Robust 3D Organ Localization with Dual Learning Architectures and Fusion

Xiaoguang Lu(✉), Daguang Xu, and David Liu

Medical Imaging Technologies, Siemens Medical Solutions, Inc., Princeton, NJ, USA
xiaoguang.lu@siemens.com

Abstract. We present a robust algorithm for organ localization from 3D volumes in the presence of large anatomical and contextual variations. The 3D spatial search space is decomposed into two components: slice and pixel, both are modeled in 2D space. For each component, we adopt different learning architectures to leverage respective modeling power on global and local context at three orthogonal orientations. Unlike conventional patch-based scanning schemes in learning-based object detection algorithms, slice scanning along each orientation is applied, which significantly reduces the number of model evaluations. Object search evidence obtained from three orientations and different learning architectures is consolidated through fusion schemes to lead to the target organ location. Experiments conducted using 499 patient CT body scans show promise and robustness of the proposed approach.

1 Introduction

Automatic 3D organ localization is essential in a wide range of clinical applications. It provides seed points to initialize subsequent segmentation algorithms. It is also useful for visual navigation, automatic windowing, semantic tagging, and organ-based lesion grouping.

Accurate localization of organs still remains a challenging task. From the local contextual perspective, the size, shape, and appearance of organs vary significantly across patients, even more so when there are pathologies or prior surgeries. Global context around each organ also varies significantly, although the context within the entire field of view such as that among multiple anatomical organs provides a cue for individual organ localization. For example, in the abdominal region, organs such as the kidney can "float" around with large degrees of freedom, therefore leading to varying appearance context. Various sizes of field of views and different body regions in clinical practice also increase the variation of global appearance.

Data-driven learning-based approaches have shown success and been widely deployed in object localization tasks. A typical search strategy in such methods uses a scanning window based scheme. A model/classifier is trained based on annotations to determine likelihood of a patch (sub-volume) being the target object. During online testing, the classifier is applied to each sub-volume

G. Carneiro et al. (Eds.): LABELS 2016/DLMIA 2016, LNCS 10008, pp. 12–20, 2016.
DOI: 10.1007/978-3-319-46976-8_2

by scanning through the entire volume. Target location is calculated by consolidating evidence collected from all scanned patches. Conventional scanning window patch-based approach is more suitable for capturing local appearance variations given its limited field of view (voxels within the sub-volume), but not global appearance variations. Many methods have been proposed in this paradigm; some focus on improving the classifiers, while others improve the scanning strategy [11], or integrates other modeling methods such as conditional random field [3] and recursive context propagation network [12].

Another category of method is based on long range regression and voting. In [1], a regression forest is trained to find the non-linear mapping from voxels to the desired anatomy location, which extracts features globally from the volume, and is shown to be effective for resolving local ambiguities. However, it has been shown in [8] that the precision of such regression methods is not as accurate as the patch based classification methods due to large context variations.

We propose a framework which models both local and global context without using patch-based scanning schemes, where two emerging learning architectures are exploited to complement each other. We use the convolution neural network (CNN) [7] to capture global context [13], and the fully convolutional network (FCN) [10] to capture local context. The local context focuses on the localization precision, while the global context helps improve robustness such as resolving ambiguities and eliminating false detections. The global context and local appearance information are integrated through a probabilistic graphical model, and we call such a learning scheme as the dual learning architecture. We show in our experiments that, with explicitly modeling and fusion of both local and global contextual information, our approach is more robust and achieves a higher accuracy compared to the state-of-the-art algorithms. In addition to the object location, a significant amount of positive seeds (within the target organ) are generated, which are useful for subsequent processes such as segmentation using graph-cut methods. Furthermore, because both CNN and FCN support multi-label tasks, our algorithm can be generalized to simultaneous multi-organ localization with limited extra run-time computational cost.

2 Methodology

2.1 Context Modeling with Dual Learning Architectures

The organ localization task is formulated as a probabilistic graphical model [6], as shown in Fig. 1. Random variable I denotes a 2D image, E represents the existence ($E = 1$) or absence ($E = 0$) of the organ of interest within image I, and L is the organ location within image I. Both E and L are hidden variables, while I is an observed variable. The joint distribution factors according to the probabilistic graphical model as follows:

$$P(I, E, L) = P(L|I, E)P(E|I)P(I). \tag{1}$$

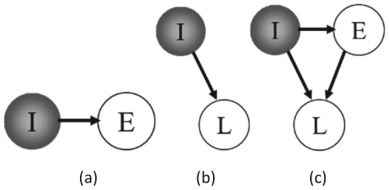

Fig. 1. Probabilistic graphical models describing the relationship between image I, the existence (E) of the organ in the image, and the location (L) of the organ in the image. From left to right: Global image classification (slice-level), local (pixel-level) classification, and the proposed global-local image classification.

