Rapid Multi-organ Segmentation Using Context Integration and Discriminative Models

Nathan Lay, Neil Birkbeck, Jingdan Zhang, and S. Kevin Zhou

Siemens Corporate Technology, 755 College Road East, Princeton NJ
{nathan.lay,neil.birkbeck,jingdan.zhang,shaohua.zhou}@siemens.com

Abstract. We propose a novel framework for rapid and accurate segmentation of a cohort of organs. First, it integrates *local and global image context* through a product rule to *simultaneously* detect multiple landmarks on the target organs. The global posterior integrates evidence over all volume patches, while the local image context is modeled with a local discriminative classifier. Through non-parametric modeling of the global posterior, it exploits sparsity in the global context for efficient detection. The complete surface of the target organs is then inferred by robust alignment of a shape model to the resulting landmarks and finally deformed using discriminative boundary detectors. Using our approach, we demonstrate efficient detection and accurate segmentation of liver, kidneys, heart, and lungs in challenging low-resolution MR data in *less than one second*, and of prostate, bladder, rectum, and femoral heads in CT scans, in *roughly one to three seconds* and in both cases with accuracy *fairly close to inter-user variability.*

Keywords: Local & global context, context integration, multi-landmark detection, discriminative learning, multi-organ segmentation.

1 Introduction

Algorithms for segmenting anatomical structures in medical imaging are often targeted to individual structures [1–4]. Instead, when the problem is posed as the joint segmentation of multiple organs, constraints can be formulated between the organs, e.g., non-overlapping, and the combined formulation allows for a richer prior model on the joint shape of the multiple structures of interest. Such multi-organ segmentation is often posed with atlas-based or level-set based formulation due to the ease at which geometric constraints can be modeled [5, 6].

However, level set methods are computationally demanding, and still require a decent initialization so as to not fall into a local minimum. Discriminative learning-based methods are often an alternative approach to initializing such segmentations (e.g., [5]), but, again, these methods often treat the initialization of each organ as an independent problem. While solving the single organ segmentation problem with learning-based methods can be fast (e.g., [2]), in order to achieve efficient multi-object segmentation, often a tree-like search structure has to be imposed on the detection order of the structures [7, 8].

J.C. Gee et al. (Eds.): IPMI 2013, LNCS 7917, pp. 450–462, 2013.

The sequential ordering is used to avoid evaluating local classifiers everywhere in the image. However, as shapes of adjacent structures are often correlated, the appearance of neighboring image patches are often consistent, meaning image patterns commonly associated with one organ, say the liver, are likely to appear next to the right kidney. Instead of modeling dependency among structures at the algorithm level, e.g., with generative models [7], the correlation between such *global image context* and the shapes can be learned directly (e.g., [4, 9, 10]).

One method to utilize global contextual cues is to regress the position of the organ bounding boxes from each voxel location in the image [4, 9, 10]. Others suggested that this global information alone may not be accurate enough, and further improved the accuracy using a cascade of locally trained regressors [11].

In this work, we propose a novel integration of both local and global discriminative information for efficient multiple organ segmentation. Unlike other learning-based approaches, we do not rely on a tree-like dependency structure of organ detections to obtain an efficient detection algorithm. Instead, our global image context is only sparsely sampled, allowing us to derive an efficient detection algorithm: global context is used to hypothesize locations that need to be evaluated with the local discriminative classifier. Our non-parametric representation of global image context models correlations in the target shape, allowing us to jointly localize landmarks on multiple target organs. We impose a constraint on the distribution of allowable shapes, enabling us to initialize a likley shape from only a few landmarks per organ. The initialized shape is then deformed using learned discriminative boundary detector to better fit image appearance. We demonstrate that the combination of local and global image context outperforms either local and global context alone, and illustrate the use of the proposed joint landmark detection, robust shape initialization, and discriminative boundary deformation to segment up to 6 organs in either CT or MR data in *roughly one to three seconds* with segmentation in MR data taking *less than one second*. The segmentation accuracy is *fairly close to inter-user variability*.

2 Method

We aim to segment C organ shapes, $\mathbf{S} = [\mathbf{S}_1, ..., \mathbf{S}_C]$, given a volumetric image $\mathbf{I}$. We denote the set of all voxels in the image $\mathbf{I}$ by Ω and its size by $|\Omega|$. We assume that there exists a set of D corresponding landmarks, $\mathbf{X} = [\mathbf{x}_1, ..., \mathbf{x}_D]$, on the multiple shapes $\mathbf{S}$ and decompose the problem into estimating (i) the landmarks given the image using the posterior $P(\mathbf{X}|\mathbf{I})$ defined in §2.1 and (ii) the shapes given the landmarks and the image using energy minimization in §2.2. We use the notation $[\mathbf{x}, ..., \mathbf{x}]_D$ to represent repeating $\mathbf{x}$ in D times.

