Nearest Neighbor 3D Segmentation with Context Features

Evelin Hristova[α,β], Heinrich Schulz[β], Tom Brosch[β], Mattias P. Heinrich[γ], Hannes Nickisch[β]

[α]Hamburg University of Applied Sciences, Hamburg, Germany;
[β]Philips Research, Hamburg, Germany;
[γ]Universität zu Lübeck, Lübeck, Germany;

ABSTRACT

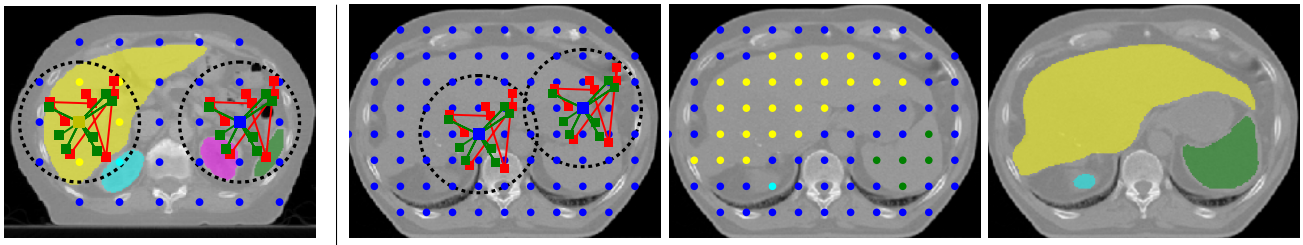
Automated and fast multi-label segmentation of medical images is challenging and clinically important. This paper builds upon a supervised machine learning framework that uses training data sets with dense organ annotations and vantage point trees to classify voxels in unseen images based on similarity of binary feature vectors extracted from the data. Without explicit model knowledge, the algorithm is applicable to different modalities and organs, and achieves high accuracy. The method is successfully tested on 70 abdominal CT and 42 pelvic MR images. With respect to ground truth, an average Dice overlap score of 0.76 for the CT segmentation of liver, spleen and kidneys is achieved. The mean score for the MR delineation of bladder, bones, prostate and rectum is 0.65. Additionally, we benchmark several variations of the main components of the method and reduce the computation time by up to 47% without significant loss of accuracy. The segmentation results are – for a nearest neighbor method – surprisingly accurate, robust as well as data and time efficient.

Keywords: Image Segmentation, Binary Context Features, Nearest Neighbor Classification, Vantage Point Tree

1. INTRODUCTION

Multi-label segmentation of 3D medical images from only a few training cases is of high importance to applications such as radiotherapy planning, because delineation of anatomical structures is challenging due to anatomical variability of organs in shape, size and appearance among patients and fuzzy boundaries between tissues. Manual classification of organs is time-consuming and subject to inter- and intra-observer variability. All existing segmentation methods (e.g., model-based segmentation,[1] atlas-based segmentation[2]) are subject to trade-offs in terms of accuracy, runtime, stability, generality, data efficiency, scalability and simplicity.

A recently proposed simple, generic, data and time efficient supervised machine learning method[3] employs a labeled training dataset to classify voxels in an unseen image. The segmentation algorithm does not rely on explicit shape models and is applicable to different organs and modalities. As the method yields surprisingly accurate results, we investigate in detail its three main components: binary context features, Vantage Point Tree (VPT)[4,5] nearest neighbor and Random Walker (RW) regularization,[6] and evaluate alternatives to each of them.



(a) Training feature extraction

(b) Test feature extraction

(c) Nearest neighbor search, labeling

(d) Label interpolation, spatial regularization

Figure 1: Nearest Neighbor Classification Pipeline

Feature vectors capture neighborhood information by comparing intensities within spatial proximity (dashed circles). Training and testing features are extracted on regular grids by using LBP (green lines) and BRIEF (red lines) sampling patterns. Classifying test features is done by finding a subset of most similar training samples (Nearest Neighbors). By retrieving their labels, probability maps are obtained for each available class in the training data. The probabilities are then interpolated across the image and spatially regularized. Finally, the labels with maximum likelihood are selected.