Our goal is to query the organ location given the image, i.e., $P(L|I)$. This can be expressed as

$$P(L|I) = P(L,I)/P(I) = \sum_E P(I,E,L)/P(I) = \sum_E P(L|I,E)P(E|I)P(I)/P(I) = \sum_E P(L|I,E)P(E|I). \tag{2}$$

By definition, $P(L|I, E = 0) = 0$ for all valid locations, and $P(L = empty|I, E = 0) = 1$. Therefore

$$P(L|I) = P(L|I, E = 1)P(E = 1|I) \tag{3}$$

for all valid pixel locations, and

$$\begin{aligned} &P(L = empty|I) \\ &= P(L = empty|I, E = 1)P(E = 1|I) + P(L = empty|I, E = 0)P(E = 0|I) \\ &= P(L = empty|I, E = 0)P(E = 0|I). \end{aligned} \tag{4}$$

The probability distribution function $P(E = 0 \text{ or } 1|I)$ poses an image categorization problem. This function is depicted in Fig. 1(a). This was often implemented by extracting global image features and training a classifier on those features. In recent years, deep Neural Networks have shown superior performance in this task. In this paper, we use the Convolutional Neural Network (CNN) [7].

The probability distribution function $P(L|I, E = 1)$ presents a pixel classification task. In contrast to $P(E|I)$, which is a global image classification problem, $P(L|I, E = 1)$ is a local pixel or patch classification problem, where the patch is centered at pixel location L. One could again use a CNN to classify each patch, but in recent literature it has been shown that the fully convolutional networks (FCN) demonstrate advantages over the CNNs for pixel-level classification. We therefore adopt the FCN for this local image classification problem. To the best of our knowledge, this is the first time an FCN is used in conjunction with a CNN in a "dual learning" architecture for solving the global-local pixel classification problem.

While the FCN is described above as a local pixel classifier, it has been used in the literature to classify pixels into multi-label masks, in which the "background" class is one of the possible labels. This means, we could have used directly the FCN to classify all the pixels without using the global CNN classifier at all. However, as we will show in the experiments, there are significant advantages of combining the FCN with the CNN, where FCN's limited receptive field [9,15] is compensated by CNNs' response. This is also evident from the above probabilistic formulation: a FCN-only pixel classifier would model directly $P(L|I)$ as shown in Fig. 1(b) without considering the hidden variable E. Therefore, our global-local model poses a stronger assumption than a typically FCN-only classifier, which does not have knowledge of the presence of the organ. For multi-organ localization tasks [4], the proposed method can be extended through multi-label training with the same architectures.

Compared with patch-based sub-window scanning in conventional object localization, in our method, one entire slice (not a sub-patch) is used as one input sample to either CNN and FCN. During online testing on a given volume, for each CNN or FCN model, the total number of image samples that are passed through CNN/FCN for evaluation is the number of slices along one orientation.

2.2 Cross-Sectional Fusion and Clustering

The dual learning architectures with respective models operate along each of the three orthogonal orientations, i.e., axial, sagittal, and coronal, resulting in three volumetric probability/score maps. These maps are generated from different orientations with different image context and therefore provide complementary information towards the target localization decision making. Typical ensemble schemes or information fusion approaches can be applied, such as majority voting, or sum rule [5], to lead to a consolidated score for each voxel. We call this scheme cross-sectional fusion.

After the consolidated probability/score map is computed, three-dimensional connected component analysis is conducted. The centroid of the largest cluster is computed as the estimated object location.

3 Experiments

Among all the organs with available expert annotations, the right kidney is one of the most challenging organs [2]. We use the right kidney as an exemplar case in our experiments. We have collected 450 patient CT body scans, one scan from each patient. For each scan, right kidney was manually delineated. At the training stage, 405 scans were selected at random for training and the remaining 45 scans (10 %) for validation. Our training data covers large variations in populations, contrast phases, scanning ranges, and pathologies. The axial slice resolution ranges between 0.5 mm and 1.5 mm. The inter-slice distance varies from 0.5 mm to 7.0 mm. Scan coverage includes abdominal regions, but can extend to head/neck and knees. After all models were trained, we collected another

