using files from [Shelhamer et al 2015]. The new version of DIGITS (version 5) already comes with native models and the user interface to the segmentation task.

7. Results

The models trained with random weights (without fine-tuning) did not converge. The raining loss maintain it initial values.Higher probability on the no convergence lies on the fact that Depth Neural Networks requires a huge amount of input to correctly adjust its parameters and the database we use is very small. The validation error (loss - val) in figure 3 as well as in the other tests were not correct, probably DIGITS confuses the calculation or a problem with its configuration. In figure 3 is possible see a fine-tuned training with validation errors that seem to not converge but the results presented by this model are about 95% of accuracy.

Results of the two approaches within the best models are on table 2. All approaches using fine-tuning have very similar results and converged in few epochs. The models trained with batch=2 don’t change results.

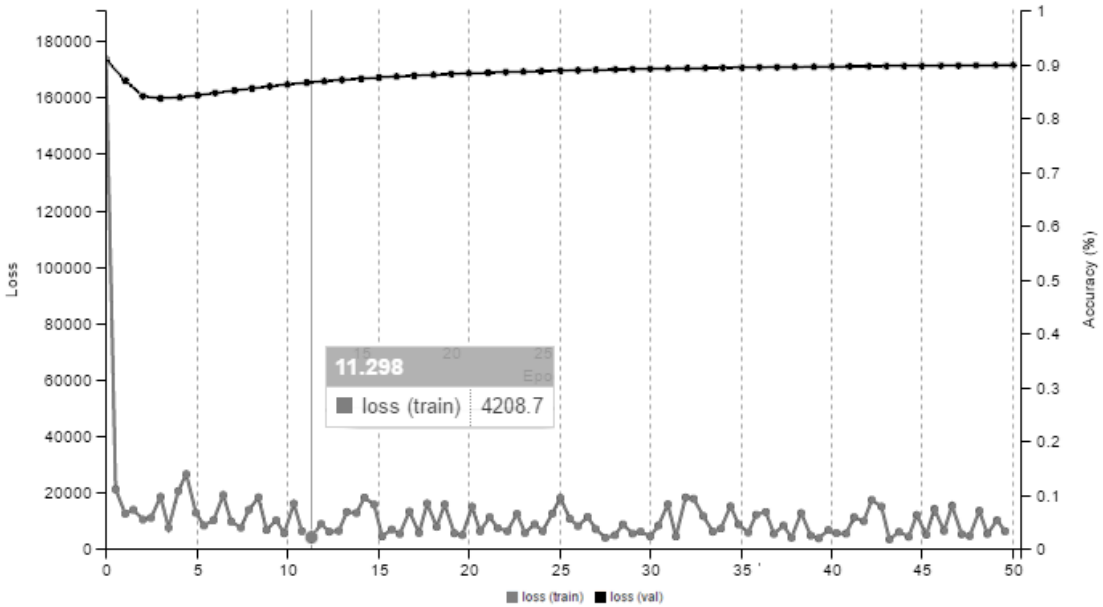

Figure 3. Training with fine-tuning. Approach HM, alpha = 1e-12. High validation error but 95% of accuracy on the segmentation results (printed screen from DIGITS).

Table 2. Results of the best models in all the approaches

Model	Batch	Epochs	Average Pixel Accuracy		
			ROI	Background	Total
HM – alpha 1e-12	1	11	97.4%	93.8%	95.5%
HM+D – alpha 1e-13	1	25	97.4%	94%	95.6%
HM+D – alpha 1e-12 batch acc 4	1	20	97%	93%	95%

The average distribution of the pixels on the test image set is well balanced between ROI (average: 49.4%, min: 38.3% and max: 64.6%) and background (average: 50.6%, min: 35.3% e max: 61.7%). The plots show the pixel accuracy distribution through the teste images. Figures 4 and 5 show some statistics results over the test set.

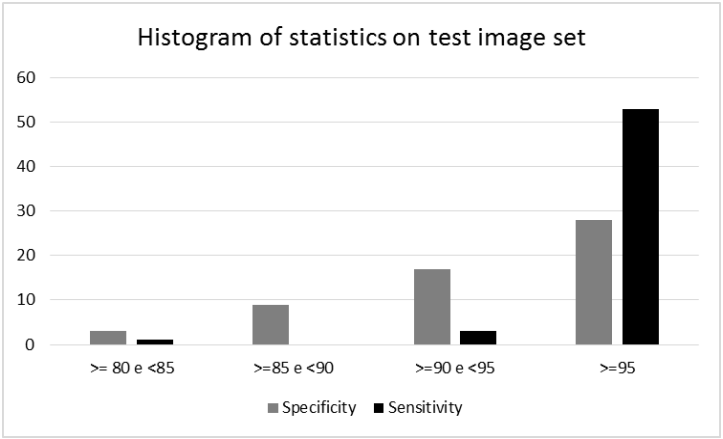

Figure 4. Histogram of ROI (sensitivity) and background (specificity) pixels correctly identified in model HM–alpha 1e-12.