2.1 Joint Landmark Detection Using Context Integration

To jointly detect the landmarks, we integrate both local and global image context using a product rule into one posterior probability $P(\mathbf{X}|\mathbf{I})$:

$$P(\mathbf{X}|\mathbf{I}) = P^L(\mathbf{X}|\mathbf{I})P^G(\mathbf{X}|\mathbf{I}), \tag{1}$$

where $P^L(\mathbf{X}|\mathbf{I})$ and $P^G(\mathbf{X}|\mathbf{I})$ are local and global context posteriors, respectively.

Local Context Posterior. Though not necessarily true, we assume that the landmarks are *locally independent*:

$$P^L(\mathbf{X}|\mathbf{I}) = \prod_{i=1}^{D} P^L(\mathbf{x}_i|\mathbf{I}). \tag{2}$$

For modeling $P^L(\mathbf{x}_i|\mathbf{I})$, we exploit the local image context to learn a discriminative detector for landmark $\mathbf{x}_i$ (using *e.g.* PBT [12]), that is,

$$P^L(\mathbf{x}_i|\mathbf{I}) = \frac{1}{Z_i}\omega_i^L(+1|\mathbf{I}[\mathbf{x}_i]), \tag{3}$$

with $\mathbf{I}[\mathbf{x}_i]$ being the local image patch centered at $\mathbf{x}_i$, $\omega_i^L(+1|\cdot)$ the local context detector for landmark $\mathbf{x}_i$ and $Z_i = \sum_{\mathbf{x}\in\Omega} \omega_i^L(+1|\mathbf{I}[\mathbf{x}_i])$ is a normalizing constant.

Global Context Posterior. We integrate global evidence from all voxels in Ω.

$$P^G(\mathbf{X}|\mathbf{I}) = \sum\nolimits_{\mathbf{y}\in\Omega} P^G(\mathbf{X}|\mathbf{I},\mathbf{y})P(\mathbf{y}|\mathbf{I}) = |\Omega|^{-1} \sum\nolimits_{\mathbf{y}\in\Omega} P^G(\mathbf{X}|\mathbf{I}[\mathbf{y}]). \tag{4}$$

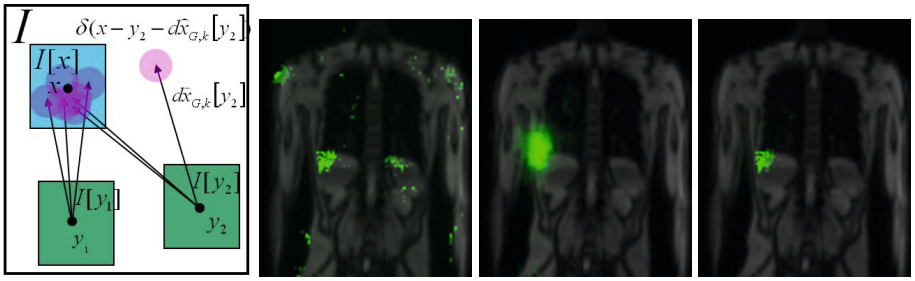
In (4), we assume a uniform prior probability $P(\mathbf{y}|\mathbf{I}) = |\Omega|^{-1}$ and $P^G(\mathbf{X}|\mathbf{I}[\mathbf{y}])$ is the probability of the landmarks at $\mathbf{X}$ when observing the image patch $\mathbf{I}[\mathbf{y}]$ at a location $\mathbf{y}$.

To learn $P^G(\mathbf{X}|\mathbf{I}[\mathbf{y}])$, we leverage annotated datasets and a 'randomized' K-nearest neighbor (NN) approach [13]. For a complete set of training images with annotated landmarks, we randomly form K subsets. From each subset of images with corresponding landmarks, we construct a training database $\{(\mathbf{J}_n, d\mathbf{X}_n)\}_{n=1}^{N}$ consisting of N pairs of image patch $\mathbf{J}$ and relative shift $d\mathbf{X}$ in an iterative fashion.

for *n=1,...,N* **do**
  Randomly sample in the subset an image say $\tilde{\mathbf{J}}$ with landmarks $\tilde{\mathbf{X}}$;
  Randomly sample a voxel location, say $\mathbf{z}$, from Ω;
  Set the image patch $\mathbf{J}_n = \tilde{\mathbf{J}}[\mathbf{z}]$;
  Set the relative shift $d\mathbf{X}_n = \tilde{\mathbf{X}} - [\mathbf{z},...,\mathbf{z}]_D$.
end

For a test image patch $\mathbf{I}[\mathbf{y}]$, we first find its NN $\hat{\mathbf{J}}_k$ from each subset; this way we find its K neighbors $\{\hat{\mathbf{J}}_1,...,\hat{\mathbf{J}}_K\}$ along with their corresponding shift vectors $\{d\hat{\mathbf{X}}_1[\mathbf{y}],...,d\hat{\mathbf{X}}_K[\mathbf{y}]\}$. How to efficiently find the NN for each subset is elaborated later. We then simply approximate $P^G(\mathbf{X}|\mathbf{I}[\mathbf{y}])$ as

$$P^G(\mathbf{X}|\mathbf{I}[\mathbf{y}]) = K^{-1} \sum\nolimits_{k=1}^{K} \delta(\mathbf{X} - [\mathbf{y},...,\mathbf{y}]_D - d\hat{\mathbf{X}}_k[\mathbf{y}]). \tag{5}$$



(a) (b) (c) (d)

Fig. 1. (a) An illustration of how image patches (green) predict the landmark location using global context and Eq. (5) and then these predictions are combined with local context at (blue) $\mathbf{x}$. (b) Detection scores for a landmark on the top left of the liver in a low resolution MR FastView 3D volume, where local context gives spurious responses. (c) Global context gives a coarse localization. (d) The integration of local and global detection gives a fine scale density.