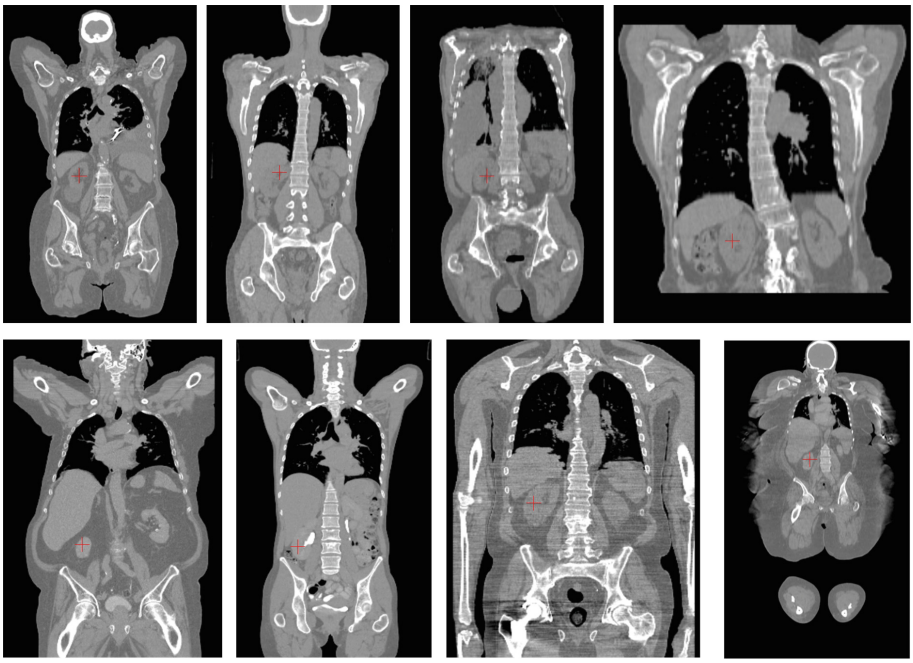

Fig. 2. Coronal slice samples in the test set. Note that the large context variations with respect to the right kidney. Red cross indicates the right kidney location automatically detected by the proposed method. (Color figure online)

49 patient CT scans from clinical sites for independent testing. Right kidney is also manually delineated in these 49 test cases to compute quantitative measurement for algorithm performance evaluation. Typical test scan samples are provided in Fig. 2.

Each CT scan contains a stack of axial slices, which were used to reconstruct a 3D volume at an isotropic resolution of $2 \times 2 \times 2\, mm^3$. All the algorithms/models in our subsequent experiments operate at this resolution. Three orthogonal orientations (axial, sagittal, and coronal) are considered for cross-sectional analysis. Only the right hand side of the body is considered in the experiments (training and testing) as the right kidney is the target object. The centroid of the delineated right kidney was used as ground-truth location. A volumetric mask was generated based on the annotations, where right kidney voxels are labeled

Table 1. Number of training images for each model.

Number of images	Axial	Sagittal	Coronal
CNN	118245	42482	90559
FCN	41276	25938	27378

as ones and all other background was labeled as zeros. This mask was used to provide the labels for FCN training. For CNN training, a two-class classification is defined, i.e., whether or not an image slice contains the right kidney.

Slice-level modeling (CNN): the AlexNet architecture [7], which contains 5 convolution layers and 3 fully connected layers, is adopted. One CNN model is

Table 2. Statistics of Euclidean distance from the automatic localization result to the ground-truth position at 2 *mm* resolution. Sum rule is applied in cross-sectional fusion. CS: cross-sectional fusion.

Dist. (voxels)	CS-(CNN+FCN)	MSL	CS-CNN	CS-FCN
Mean	3.9	12.8	9.1	9.0
Std	4.7	10.7	11.4	23.0
Median	2.3	10.9	5.4	1.9