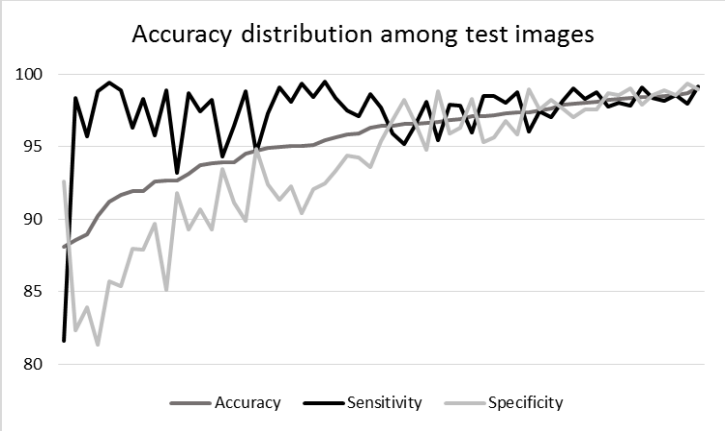

Figure 5. Pixel Accuracy distribution through test set in model HM–alpha 1e-12.

Next some samples (good, bad and curious) of the input images (figures 6, 7 and 8) can be seen, the segmentation result and its differences to the ground truth. The number images are the infrared breast exams in grayscale, GT prefix indicates the ground truth and the SEG prefix indicates the segmentation here proposed.

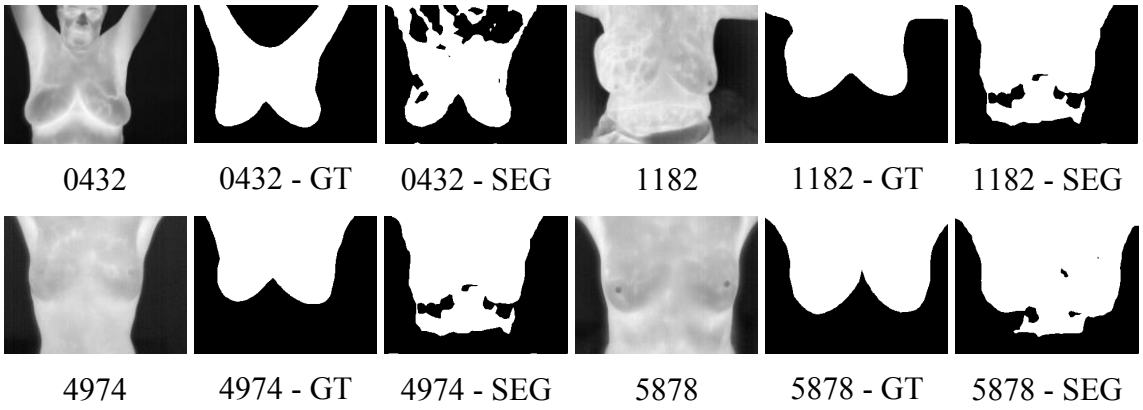

Figure 6. Samples of bad segmentations (accuracy less or equal to 90%). Sensitivity or Specificity near 80%.

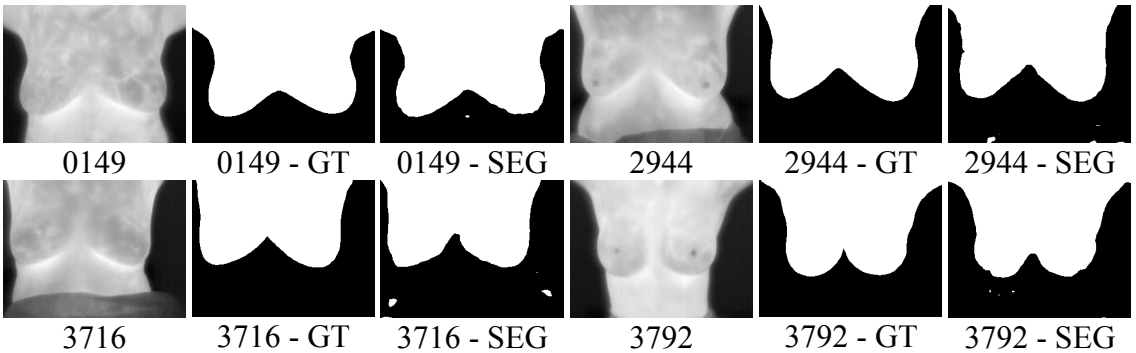

Figure 7. Samples of good segmentations (accuracy greater than 98%).