Figure 1 graphically illustrates how the approach works. It also gives an example of the local, global and joint posteriors. Even though the local detector may be inaccurate, it is only being applied at locations predicted from the global context, meaning it is possible to get a highly peaked posterior when integrating evidence from local and global context.

MMSE and MAP Estimate for Landmark Location. The expected landmark location $\bar{\mathbf{X}}$, also the minimum mean square error (MMSE) estimate, is computed as

$$\begin{aligned}\bar{\mathbf{X}} &= \sum_{\mathbf{X}} \mathbf{X}\ P(\mathbf{X}|\mathbf{I}) = \sum_{\mathbf{X}} \mathbf{X}\ P^L(\mathbf{X}|\mathbf{I})P^G(\mathbf{X}|\mathbf{I}) \qquad (6)\\ &= \frac{1}{K|\Omega|}\sum_{\mathbf{X}}\sum_{\mathbf{y}\in\Omega}\sum_{k=1}^{K}\mathbf{X}\ \prod_{i=1}^{D}\frac{1}{Z_i}\omega_i^L(+1|\mathbf{I}[\mathbf{x}_i])\delta(\mathbf{X}-[\mathbf{y},...,\mathbf{y}]_D - d\hat{\mathbf{X}}_k[\mathbf{y}]).\end{aligned}$$

where $Z_i = \sum_{\mathbf{x}} \omega_i^L(+1|\mathbf{I}[\mathbf{x}_i])$ is a normalizing constant. Using the local independence and vector decomposition, it can be shown that the expected location $\bar{\mathbf{x}}_i$ for a single landmark is computed as

$$\bar{\mathbf{x}}_i = Z^{-1}K^{-1}|\Omega|^{-1}\sum_{\mathbf{y}\in\Omega}\sum_{k=1}^{K}(\mathbf{y}+d\hat{\mathbf{x}}_{k,i}[\mathbf{y}])\omega_i^L(+1|\mathbf{I}[\mathbf{y}+d\hat{\mathbf{x}}_{k,i}[\mathbf{y}]]). \qquad (7)$$

where $Z = \sum_{\mathbf{y}\in\Omega}\sum_{k=1}^{K}\omega_i^L(+1|\mathbf{I}[\mathbf{y}+d\hat{\mathbf{x}}_{k,i}[\mathbf{y}]])$ is a normalizing constant. Eq. (7) implies an efficient scheme – evaluating the local detector only for the locations predicted from the global context posterior instead of the whole image! Since the predicted locations are highly clustered around the true location, this brings the first significant reduction in computation.

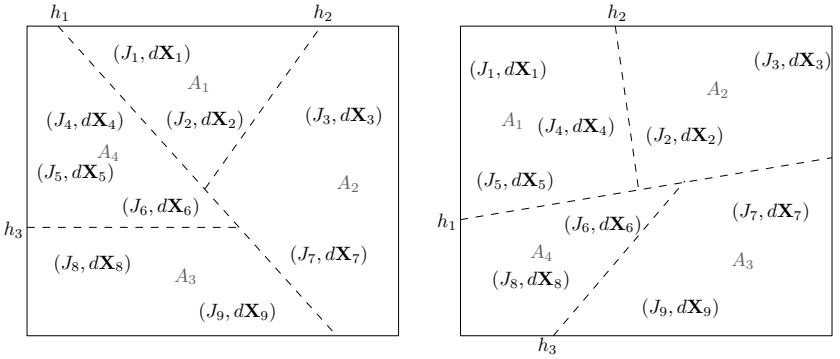


Fig. 2. Two BSP trees for different subsets of the training data are used to partition the space into convex regions (e.g., the leafs), A_j, using a set of hyperplanes h_i. Instead of searching over all entries within a leaf of the tree to find an exact NN, we simply store the average relative offset vector for the training samples that fell into the leaf.

Similarly, the maximum a posterior (MAP) estimate $\hat{\mathbf{x}}_i$ can be derived as

$$\hat{\mathbf{x}}_i = \arg\max_{\mathbf{x}} \omega_i^L(+1|\mathbf{I}[\mathbf{x}]) \sum_{\mathbf{y}\in\Omega} \sum_{k=1}^{K} \delta(\mathbf{x} - \mathbf{y} - d\hat{\mathbf{x}}_{k,i}[\mathbf{y}]). \tag{8}$$

Sparsity in Global Context. The global context from all voxels is highly redundant as neighboring patches tend to predict nearby landmark locations. Therefore, we can 'sparsify' the global context by constructing the subset Ω_ℓ from the full voxel set Ω; for example, we can skip every other l voxels. This brings the second significant reduction in computation complexity by $O(l^3)$!