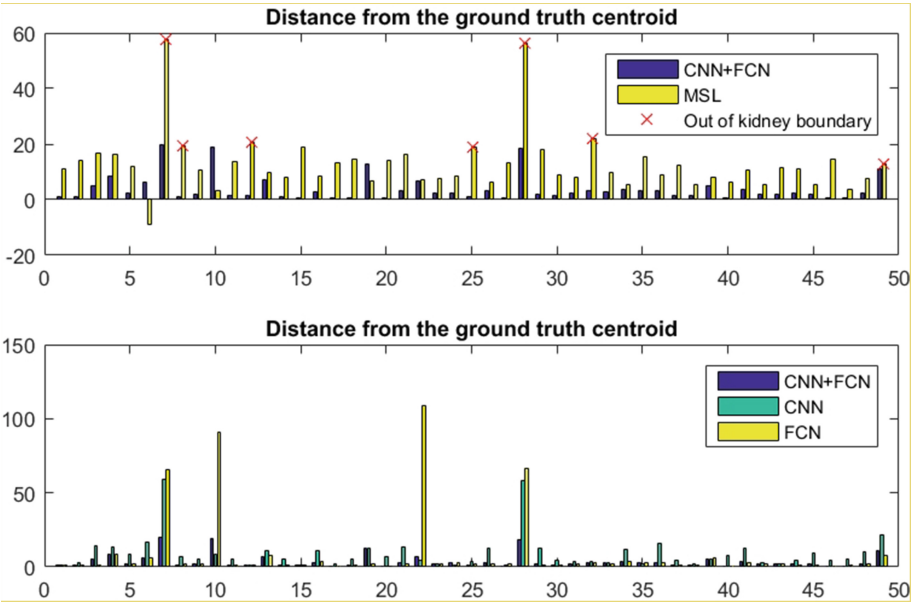

Fig. 3. Euclidean distance between the calculated right kidney location and the ground-truth location for each of the 49 test cases (horizontal axis is case index) in number of voxels at the isotropic 2 *mm* resolution. Negative distance (case 6 in Top) indicates that the corresponding localization algorithm does not generate any detection results, and the absolute distance value in this case is nominal for visualization purposes. Top: comparison of the proposed method (blue) and MSL (yellow), where a red cross indicates the localization result is out of the actual kidney boundary. Bottom: comparison of the proposed method (blue), CNN only (green), and FCN only (yellow). Results of CNN, FCN, and CNN+FCN are all calculated through cross-sectional fusion. (Color figure online)

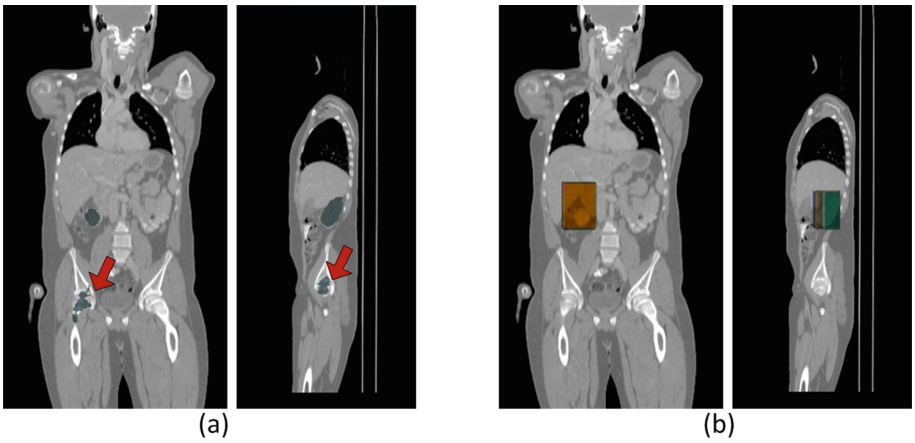

Fig. 4. Example of model responses (color overlaid) from FCNs (a) and CNNs (b) after cross-sectional fusion. Responses are presented after fusion across three orientations. Each group contains one sagittal view and one coronal view. Red arrows indicate false alarms detected by FCNs. CNNs response maps show inferior localization precision on the same cluster. Combining both responses through fusion leads to successful right kidney localization. (Color figure online)

trained for each cross-section orientation using the same learning architecture. Pixel-level modeling (FCN): the VGG-FCN8s architecture [10] is adopted, which is an end-to-end network with 7 levels of convolution layers, 5 pooling layers and 3 deconvolution layers. One FCN model is learned for each cross-section orientation with the same network architecture. Table 1 lists the number of training images/slices used for each model.

For comparison, we implemented a 3D patch-based scanning window approach based on the method proposed by Zheng et al. [14], and applied it on the same test set. We refer to their approach as marginal space learning (MSL). Quantitative performance evaluation against the ground-truth is provided in Table 2 and Fig. 3. Figure 4 presents an example to demonstrate complementary information extraction from the dual learning architectures.

Although the focus of the proposed method is on organ localization, one typical use case of organ localization is for organ segmentation. We evaluate the impact of our kidney localization on the accuracy of kidney segmentation. As the MSL method together with active shape models has shown to provide good cardiac segmentation results [14], we adopt it for right kidney segmentation. Our automatic localization led to similar segmentation error rates compared to using the ground-truth locations. Using our automatic localization results as input for segmentation, the [mean, std., median, 80 percentile] of point-to-mesh errors (used in [14]) in *mm* are [2.32, 1.23, 1.91, 2.22], while the ground-truth locations led to error rates of [2.00, 0.48, 1.85, 2.20].