About the results in figure 6 (0432) is important to point out that the face of the patients appear in only three over the 285 images, in only fifteen the neck appears. Another important thing to note in many of the segmentations of figures 7 (images 0149, 3716, 3792) and 8 (image 5140) is that some little white islands were wrongly identified

as ROI parts). A simple pos processing could solve this problem and improve the results but in this work we focused in only explore the potentials of the FCN in an end-to-end application.

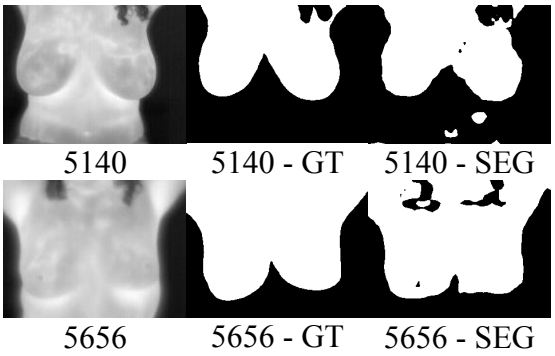

Figure 8. Curious samples: 95% of accuracy on 5140 and 92.7% on 5656.

Figure 8 shows the only two images from the database that presents the hair of the patients over the ROI. Quite curious actually is that on the ground-truth of the image 5140 the hair area is removed from the ROI but this doesn't happened on image 5656, possibly a misjudgment of one or more of the three specialists. The method here proposed, even without receiving any image with hair as input on the training (once the only two images of the database were on the test set) decide to remove the area of the two images as the segmentation result. The net result can even be considered more accurate than the ground truth in these cases. Table 3 shows the comparison between the best results obtained using a traditional image processing method including pre and pos processing and our proposal of using only a Deep Convnet. There are only three images because these was the intercept of the results showed in [Conci et al 2015], that, in this table, is identified as LSF, and our randomically choosed test.

Table 3. Comparing results of Accuracy, Sensitivity and Specificity obtained in model HM+D – alpha 1e-13 of this work with the LSF.

Image	Accuracy (ACC)		Sensitivity (SEN)		Specificity (ESP)	
	FCN	LSF	FCN	LSF	FCN	LSF
IR_3830	97.49	98.50	97.41	98.40	97.60	98.70
IR_3921	93.85	97.10	98.24	99.60	89.32	94.70
IR_5870	97.17	98.30	98.47	98.10	95.65	98.60
Max (all samples)	98.87	98.70	99.73	99.90	98.95	99.00
Min (all samples)	88.58	95.90	82.24	93.70	82.15	91.50
Average (all samples)	95.61	-	97.44	-	94.01	-

The three results of FCN on table 3 have ACC, SEN and ESP lower than LSF method but the results are quite promising since none pre or pos processing was used and the input data is very small for the general idea of convnets. It is important to note that the maximum values of accuracy considering all the samples exposed on papers are greater on our proposal. From last line of table 3 we see the average for each statistical measured for all our test set. Table 4 shows the best results for all the images of the test set and table 5 shows the worst results. The time to segment the ROI was 100 milliseconds in all models.

Table 4. Ten results of best Accuracy (ACC) on model HM+D – alpha 1e-13.

Image	Accuracy (ACC)	Sensitivity (SEN)	Specificity (ESP)
IR_0149	98.87	99.37	98.05
IR_2944	98.83	98.83	98.84
IR_3792	98.72	98.46	98.95
IR_3849	98.66	98.47	98.83
IR_3716	98.6	99.03	98.25
IR_0225	98.23	98.21	98.26
IR_3951	98.19	97.85	98.59
IR_3730	98.18	98.31	98.05
IR_3437	98.05	98.93	97.36
IR_1336	98.01	97.8	98.19

Table 5. Ten results of worst Accuracy (ACC) on model HM+D – alpha 1e-13.