Efficient Approximate NN Search. Computing the expected landmark location in (7) relies on the ability to compute the NN from the training database of $\{\mathbf{J}_n, d\mathbf{X}_n\}_{n=1}^N$ for one subset of training images. The time and space efficiency of this operation is influenced by two factors: the size of the database, N, and the dimension of the points, D, in the database. With, for example 100 training volumes of dimension 128^3, we have a potential database size of $N = 128^3 \times 100 > 209$ million. Furthermore, in order to have enough contextual information, an image patch J_n, with size up to 32^3 voxels, is used, meaning that the NN query must be performed in a high dimensional space of up to $D = 32768$.

For efficiency, we relax the requirement of finding the exact nearest Euclidean neighbor to that of finding an approximate NN. We then take a similar approach as local sensitive hashing [14] and build multiple hash indexes on the data (Fig. 2). However, instead of using a hash function, we construct a random Binary Space Partition (BSP) tree that is similar to a random projection tree [15]. At each node of our BSP tree, we choose a random hyperplane to split the data. Unlike random projection trees, which choose the split hyperplane uniformly random on a D-dimensional hypersphere, we restrict the hyperplanes to

be Haar wavelets. We have two reasons for doing this: 1) Haar wavelets provide a class of features often used to discriminate appearance in classification problems, and 2) any Haar feature can be instantaneously evaluated using an integral image. Further, instead of storing all training sample patches in their respective leaf nodes within the tree, we choose a single representative relative shift vector–this way the space requirements are dependent on the size of the tree instead of $O(ND)$.

In our experiments, we typically form $K = 10$ subsets and hence train 10 BSP-trees with each tree built up to depth 10. This means that an approximate NN match for a single tree is computed using at most 10 Haar wavelet evaluations, and all $K = 10$ approximate neighbors can be found in as little as 100 Haar wavelet evaluations.

2.2 Shape Initialization Using Robust Model Alignment

An initial segmentation for each organ is then aligned to the sparse detected landmarks through the use of a statistical model of shape variation. Here we use a point distribution model, where each organ shape is represented as a mean shape or mesh with M mesh nodes, $\bar{\mathbf{V}} = [\bar{\mathbf{v}}_1, \bar{\mathbf{v}}_2, \ldots, \bar{\mathbf{v}}_M]$, plus a linear combination of a set of N eigenmodes, $\mathbf{U}_n = [\mathbf{u}_{1,n}, \mathbf{u}_{2,n}, \cdots \mathbf{u}_{M,n}]$, with $1 \leq n \leq N$.

As a complete organ shape is characterized by only a few coefficients that modulate the eigenmodes, the point distribution model can be used to infer a shape from a sparse set of landmark points. Given a set of detected landmarks, $\{\mathbf{x}_i\}$, the best fitting instance of the complete shape is found by minimizing the following robust energy function:

$$(\beta, \{a_n\}) = \text{argmin}_{\beta,\{a_n\}} \sum_i \psi \left(\|\mathbf{x}_i - T_\beta\{\bar{\mathbf{v}}_{\pi(i)} + \sum_{n=1}^{N} a_n \mathbf{u}_{\pi(i),n}\}\|^2 \right) + \sum_{n=1}^{N} a_n^2/\lambda_n \tag{9}$$

where the function $\pi(i)$ maps the landmark $\mathbf{x}_i$ to the corresponding mesh index in $\bar{\mathbf{V}}$, the function $T_\beta\{\cdot\}$ is a 9D similarity transform parameterized by the vector $\beta = [t_x, t_y, t_z, \theta_x, \theta_y, \theta_z, s_x, s_y, s_z]$, and λ_n are the corresponding eigenvalues. The first term measures the difference between a predicted shape point under a hypothesis transformation from the detected landmark, and the second term is a prior keeping the eigenmodes responsible for smaller variation closer to zero. As we typically only have a few landmarks, and have a PCA model for a larger number of vertices, using no prior term gives rise to an ill-posed problem. Finally, ψ is a robust norm, reducing the effect of outliers. We use $\psi(s^2) = s$.

2.3 Discriminative Boundary Refinement

Using the initialization from §2.2, a fine refinement of the points on the surface mesh is obtained by iteratively displacing each vertex along its surface normal, $\mathbf{v}_i \leftarrow \mathbf{v}_i + \mathbf{n}_i\hat{\tau}_i$. The best displacement for each point is obtained by maximizing the output of a discriminative classifier [3]:

Table 1. Accuracy (measured in mm) and timing results for the landmark detection using local, global, and local + global context posterior

