In MICCAI 2014 Workshop on
Image-Guided Adaptive Radiation

On the Importance of Location and Features for the Patch-Based Segmentation of Parotid Glands

Christian Wachinger[1,2], Matthew Brennan[1], Greg C. Sharp[2], Polina Golland[1]

[1]Computer Science and Artificial Intelligence Lab, MIT
[2]Massachusetts General Hospital, Harvard Medical School

Abstract. The segmentation of parotid glands in CT scans of patients with head and neck cancer is an essential part of treatment planning. We introduce a new method for the automatic segmentation of parotid glands that extends existing patch-based approaches in three ways: (1) we promote the use of image features in combination with patch intensity values to increase discrimination; (2) we work with larger search windows than established methods by using an approximate nearest neighbor search; and (3) we demonstrate that location information is a crucial discriminator and add it explicitly to the description. In our experiments, we compare a large number of features and introduce a new multi-scale descriptor. The best performance is achieved with entropy image features in combination with patches and location information.

1 Introduction

Radiation therapy planning aims to maximize the radiation dose in the target region while minimizing the dose in surrounding tissue. In intensity modulated radiation therapy, experts delineate the most critical structures, also known as organs at risk, and use the generated segmentations to reduce the irradiation of healthy tissue and potential side effects. The parotid glands are critical salivary glands and organs at risk for treating patients with head and neck cancer. The irradiation of the parotid glands can lead to xerostomia, a condition that interferes with mastication, deglutition, and speech in patients. The automatic segmentation of parotid glands is particularly challenging due to the low soft-tissue contrast in CT images and the high anatomical variability of the glands among patients.

In this study, we propose a new atlas-based method to automatically segment the parotid glands of patients with head and neck cancer. Instead of using deformable registration to create correspondences between test and training images, as is common in atlas-based approaches, we establish correspondences by directly comparing the image content of small regions. If image intensities are used to represent the image content, this leads to patch-based segmentation methods. Intensity values are, however, just one possible way to describe image content. We present a natural generalization of patch-based approaches using image features to extract additional discriminative information. We investigate the optimal selection and integration of these features.

We build upon the non-local means (NLM) framework [2] for patch-based segmentation, which produces state-of-the-art segmentation results [4, 14]. The idea behind NLM is to compare patches across the entire image domain and to let the comparison only depend on patch intensity values, and not on location. In the actual implementation of NLM for image denoising, the search window is restricted to 21×21 pixels to address computational concerns [2]. Similarly, [4, 14] restrict the search window to range from $9 \times 9 \times 9$ to $15 \times 15 \times 15$ voxels to improve computational efficiency. In our study, we employ an efficient approximate nearest

neighbor search allowing us to work with larger search windows that contain the entire parotid gland. Counter-intuitively, *larger search windows lead to less accurate segmentations*. This suggests that the spatial information implicitly incorporated into the comparison of patches on restricting the search to small windows not only improves computational efficiency but also has a direct influence on segmentation accuracy. Contrary to the idea behind NLM, we explicitly include location information in the comparison of patches as a descriptor, acting as a soft constraint towards spatially closer patches. We find that considering location explicitly in this way yields a significant improvement in segmentation results. This finding demonstrates the importance of spatial information in patch-based segmentation and reaffirms our conclusion that small search windows have a positive influence on segmentation accuracy.

The contributions of this work are: (1) an evaluation of different image features as descriptors in a patch-based segmentation approach; (2) a modification of the NLM framework for patch-based segmentation to use larger search windows in an approximate nearest neighbor search; and (3) an explicit integration of location information as a descriptor. All of our experiments segment the parotid glands of patients undergoing radiation therapy.