2. METHODS

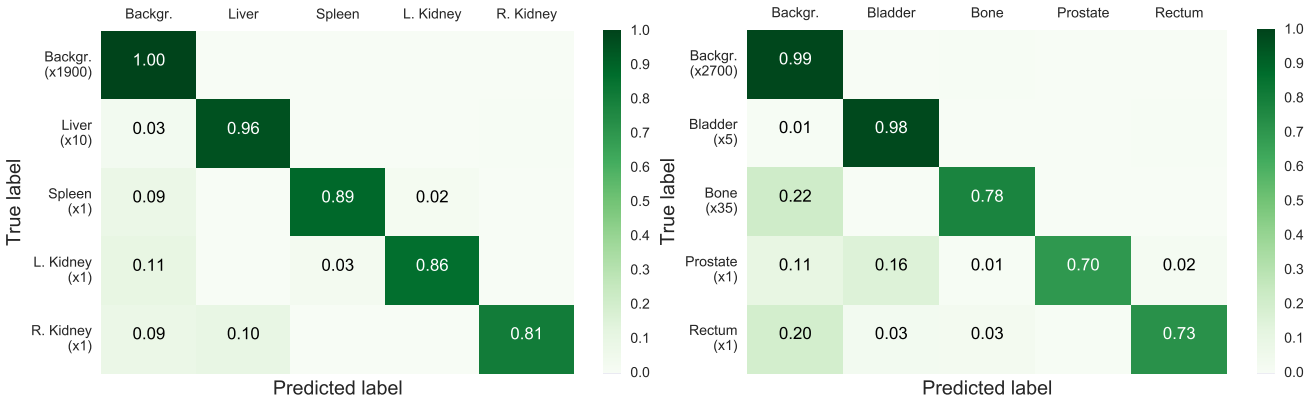
Our supervised learning approach has two independent phases: training and testing as illustrated in Figure 1. During training, the algorithm loads images with dense voxel annotations, paired with binary body contour masks that indicate regions relevant for feature sampling. After reducing noise by pre-smoothing the images, features are extracted on regular grids within the masks (1(a)). These vectors encode neighborhood information by comparing intensity pairs in spatial proximity (dashed circles in 1(a)). An intensity difference at location x is defined as $\Delta_i(x) := p(x + y_i) - p(x + z_i)$ where $p(x)$ is the pixel intensity. Using a sampling pattern $(y_i, z_i)\ \ i = 1..n$ relative to a location x, an n-dimensional binary feature vector $h(x)$ can be constructed by $h_i(x) = \text{sign}(\Delta_i(x))/2 + 1/2$. Using the sign, instead of the actual difference, and hence binary features, makes the descriptors robust to monotonic gray-level changes and also enables fast matching by calculating the Hamming distance $d_H(h, l) = ||h - l||_1$ between two vectors h and l.* A combination of two different feature descriptors is implemented here. On the one hand, Local Binary Patterns (LBP)[7] focus on relations around the central region (by setting $y_i = 0$) by comparing the intensity at x with intensities randomly selected in its spatial proximity (green lines in 1(a)). On the other hand, Binary Robust Independent Elementary Features (BRIEF)[8] capture interactions among neighboring structures by evaluating the intensities of the patches around (red lines in 1(a)).

In the test phase, classification of a previously unseen image requires the extraction of features with the same combination of sampling patterns (1(b)). Assigning labels to these test features is then based on the subset of training features with shortest distance to them, known as Nearest Neighbors (NNs). Finding these features can be achieved by going through all possible solutions (Exhaustive Search), however, in high dimensional classification, it would be inefficient. Instead, hierarchical tree structures can be exploited to optimize this procedure. The Vantage Point Tree (VPT)[4] has been shown to perform well in high dimensional nearest neighbor search[3,9] and is thus adopted here. VPTs are constructed by recursively splitting the data points using absolute distances from randomly chosen centers. These centers, called vantage points, partition the data at each iteration in such a way that approximately half of the points are within a distance threshold, and the other half are beyond it. This results in a structure where data points in the same sub-branch also tend to be neighbors in space and thus, searching for them is more efficient. Moreover, ensembles of VPTs increase accuracy without increasing search time too much. The labels of multiple NNs are retrieved to create probability maps for each available class in the training data (1(c)). These are then linearly interpolated across the whole image. Subsequently, the labels are spatially regularized by a multi-label Random Walker.[6] The RW combines the computed probability maps and the image intensities to obtain smooth segmentation. Finally, the labels with highest probabilities are selected (1(d)).

The above presented framework (with $h_i(x) = \text{sign}(\Delta_i(x))/2 + 1/2$, VPT classification and RW post processing) following the suggestions of the initial paper[3] is referred to as the "Defaults" pipeline here. In the search of higher accuracy and faster execution, for each of the main three steps two alternative solutions are proposed. Each of these modifications adopts the standard default configuration and replaces only one component at a time.