	Global		Local		Local + Global	
Spacing	Time	Median	Time	Median	Time	Median
Spacing	Time	Median	Time	Median	Time	Median
1 (5mm)	2.76s	**25.0 ± 17.4**	1.91s	16.4 ± 10.6	-	-
5 (25mm)	0.92s	39.9 ± 33.4	-	-	2.11s	**12.9 ± 7.52**
7 (35mm)	0.91s	54.1 ± 54.1	-	-	0.91s	13.0 ± 7.56
15 (75mm)	**0.89s**	79.0 ± 85.6	-	-	**0.23s**	14.1 ± 8.25

$$\hat{\tau}_i = \text{argmax}_{\tau_i} \omega^B(+1|\mathbf{v}_i + \mathbf{n}_i\tau_i). \tag{10}$$

Here, $\omega^B(+1|\cdot)$ is the boundary detector that scores whether the point, $\mathbf{v}_i + \mathbf{n}_i\tau_i$, is on the boundary of the organ being segmented. Regularity is incorporated in the previously independent estimated displacements by projecting the resulting mesh onto the linear subspace spanned by the linear shape model, as in the active shape model [16].

3 Results

Our system was implemented in C++ using OpenMP and compiled using Visual Studio 2008. In the experiments below, timing results are reported for an Intel Xeon 64-bit machine running Windows Server 2008 and using 16 threads. We illustrate the results on segmenting 6 organs in MR scans (§3.1) and 5 organs in CT (§3.2).

3.1 Lungs, Heart, Liver, and Kidneys in MR Localizer Scans

We tested our approach on a challenging set of MR localizer scans acquired using a fast continuously moving table technique (syngo TimCT FastView, Siemens). Such scans are often used for MR examination planning to increase scan reproducibility and operator efficiency. A total of 185 volumes having 5mm isotropic spacing were split into a training set of 135 and test set of 50. This data is challenging due to the low resolution, weak boundaries, inhomogeneous intensity within scan, and varying image contrast across scans. For this example, we used $K = 10$ NN. The local detectors were also trained on 5mm resolution using a PBT [12] and a combination of Haar and image gradient features. A total of 33 landmarks were selected on the 6 organs, with 6 landmarks each on the liver and the lungs, and 5 landmarks each on the kidneys and heart.

First, we demonstrate the effectiveness of integrating local and global context with respect to accuracy and evaluation time. Table 1 illustrates median errors for all landmark positions averaged over the testing set. For the local context detector and local+global posterior, we used the MMSE estimate. While it is

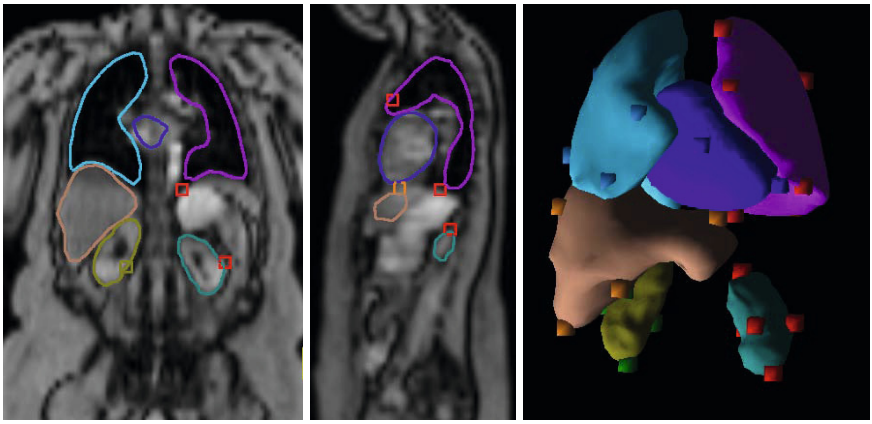


Fig. 3. An illustration of the landmarks in 3D and automatic segmentation results. Our method is robust to a few failed landmarks.

possible to get better speed-up with a sparse sampling of the global context when computing the expected value, we noticed that the MAP estimate gave better results as we reported in the table. Obtaining the MAP estimate requires populating a probability image and scanning through the image to get the MAP estimate (this is proportional to the number of landmarks, which is why no speedup is reported in the table). Besides, the accuracy of the global context posterior suffers from sparse sampling, and even with dense sampling it still performs worse than the local + global method. On the other hand, it is evident that a sparser sampling of the volume has little impact on the accuracy of the local+global method. The local classifier is computed using a constrained search over the volume (e.g., using bounds for the landmark positions relative to the image [2]), but achieves worse accuracy and is still slower than our combined local+global posterior modeling.

The shape landmarks are used to infer the shape of all the organs (see Fig. 3). We compare the resulting segmentation results at several phases to a state-of-the-art hierarchical detection using marginal space learning (MSL) [2] that is known as both fast and accurate. For the MSL setup, the kidneys were predicted from the liver bounding box, meaning the kidney search range was more localized allowing the detection to be faster (the lungs were predicted relative to the heart in a similar manner). Table 2 illustrates the timing and accuracy results for the 50 unseen test cases using both MSL and our method. The accuracy is gauged by symmetric surface-to-surface distance. Figure 4 illustrates two qualitative results.