Image	Accuracy (ACC)	Sensitivity (SEN)	Specificity (ESP)
IR_0432	88.58	82.24	92.99
IR_4974	89.08	97.79	83.52
IR_5479	89.89	99.73	83.28
IR_1182	90.4	95.64	86.48
IR_5878	90.59	98.71	82.15
IR_3819	90.88	99.03	83.74
IR_5656	91.07	91.29	90.67
IR_5624	91.59	98.52	87.17
IR_5336	92.52	96.09	89.31
IR_2931	93.16	98.41	89.51

8. Concluding remarks

The results shown that even with small datasets FCN could achieve good accuracy (95% average) using fine-tuning and some data augmentation strategies. To obtain the segmentation once the net is trained takes only 100 ms, which is very good to real time requirements. Future works include improving the results with simple pre and pos processing filters.

Acknowledgments

The author R.H.C.M. thanks CAPES for the financial support. A.C. is partially supported by CNPq , MACC and SIADE2. C.N.V. would like to thank NVIDIA for the donation of the Titan X used in this research.

References

Amalu, W. C., Hobbins, W. B., Head, J. F., Elliot, R. L. (2006) “Medical Devices and Systems”. In Infrared Imaging of the Breast – An Over View, In: Bronzino, J. D., The Biomedical Engineering Handbook, Third edition, pages 25.1 – 25.20. CRC Press.

Arora, N. M. D.; Martins, D. B. S.; Ruggerio, D. B. S.; Tousimis, E. M. D.; Swistel, A. J. M. D.; Osborne, M. P. M. D. and Simmons, R. M. M. D. (2008), "Effectiveness of a noninvasive digital infrared thermal imaging system in the detection of breast cancer", In: The American Journal of Surgery, pages 196, 523–526.

Ciresan, D. C., Giusti, A., Gambardella, L. M., and Schmidhuber, J. (2012) "Deep neural networks segment neuronal membranes in electron microscopy images" in NIPS, 2012, pp. 2852–2860. 1, 2, 4, 7

Conci, A., Sanchez, A., Liatsis, P., Usuki, H. (2013) "Signal Processing Techniques for Detection of Breast Diseases", Signal Processing. Vol. 93, pp. 2784-2788.

Conci, A., Galvão, S., Sequeiros, G. O., Saade, D.C.M., Machenry, T. (2015) "A new measure for comparing biomedical regions of interest in segmentation of digital images", Discrete Applied Mathematics, v. 1, p. 1.

Everingham, M., Van Gool, L., Williams, C. K. I., Winn, J., and Zisserman, A. (2011) "The PASCAL Visual Object Classes Challenge 2011 (VOC2011) Results" URL: http://host.robots.ox.ac.uk/pascal/VOC/voc2011/index.html.

Farabet, C., Couprie , C., Najman, L., and LeCun, Y. (2013) "Learning hierarchical features for scene labeling", PAMI, 1, 2, 4, 7, 8.

Gupta, S., Girshick, R., Arbelaez, P., and Malik, J. (2014) "Learning rich features from RGB-D images for object detection and segmentation," in ECCV, 1, 2, 8.

Hariharan, B., Arbelaez, P., Girshick, R., and Malik, J. (2014) "Simultaneous detection and segmentation," in ECCV, 1, 2, 4, 5, 7, 8, 9.

Ng, E.Y.-K. (2009) "A review of thermography as promising non-invasive detection modality for breast tumor", International Journal of Thermal Sciences, Volume 48, Issue 5, Pages 849-859, ISSN 1290-0729.

NVIDIA DIGITS, (2016) "NVIDIA Deep Learning GPU Training System". URL: https://developer.nvidia.com/digits. Acessed on September 10th of 2016.

Pinheiro, P. H. and Collobert, R. (2014) "Recurrent convolutional neural networks for scene labeling," in ICML, 1, 2, 4, 7, 8

Shelhamer, E., Long, J. and Darrell, T. (2016) "Fully Convolutional Networks for Semantic Segmentation". URL: https://arxiv.org/pdf/1605.06211.pdf. Acessed on 15th of 2016.

Shelhamer, E., Long, J. and Darrell, T. (2015) "Pre-trained weights of the FCN, Python Code and Caffe", URL: https://github.com/shelhamer/fcn.berkeleyvision.org. Acessed on October 8th of 2016.

Silva, L. F., Saade, D. C. M., Sequeiros, G. O., Silva, A. C., Paiva, A. C., Bravo, R. S., Conci, A. (2014) "A New Database for Breast Research with Infrared Image", Journal of Medical Imaging and Health Informatics, v.4, p.92 - 100.

Visual Lab, Database of infrared images and segmentations done by specialists. Ground Truth of 285 images. URL: http://visual.ic.uff.br/proeng/software.php. Acessed on October 26th of 2016.