In the feature extraction phase, first a minimum intensity threshold τ is introduced yielding $h_i(x) = \text{sign}(\Delta_i(x) - \tau)/2 + 1/2$ ("Threshold"). Then experiments are also performed on the absolute intensity differences between the sampling pairs where $h_i(x) = \Delta_i(x)$ ("Difference"). In the nearest neighbor search step, the Vantage Point Tree is replaced by two other classification methods:[9] k-Means tree ("K-means") and k-d tree ("Kd tree"). All classifier methods are generic in the distance metric they use to match the feature vectors. Therefore, when absolute intensity differences are used to construct the features, the standard L_2 norm is calculated. In case of binary vectors, the Hamming distance is measured. Finally, the post processing (regularization) step is evaluated with a $3 \times 3 \times 3$ median filter ("Median filter"). Additionally, we explore a new strategy without spatial regularization ("No Regularization") in which under-classification is compensated. In this pipeline, background labels are only assigned if their probability is > 0.4. Otherwise, the label with second highest probability is chosen.

*The Hamming distance can be computed efficiently with the use of the `POPCNT` instruction in x86_64 architectures.



(a) CT (b) MR

Figure 2: Confusion Matrices

The two confusion matrices summarize the average outcome of the CT and MR data cross validations. The approximate amount of features compared to the organ with the least samples in each experiment is indicated below each class. The strong diagonals of the confusion matrices give an insight that most of the times voxels are correctly classified. The confusion with background can be explained by the prevalence of background training data. It can also be seen that the algorithm occasionally confuses some organs with others, e.g., bladder and prostate.

3. EXPERIMENTS AND RESULTS

3.1 Analyzed Data and Parameters

The default pipeline, as well as the alternative solutions, are applied to images from two different modalities, each with four different annotated organs. First, 70 abdominal CT images with liver, spleen, left and right kidneys segmentation, of size $512 \times 512 \times 394$ voxels ($1.37 \times 1.37 \times 1.36$ mm) are used for a 5-fold cross validation. Then, 42 male pelvic MRI T1 images ($528 \times 528 \times 120$ voxels, $1.05 \times 1.05 \times 2.5$ mm) with dense bladder, bone, prostate and rectum labels are used for a 7-fold cross validation.

For each of the pipelines, 1280 intensity comparisons are used to construct the feature vector, where 20-40% employ a BRIEF sampling pattern and 40-60% - LBP. The displacement distribution of the randomly selected points around the voxel of interest is normal with a standard deviation of [20/30; 80].† In the MR data, training features are extracted from each 4^{th} voxel (see Figure 1) and in the CT from each 6^{th} as the images are larger. The classification is based on the 20 NNs of the feature vectors extracted from each 2^{nd} (MR)/4^{th} (CT) voxel from the test images.

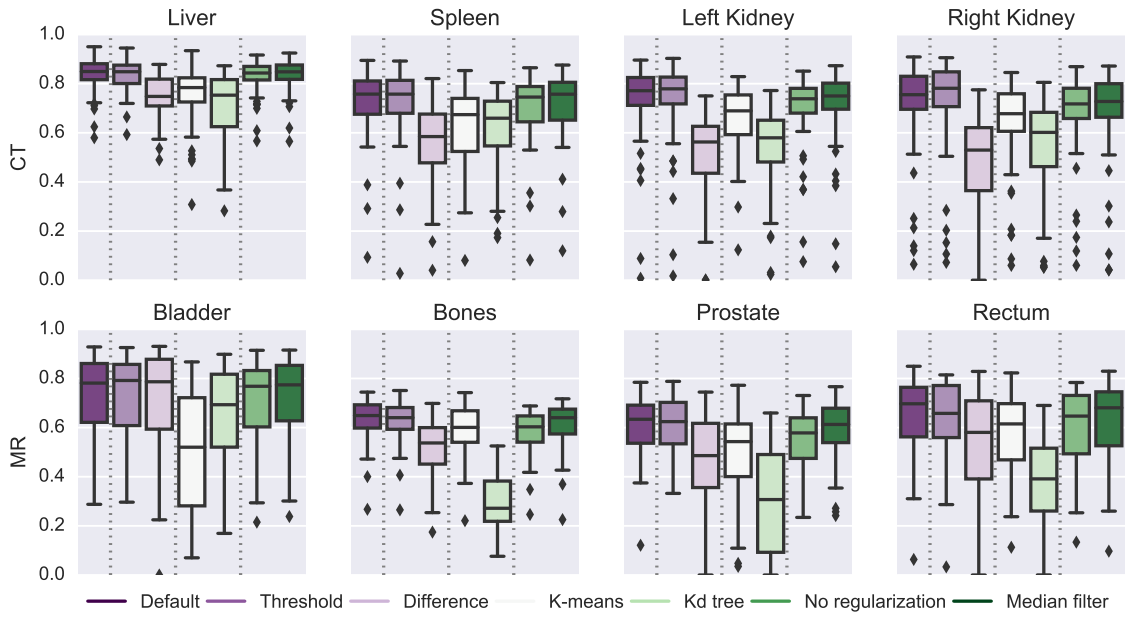
For each experiment 15 VP trees, containing the features, are either fully grown or terminated at a fixed leaf size of 15.