The fast landmark detection and robust shape initialization can provide an approximate shape in as little as 0.33s (for spacing of 75mm, e.g., 15 voxels). The improvement of our initialization on the liver and lungs over the MSL approach is likely due to our use of more landmarks to capture more variations associated with complex anatomies than MSL that fits shapes of varying complexities into a rigid bounding box. On the other hand, for both kidneys with less variations

Table 2. Accuracy (measured in mm) and timing for segmentation results using our approach compared to the state-of-the-art MSL model on the MR FastView data

			Detection & Shape initialization					
	Skip (mm)	Time	Liver	R. Kidney	L. Kidney	R.Lung	Heart	L. Lung
MSL	-	5.50s	9.21 ± 1.82	**3.44 ± 1.16**	**3.08 ± 1.21**	7.29 ± 1.64	5.98 ± 1.59	7.42 ± 1.71
Local+Global	25mm	2.21	**7.41±1.91**	4.10±1.34	4.31±1.81	**6.60±1.74**	**5.64±1.41**	**6.72±1.55**
	35mm	1.01	7.43±1.95	4.18±1.39	4.39±1.89	6.67±1.79	5.69±1.40	6.78±1.53
	50mm	0.55	7.55±2.03	4.36±1.43	4.57±1.93	6.77±1.86	5.78±1.48	6.83±1.64
	60mm	0.39	7.63±1.95	4.59±1.52	4.70±1.98	6.86±1.91	5.92±1.53	6.91±1.68
	75mm	**0.33**	7.94±2.21	5.13±1.77	5.38±2.90	6.97±1.95	5.98±1.57	6.88±1.75
			With boundary refinement					
MSL	-	6.36s	4.87 ± 1.46	**2.26 ± 0.61**	**2.12 ± 0.68**	3.67 ± 0.95	**3.99 ± 1.36**	3.55 ± 0.97
(BSP)	25mm	2.89	**4.07±0.99**	2.33±0.68	2.41±1.61	**3.56±0.96**	4.02±1.50	**3.35±0.83**
	35mm	1.60	4.08±0.99	2.37±0.69	2.47±1.72	3.57±0.98	4.02±1.52	**3.35±0.83**
	50mm	1.13	4.09±1.01	2.37±0.73	2.48±1.66	3.57±0.95	4.06±1.62	3.36±0.83
	60mm	0.97	4.08±1.00	2.42±0.79	2.42±1.57	3.57±0.97	4.07±1.63	3.35±0.84
	75mm	**0.89**	4.17±1.14	2.51±1.00	2.84±2.51	3.57±0.95	4.11±1.64	3.37±0.83
Inter-user variability			4.07±0.93	1.96±0.43	2.10±0.51	3.79±0.36	4.54±0.88	3.52±0.63

in the shape but more in the appearance, MSL performs better as it considers kidney as a whole. The discriminative boundary deformation significantly improves the segmentation accuracy for both approaches, which yield comparable overall accuracy for all organs. Our approach is more efficient, e.g., over 5 times faster if we skip every 12th voxel (65mm) in the global context. With a skipping factor of 75mm, we achieved segmentation of 6 organs *within one second* and with accuracy almost as good as the best quality! Both methods perform fairly close to inter-user variability[1].

One potential concern with relying on far away global context information is that the reliability of the detection and segmentation may degrade or vary when given a subvolume. To investigate this, we evaluated the lung, liver, and heart segmentation accuracy on the same subset of unseen volumes, but this time we cropped the volumes 10cm below the lung and heart, meaning that the kidneys and liver are not present. In these cropped volumes, using a spacing factor of 50mm, we find the accuracy of our local+global method to be consistent with that in Table 2, where right lung accuracy was 3.57 ± 1.32, heart accuracy was slightly worse at 4.53 ± 2.39, and the left lung was 3.22 ± 1.02. Although the global model may predict instances of missing organs (e.g., the kidney and liver), these detections can be pruned by thresholding the local classifier scores or by identifying missing organs as those with a low average boundary detector score.

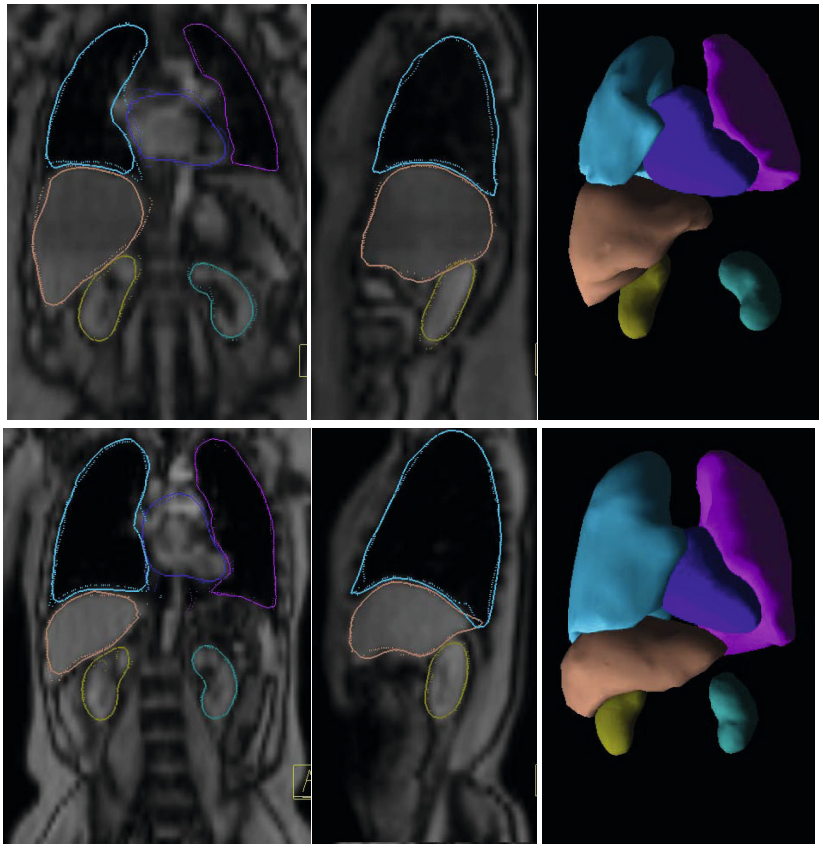
3.2 Prostate, Bladder, Rectum, Femoral Heads in CT Scans