1.1 Related Work

The atlas-based segmentation of parotid glands with deformable registration was applied in [7, 13]. In [3], atlas images were used to train an active shape model used to segment the parotid glands. The refinement of head and neck segmentations based on classification with features was proposed in [12]. In [6], label fusion was used to initialize a segmentation pipeline employing statistical appearance models and geodesic active contours. Patch-based segmentation approaches as described within the NLM framework were proposed in [4,14]. In previous work, we used a patch-based approach to segment the parotid glands using the NLM framework and a random forest classifier [16]. We also refined the initial segmentations based on image contours with a Gaussian process regression. Sparse coding is a related extension of patch-based segmentation which was combined with the Haar-wavelet, histogram of oriented gradients and local binary patterns image features in [9].

2 Method

2.1 Review of Non-Local Means Segmentation

Given an atlas $\mathcal{A} = (\mathcal{I}, \mathcal{S})$ that contains images $\mathcal{I} = \{I_1, \ldots, I_n\}$ and their corresponding segmentations $\mathcal{S} = \{S_1, \ldots, S_n\}$ over a common image domain Ω, our objective is to compute the segmentation S of a new image I. Patch-based methods are based on the rationale that locations with similar image content should have similar segmentations, where local image content is represented by the intensity values in a patch centered at each voxel. Consider a patch $P(\mathbf{x})$ from the test image at a location $\mathbf{x} \in \Omega$ and the collection of all patches in the training images $\mathcal{P}$. We find the closest patch $P_{\mathbf{x}}$ in the training set

$$P_{\mathbf{x}} = \arg\min_{P \in \mathcal{P}} \|P(\mathbf{x}) - P\|_2. \tag{1}$$

Associated to the image patch $P_{\mathbf{x}}$ is the segmentation patch $S_{\mathbf{x}}$, which is used to infer the segmentation $S(\mathbf{x})$ in the test image around that location. In addition to finding the nearest neighbor $P_{\mathbf{x}} = P^1_{\mathbf{x}}$, we identify the full set of k-nearest neighbor patches $P^1_{\mathbf{x}}, \ldots, P^k_{\mathbf{x}}$. We differentiate between two methods of label propagation: (1) point-wise (PW) estimation which only considers the center location of the patch $S_{\mathbf{x}}[\mathbf{x}]$; and (2) multi-point (MP) estimation [14],

which considers the entire segmentation patch $S_{\mathbf{x}}$. The label map is computed under the two approaches as

$$L^{\mathrm{PW}}(\mathbf{x}) = \frac{\sum_{i=1}^{k} w(P(\mathbf{x}), P_{\mathbf{x}}^{i}) \cdot S_{\mathbf{x}}^{i}[\mathbf{x}]}{\sum_{i=1}^{k} w(P(\mathbf{x}), P_{\mathbf{x}}^{i})}, \tag{2}$$

$$L^{\mathrm{MP}}(\mathbf{x}) = \frac{\sum_{\mathbf{y}\in\mathcal{N}_{\mathbf{x}}} \sum_{i=1}^{k} w(P(\mathbf{y}), P_{\mathbf{y}}^{i}) \cdot S_{\mathbf{y}}^{i}[\mathbf{x}]}{\sum_{\mathbf{y}\in\mathcal{N}_{\mathbf{x}}} \sum_{i=1}^{k} w(P(\mathbf{y}), P_{\mathbf{y}}^{i})}, \tag{3}$$

where $\mathcal{N}_{\mathbf{x}}$ is the patch neighborhood around $\mathbf{x}$ and $S_{\mathbf{y}}[\mathbf{x}]$ is the label on the location $\mathbf{x}$ of the segmentation patch $S_{\mathbf{y}}$ centered at $\mathbf{y}$. The weight w is defined as

$$w(P, P') = \exp\left(-\frac{\|P - P'\|_2^2}{2\sigma^2}\right). \tag{4}$$

To obtain the segmentation of the image I, a label map is calculated for each parotid gland and each voxel is assigned the label with the most votes.