3.2 Evaluation

The evaluation of the segmentation is based on the spatial overlap with the ground truth classification. A summary of the outcomes from the cross validation procedures by the default pipeline can be seen in Figure 2. The confusion matrices give an overview on whether (and if true, how often) certain classes get confused with others. The high values on the diagonals in both cases indicate that voxels belonging to the organs are correctly classified in most cases. It can be observed that organs are occasionally labeled as background some of the time mainly due to the fact that background features dominate in the training data. Inter-organ confusion happens in very few cases for the CT experiments and slightly more often in the MR scenario.

The accuracy is then measured independently for foreground/background classes except for the background using the Dice overlap score $D = 2TP/(2TP + FP + FN)$, see Figure 3, where TP are the correctly positively

†The min and max values correspond to the std. dev. of the displacement distribution of the sampling points measured by voxels. These are adapted for images of size $512 \times 512 \times 394$. The pattern used for the MR data is adjusted to match the size and the anisotropy of the voxels.



	Liver	Spleen	L. Kidney	R. Kidney	Bladder	Bone	Prostate	Rectum
Default	**0.84**	**0.73**	0.73	0.72	**0.73**	**0.63**	**0.61**	**0.64**
Threshold	0.83	**0.73**	**0.74**	**0.73**	0.72	**0.63**	**0.61**	0.63
Difference	0.74	0.56	0.52	0.47	0.72	0.52	0.47	0.54
K-means	0.76	0.63	0.66	0.64	0.50	0.58	0.50	0.57
Kd tree	0.70	0.61	0.54	0.54	0.66	0.29	0.29	0.38
No regularization	0.83	0.70	0.70	0.68	0.70	0.58	0.55	0.59
Median filter	0.83	0.72	0.72	0.69	0.72	0.61	0.58	0.62

Figure 3: Cross Validation Evaluation, Dice Score Distribution

The Dice score distribution of the default framework shows satisfactory results for the segmentation of liver, spleen, kidneys (CT) and bladder (MR). The proposed alternative solutions do not improve the accuracy except in few cases.

classified voxels, FP are the falsely positively labeled and FN are the false negative ones. It can be observed that the distribution of the Dice score differs among the different modalities and organs even when the same framework is used. The relatively small spread of the results of the CT segmentation for liver and kidneys by the "Default", "Threshold", "No regularization" and "Median filter" pipelines shows that the algorithm is stable across these organs. In the MR classification of rectum and prostate, however, the wider spread indicates that the method achieves good segmentation for some of the images, but it is not robust. Nevertheless, the default choice of using binary features and VPTs performs best on average with a Dice of 0.76 for the CT images and 0.65 for the MR images. Some of the alternative solutions also deliver competitive results, e.g. "Threshold" and "Median filter".

3.3 Time Complexity

Runtime is another relevant factor when assessing the performance of the different pipelines, because time sensitive scenarios often require fast segmentation. Each test naturally depends on the amount and size of the data. The length of each step can also vary depending on the machine specification and computing power. Figure 4 gives an insight into the duration of the main steps of the segmentation for the given complete data set with the default pipeline. It can be observed that the runtime for both MR and CT images is approximately 2 minutes, but the intermediate steps differ between the two modalities. The CT images used here are larger as compared to the MR data, but less feature vectors are extracted from them (CT: training every 6^{th}, testing every 4^{th} voxel; MR: training every $4th$, testing every 2^{nd}). Therefore, extracting the features as well as constructing and