In this second data set, we detect the prostate, bladder, rectum, and femoral heads in CT scans. The detection and segmentation of these structures is useful

[1] The inter-user variability was measured over 10 randomly selected unseen test cases.

Table 3. Accuracy and timing for segmentation results using our model compared to the state of the art MSL model on CT prostate, bladder, rectum and femoral heads

Detection, shape initialization, & boundary refinement							
	Skip(mm)	Time	Prostate	Bladder	Rectum	R.Fem	L.Fem
MSL	-	9.67s	3.57±2.01	**2.59±1.70**	4.36±1.70	1.89±0.99	2.05±1.27
BSP	10 (30mm)	1.76s	**3.35±1.40**	3.08±2.25	**3.97±1.43**	**1.88±0.78**	**1.90±1.18**
	12 (36mm)	1.36s	3.48±1.53	3.17±2.28	3.98±1.49	1.93±1.00	2.23±1.76
	15 (45mm)	1.09s	3.70±1.64	3.28±2.42	4.03±1.48	2.04±1.18	2.25±2.04
	Inter-user variability		3.03±1.15	2.03±0.11	2.93±1.10	1.29 ± 0.12	1.16±0.21



Front Side 3D Shapes

Fig. 4. Qualitative results of the MR FastView segmentation (solid) on unseen cases with ground truth (dotted)

for radiation therapy planning. This data exhibits challenges in weak boundaries between soft tissues, complex shapes in rectum and femoral head, large scale variation in bladder, etc. A total of 145 cases were used, with 100 randomly selected for training and the remaining 45 used in testing. The volumes were isotropically resampled to have a resolution of 3mm. Six manually selected landmarks were identified on each of the objects, with the exception of the bladder which used 7 as it had large variability. And a similar configuration as described in the previous section was used to train the local and global context models.

Table 3 shows the timing and accuracy results for the final segmentation compared to an MSL pipeline. Even with a spacing factor of 36mm, our local+global model behaves similarly to or better than MSL on all organs except for the bladder while giving an overall speedup of 6 times over MSL. MSL seems to better handle the large scale variability observed in the bladder. Our approach significantly outperforms MSL for rectum, possibly because of the aforementioned reason – the rectum shape varies a lot and landmark-based shape initialization is better. Both approaches achieved accuracy fairly close to the inter-user

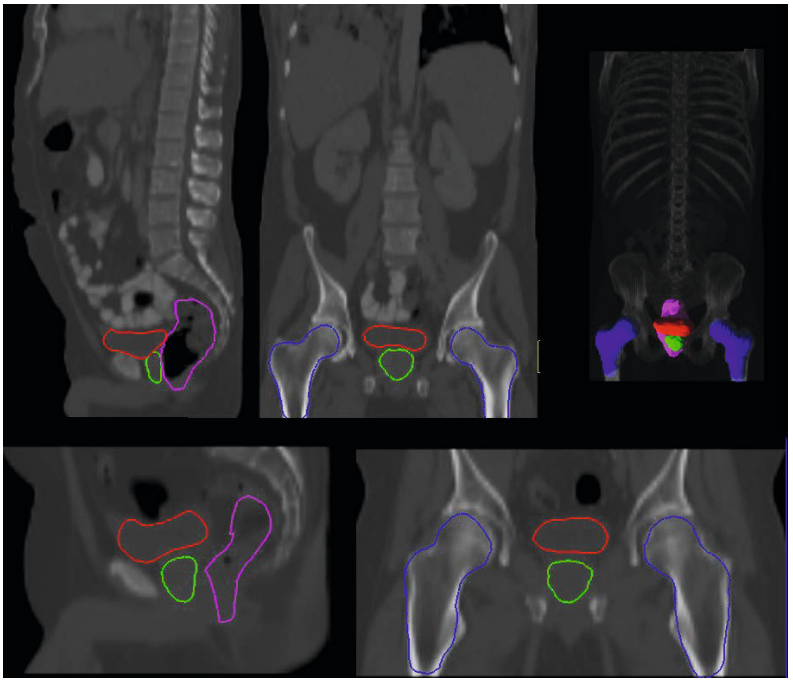


Fig. 5. An illustration of the segmentation results on two of the CT prostate data sets. The data has wildy varying dimensions, some being full body scans, and others localized near the prostate. Our method works well across this variation and handles large variability in shape and appearance of the organs, such as drastic changes in appearance in the rectum.