4 Conclusions

We have presented a robust 3D organ localization algorithm. We approach the 3D localization task through cross-sectional 2D modeling, exploiting two learning architectures that model various context for localizing the target organ. Contextual information extracted by the two learning schemes is complementary and integrated for improved robustness. Because FCN and CNN are capable of learning multiple targets/labels, our method can be extended for simultaneous multi-organ localization. Although CT body scans are used in the experiments, the proposed method is not limited to specific imaging modalities.

References

1. Criminisi, A., Shotton, J., Robertson, D., Konukoglu, E.: Regression forests for efficient anatomy detection and localization in CT studies. In: Menze, B., Langs, G., Tu, Z., Criminisi, A. (eds.) MICCAI 2010. LNCS, vol. 6533, pp. 106–117. Springer, Heidelberg (2011)
2. Cuingnet, R., Prevost, R., Lesage, D., Cohen, L.D., Mory, B., Ardon, R.: Automatic detection and segmentation of kidneys in 3D CT images using random forests. In: Ayache, N., Delingette, H., Golland, P., Mori, K. (eds.) MICCAI 2012, Part III. LNCS, vol. 7512, pp. 66–74. Springer, Heidelberg (2012)
3. Farabet, C., Couprie, C., Najman, L., LeCun, Y.: Learning hierarchical features for scene labeling. IEEE TPAMI **35**(8), 1915–1929 (2013)
4. Gauriau, R., Cuingnet, R., Lesage, D., Bloch, I.: Multi-organ localization combining global-to-local regression and confidence maps. In: Golland, P., Hata, N., Barillot, C., Hornegger, J., Howe, R. (eds.) MICCAI 2014, Part III. LNCS, vol. 8675, pp. 337–344. Springer, Heidelberg (2014)
5. Kittler, J., Hatef, M., Duin, R., Matas, J.: On combining classifiers. IEEE TPAMI **20**(3), 226–239 (1998)
6. Koller, D., Friedman, N.: Probabilistic Graphical Models: Principles and Techniques. Lippincott Williams & Wilkins, Philadelphia (2009)
7. Krizhevsky, A., Sutskever, I., Hinton, G.: Imagenet classification with deep convolutional neural networks. In: Proceedings of the NIPS (2012)
8. Lay, N., Birkbeck, N., Zhang, J., Zhou, S.K.: Rapid multi-organ segmentation using context integration and discriminative models. In: Joshi, S., Pohl, K.M., Wells, W.M., Zöllei, L., Gee, J.C. (eds.) IPMI 2013. LNCS, vol. 7917, pp. 450–462. Springer, Heidelberg (2013)
9. Liu, W., Rabinovich, A., Berg, A.C.: Parsenet: looking wider to see better (2015). arXiv:1506.04579v2
10. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the CVPR (2015)
11. Roth, H.R., Lu, L., Farag, A., Shin, H.C., Liu, J., Turkbey, E.B., Summers, R.M.: Deeporgan: multi-level deep convolutional networks for automated pancreas segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A. (eds.) MICCAI 2015. LNCS, vol. 9349, pp. 556–564. Springer, Heidelberg (2015)
12. Sharma, A., Tuzel, O., Liu, M.Y.: Recursive context propagation network for semantic scene labeling. In: Proceedings of the NIPS (2014)

13. Sun, Y., Wang, X., Tang, X.: Deep convolutional network cascade for facial point detection. In: Proc. CVPR (2013)
14. Zheng, Y., Barbu, A., Georgescu, B., Scheuering, M., Comaniciu, D.: Four-chamber heart modeling and automatic segmentation for 3D cardiac CT volumes using marginal space learning and steerable features. IEEE TMI **27**(11), 1668–1681 (2008)
15. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Object detectors emerge in deep scene CNNs. In: Proceedings of the ICLR (2015)

http://www.springer.com/978-3-319-46975-1

Deep Learning and Data Labeling for Medical Applications
First International Workshop, LABELS 2016, and Second International Workshop, DLMIA 2016, Held in Conjunction with MICCAI 2016, Athens, Greece, October 21, 2016, Proceedings
Carneiro, G.; Mateus, D.; Loïc, P.; Bradley, A.; Tavares, J.M.R.S.; Belagiannis, V.; Papa, J.P.; Nascimento, J.C.; Loog, M.; Lu, Z.; Cardoso, J.S.; Cornebise, J. (Eds.)
2016, XIII, 280 p. 115 illus., Softcover
ISBN: 978-3-319-46975-1