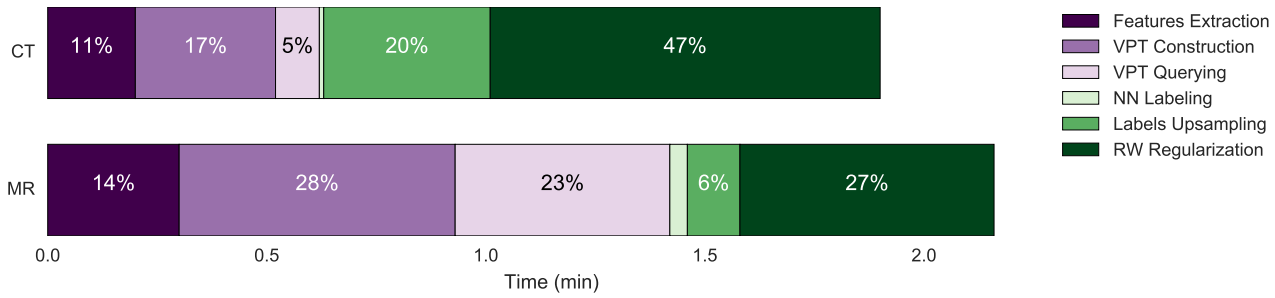


Figure 4: Default Pipeline Segmentation Duration

The duration of the segmentation phases is dependent on the size of the data. The CT images used here have more voxels, but less features are extracted from them. Hence, construction and querying of VPTs is faster, but up-sampling the probability maps and their spatial regularization takes longer time.

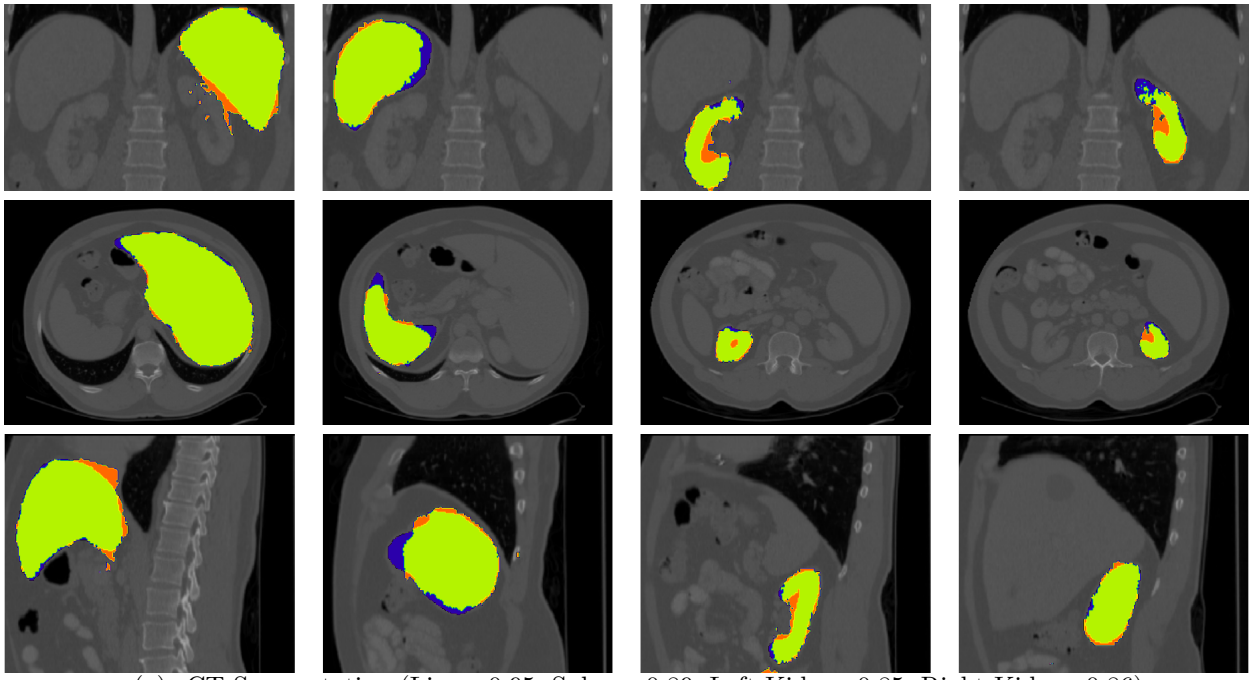
querying the VPTs is relatively fast. Due to the large number of voxels, however, the up-sampling of the label probability maps and especially the regularization are more time-consuming. The RW takes approximately 47% of the CT and up to 27% of the MR classification duration and yet it does not significantly improve the accuracy compared to median filtering. The results in 3 show that omitting the regularization and compensating the under-segmentation reduces the Dice score by no more than 6% (on average). Hence, using the "No Regularization" pipeline would be more suitable when fast coarse classification is of interest. Nonetheless, both frameworks are feasible using a standard computer and hence, meet the constraints of a typical clinical setting.