variability except the rectum[2]. We achieved a speed of just over one second by skipping every 16^{th} voxel with decent accuracy. Figure 5 illustrates two of the automatic CT segmentations on unseen images. The femoral accuracy of the femoral head is limited by the low resolution of our mesh and due to using a 3mm isotropic resolution. However, this serves as a good initialization for voxel-based refinement using graph cut or random walker.

4 Conclusion

In this work we proposed a fusion of local and global context, coupled with discriminative models, for rapid multi-organ segmentation. Exploiting sparsity of the non-parametric global context led to a fast algorithm: the global context is only evaluated at sparse regions and is used to predict hypotheses for all landmarks simultaneously. By robustly fitting statistical shape model to these landmarks and deforming the fitted shape using learned boundary detector, we achieved segmentation accuracy comparable to inter-user variability.

Although our approach is already efficient, we feel that there is still room for improvement. Specifically, the local detectors often get evaluated on the same voxel multiple times; a simple caching of classifier results could be used to improve efficiency. Along a similar line, if results are cached, there may also be benefit in having a multi-class classifier be used to model the local posterior. We will also investigate how to further improve the segmentation accuracy for organs with simple shape but large variability in appearance like kidney or in scale like bladder.

References

1. Yang, J., Duncan, J.S.: 3D image segmentation of deformable objects with joint shape-intensity prior models using level sets. Medical Image Analysis 8(3), 285–294 (2004)
2. Zheng, Y., Georgescu, B., Ling, H., Zhou, S.K., Scheuering, M., Comaniciu, D.: Constrained marginal space learning for efficient 3D anatomical structure detection in medical images. In: CVPR, pp. 194–201. IEEE (2009)
3. Ling, H., Zhou, S.K., Zheng, Y., Georgescu, B., Suehling, M., Comaniciu, D.: Hierarchical, learning-based automatic liver segmentation. In: CVPR (2008)
4. Zhou, S.K.: Shape regression machine and efficient segmentation of left ventricle endocardium from 2D b-mode echocardiogram. Medical Image Analysis 14(4), 563–581 (2010)
5. Kohlberger, T., Sofka, M., Zhang, J., Birkbeck, N., Wetzl, J., Kaftan, J., Declerck, J., Zhou, S.K.: Automatic multi-organ segmentation using learning-based segmentation and level set optimization. In: Fichtinger, G., Martel, A., Peters, T. (eds.) MICCAI 2011, Part III. LNCS, vol. 6893, pp. 338–345. Springer, Heidelberg (2011)
6. Shimizu, A., Ohno, R., Ikegami, T., Kobatake, H., Nawano, S., Smutek, D.: Segmentation of multiple organs in non-contrast 3D abdominal CT images. International Journal of Computer Assisted Radiology and Surgery 2, 135–142 (2007)

[2] The inter-user variability was measured over 5 randomly selected unseen test cases.

7. Sofka, M., Zhang, J., Zhou, S.K., Comaniciu, D.: Multiple object detection by sequential Monte Carlo and hierarchical detection network. In: CVPR, June 13-18 (2010)
8. Liu, D., Zhou, S.K., Bernhardt, D., Comaniciu, D.: Search strategies for multiple landmark detection by submodular maximization. In: CVPR. IEEE (2010)
9. Criminisi, A., Shotton, J., Bucciarelli, S.: Decision forests with long-range spatial context for organ localization in ct volumes. In: MICCAI-PMMIA Workshop (2009)
10. 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 Workshop MVC. LNCS, vol. 6533, pp. 106–117. Springer, Heidelberg (2011)
11. 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)
12. Tu, Z.: Probabilistic boosting-tree: Learning discriminative models for classification, recognition, and clustering. In: ICCV, pp. 1589–1596 (2005)
13. Friedman, J., Hastie, T., Tibshirani, R.: The elements of statistical learning. Springer Series in Statistics, vol. 1 (2001)
14. Datar, M., Indyk, P.: Locality-sensitive hashing scheme based on p-stable distributions. In: SCG 2004: Proceedings of the Twentieth Annual Symposium on Computational Geometry, pp. 253–262. ACM Press (2004)
15. Dasgupta, S., Freund, Y.: Random projection trees and low dimensional manifolds. In: Proceedings of the 40th Annual ACM Symposium on Theory of Computing, STOC 2008, pp. 537–546. ACM, New York (2008)
16. Cootes, T.F., Taylor, C.J., Cooper, D.H., Graham, J.: Active shape models their training and application. Comput. Vis. Image Underst. 61, 38–59 (1995)