3.4 Segmentation Visualization

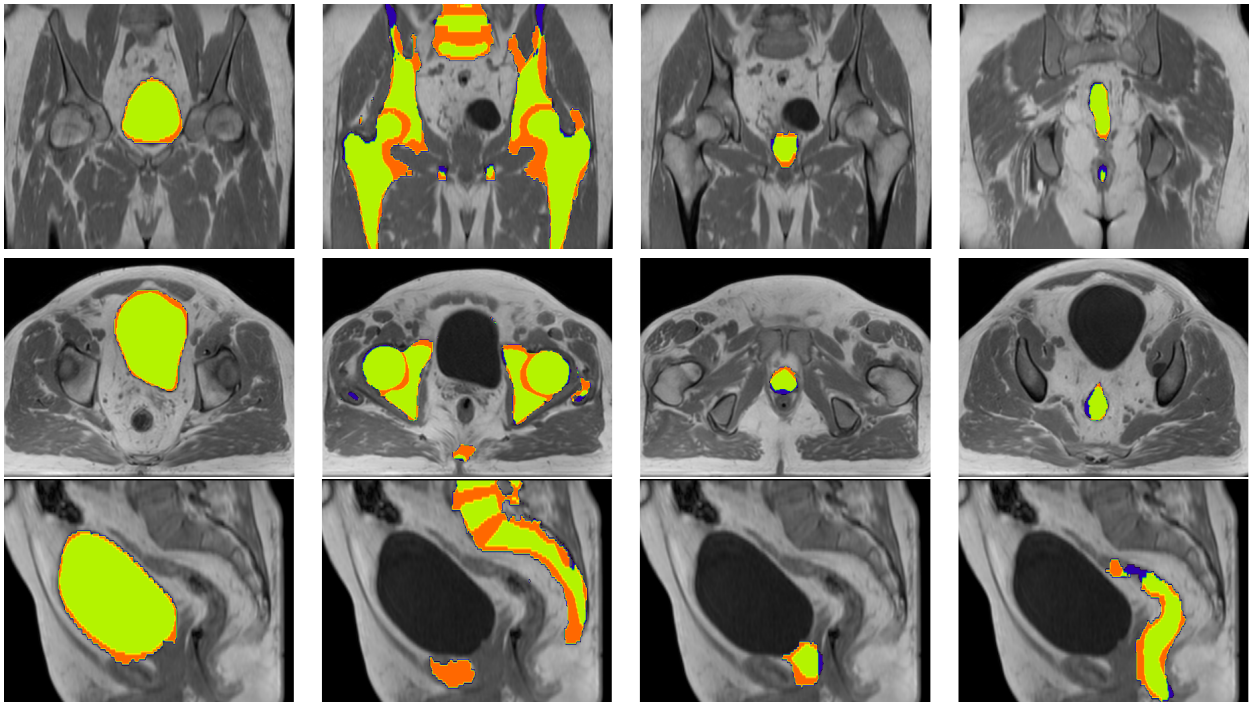
The results achieved by the default segmentation method for one CT and one MR image can be seen in Figure 5. For each organ in both modalities, sagittal, axial and coronal views are given. It can be observed that most of the voxels are correctly classified (green in 5). Occasionally some background voxels are labeled as organs (red in 5). Under-segmentation is rare (blue in 5). Additionally, Figure 6 illustrates two outlier cases (1 CT and 1 MR) where the segmentation has relatively lower Dice scores. The segmentation method successfully identifies the locations of all organs. However, as the organs often vary in shape, size and appearance, the algorithm has difficulties in finding the exact boundaries.

4. CONCLUSIONS

We evaluate a fully automatic multi-organ segmentation method on a large dataset and benchmark variations of it. The default VPT framework with RW regularization is able to classify binary features extracted from different imaging modalities with high accuracy. The hypothesis that binarization yields higher robustness against contrast variations is confirmed when comparing the results of "Difference". Alternative solutions to the feature construction, the classification method and post processing do not result in higher precision. However, an alternative post processing technique reduces computation time by up to 47% which underlines the clinical relevance of this generic approach.



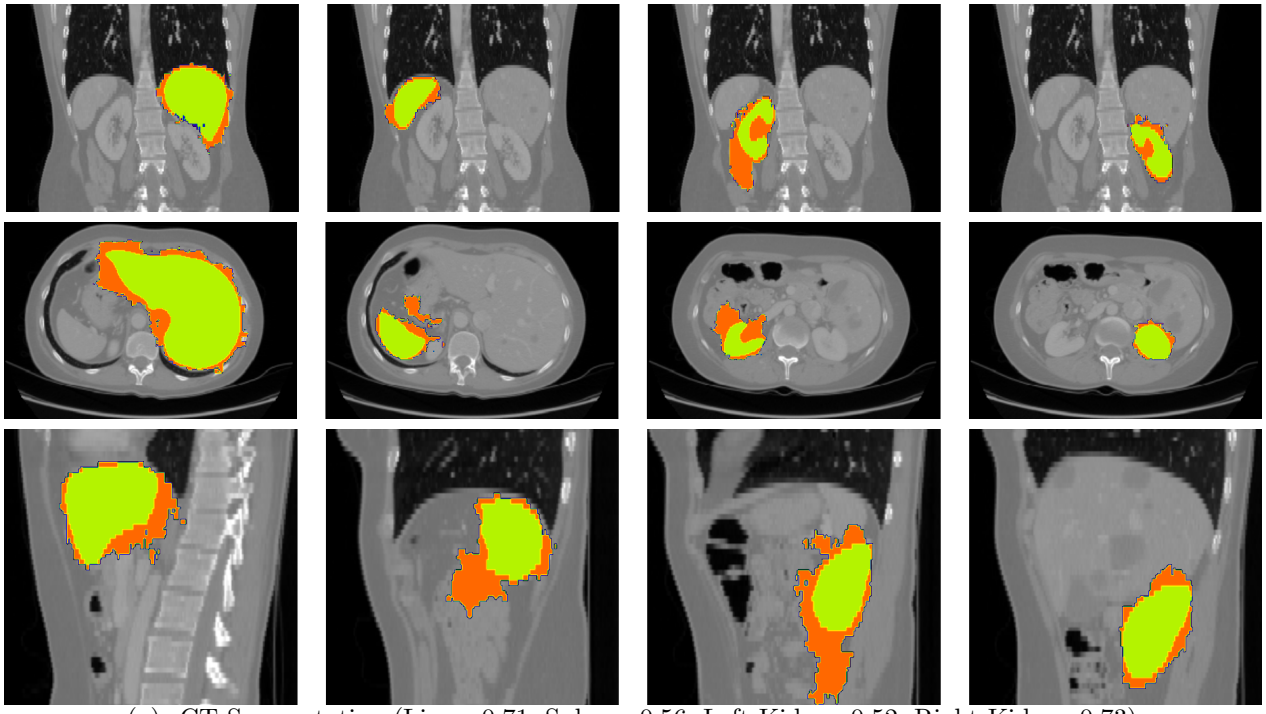
(a): CT Segmentation (Liver: 0.95, Spleen: 0.89, Left Kidney 0.85, Right Kidney 0.86)



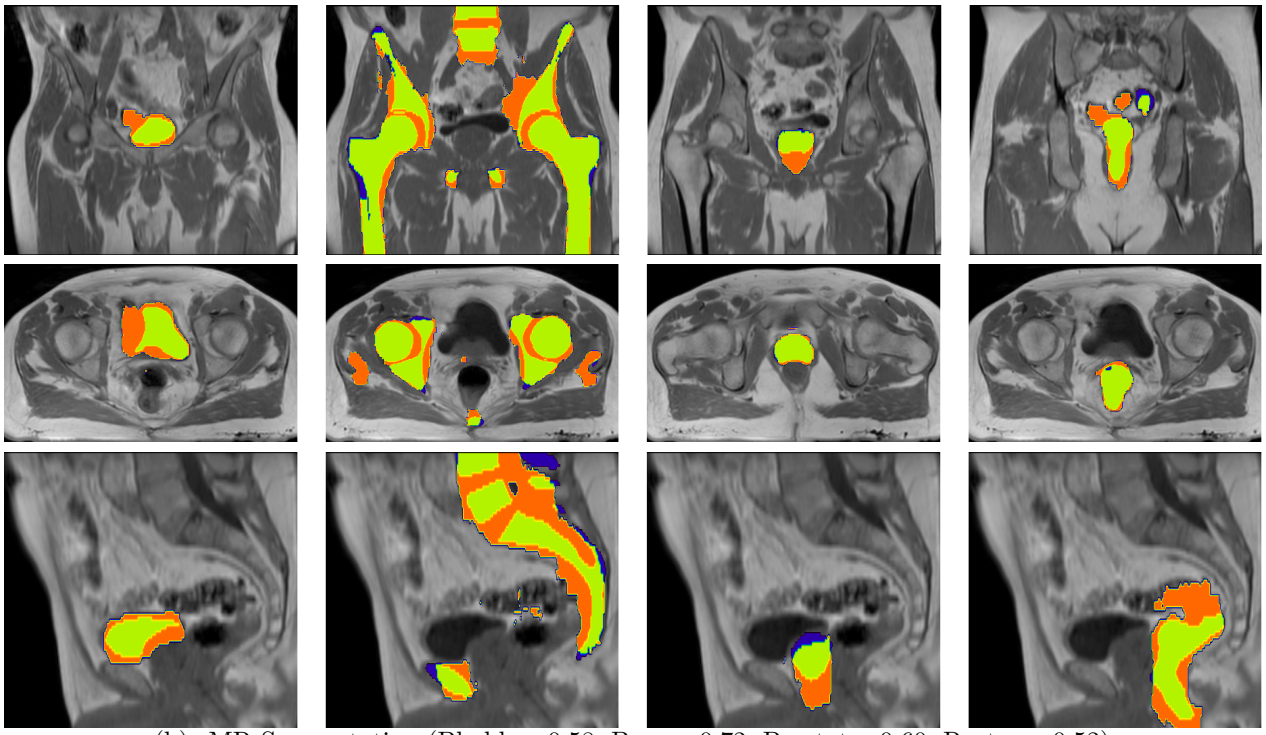
(b): MR Segmentation (Bladder: 0.91, Bones: 0.72, Prostate: 0.66, Rectum: 0.77)

Figure 5: Ground Truth vs. Algorithm Segmentation

Segmentation by the default pipeline on CT and MR images and Dice score per organ. Outcomes of the classification are indicated by colors: TP in green, FP in red, FN in blue. For each organ in both modalities, sagittal, axial and coronal views are given.



(a): CT Segmentation (Liver: 0.71, Spleen: 0.56, Left Kidney 0.52, Right Kidney 0.73)



(b): MR Segmentation (Bladder: 0.58, Bones: 0.72, Prostate: 0.60, Rectum: 0.52)

Figure 6: Ground Truth vs. Algorithm Segmentation (**Outlier**)

Segmentation by the default pipeline on CT and MR images and Dice score per organ. Outcomes of the classification are indicated by colors: TP in green, FP in red, FN in blue. For each organ, sagittal, axial and coronal views are given.

REFERENCES

[1] Ecabert, O., Peters, J., Schramm, H., Lorenz, C., von Berg, J., Walker, M. J., Vembar, M., Olszewski, M. E., Subramanyan, K., Lavi, G., et al., "Automatic model-based segmentation of the heart in ct images," *IEEE transactions on medical imaging* **27**(9), 1189–1201 (2008).

[2] Aljabar, P., Heckemann, R. A., Hammers, A., Hajnal, J. V., and Rueckert, D., "Multi-atlas based segmentation of brain images: atlas selection and its effect on accuracy," *Neuroimage* **46**(3), 726–738 (2009).

[3] Heinrich, M. P. and Blendowski, M., "Multi-organ segmentation using vantage point forests and binary context features," in [*International Conference on Medical Image Computing and Computer-Assisted Intervention*], 598–606, Springer (2016).

[4] Yianilos, P. N., "Data structures and algorithms for nearest neighbor search in general metric spaces," in [*SODA*], **93**(194), 311–21 (1993).

[5] Uhlmann, J. K., "Satisfying general proximity/similarity queries with metric trees," *Information processing letters* **40**(4), 175–179 (1991).

[6] Grady, L., "Multilabel random walker image segmentation using prior models," in [*Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*], **1**, 763–770, IEEE (2005).

[7] Ahonen, T., Hadid, A., and Pietikainen, M., "Face description with local binary patterns: Application to face recognition," *IEEE transactions on pattern analysis and machine intelligence* **28**(12), 2037–2041 (2006).

[8] Calonder, M., Lepetit, V., Strecha, C., and Fua, P., "Brief: Binary robust independent elementary features," in [*European conference on computer vision*], 778–792, Springer (2010).

[9] Kumar, N., Zhang, L., and Nayar, S., "What is a good nearest neighbors algorithm for finding similar patches in images?," in [*European conference on computer vision*], 364–378, Springer (2008).

[10] Muja, M. and Lowe, D. G., "Fast matching of binary features," in [*Computer and Robot Vision (CRV), 2012 Ninth Conference on*], 404–410, IEEE (2012).

[11] Muja, M. and Lowe, D. G., "Scalable nearest neighbor algorithms for high dimensional data," *IEEE Transactions on Pattern Analysis and Machine Intelligence* **36**(11), 2227–2240 (2014).