2.2 Descriptor-Based Segmentation

We extend patch-based segmentation to *descriptor-based segmentation* by including image features and location information as further descriptors of image content. Image features can capture additional information about contours, gradients, and texture in the image. The evaluated features are described in section 2.4. We also include location information in the descriptor by adding the xyz-coordinates of the voxel $\mathbf{x}$. Location information imposes a soft spatial constraint on the nearest neighbor search which is especially important when working with large search windows, as described in section 2.3. The descriptor vector $D(\mathbf{x})$ is the concatenation of a patch $P(\mathbf{x})$, an image feature $F(\mathbf{x})$, and location information $L(\mathbf{x})$

$$D(\mathbf{x}) = \begin{pmatrix} \frac{1}{|P(\mathbf{x})|\cdot\sigma_P^2} P(\mathbf{x}) \\ \frac{f}{|F(\mathbf{x})|\cdot\sigma_F^2} F(\mathbf{x}) \\ \frac{\ell}{|L(\mathbf{x})|\cdot\sigma_L^2} L(\mathbf{x}) \end{pmatrix}, \tag{5}$$

where f and ℓ are positive weights and each sub-vector is normalized by dividing by the number of entries $|\cdot|$ and the corresponding variance $\sigma_\cdot^2$. These variances are calculated for each sub-vector over the entire training set. This normalization permits direct control over the expected contributions of each descriptor type to the magnitude of the squared distance $\|D - D'\|_2^2$ by varying f and ℓ. The weight w in the label propagation is calculated with

$$w(D, D') = \exp\left(-\frac{\|D - D'\|_2^2}{2(1 + f + \ell)}\right). \tag{6}$$

2.3 Nearest Neighbor Search

We evaluate three approaches to performing the k-nearest neighbor search in Eq. (1): a full, a bounded and an approximate k-nearest neighbor search. The full nearest neighbor (FNN) search searches over all locations in the domain of the organ to find nearest neighbors, following the original idea behind non-local means. The bounded nearest neighbor (BNN) search searches over all locations $\mathbf{y}$ within a sphere of radius r ($\|\mathbf{y} - \mathbf{x}\|_2 \leq r$). This approximates

the search windows used in [4,14], where search is restricted to boxes of between $9 \times 9 \times 9$ and $15 \times 15 \times 15$ voxels to reduce computation time. To achieve a similar behavior, we set $r = 5$.

Disadvantages to these approaches include the high computational complexity of FNN and the hard spatial cut-off imposed by BNN during search. As a compromise, we consider an unbounded approximate nearest neighbor (ANN) search. We use the randomized kd-tree algorithm implemented in FLANN [11]. Although the kd-tree algorithm is a frequently used ANN, its performance generally decreases with high dimensional data. While this is true for randomly generated data, it has been shown that the performance of kd-trees is high for high dimensional data from image patches, likely due to strong correlations in images [11]. The randomized kd-tree algorithm splits data on a dimension randomly chosen among the dimensions of highest variance, rather than that of highest variance as in the classic kd-tree algorithm. Searching over multiple randomized kd-trees generally improves performance [11]. The randomized kd-tree algorithm in FLANN commonly provides more than 95% of the correct neighbors and is two or more orders of magnitude faster than the exact search [11].

2.4 Image Features

For most of the image features considered, we process the entire image first to produce a feature image and then extract a patch from the feature image, *e.g.*, for filtering, the feature corresponds to a patch of the filtered version of the image. The sizes of patches taken from the feature image range from $1 \times 1 \times 1$ to $5 \times 5 \times 5$. For some features, we use the size of the intensity patch $P(\mathbf{x})$, which is $9 \times 9 \times 5$ for most experiments.

Multi-Scale Patch: A disadvantage of patch-based approaches is the limited information about spatial context, which leads to undesirable pairings in the k-nearest neighbor search. We propose a new multi-scale patch that combines high resolution at its center and low resolution in the surrounding area (see figure on the right). In addition to the standard intensity patch $P(\mathbf{x})$ in the center, we consider a $3 \times 3 \times 3$ grid of blocks of the same size as $P(\mathbf{x})$ centered at $\mathbf{x}$ and the mean intensities of each block. This yields a vector of length 27 containing these mean values which we concatenate with the vector of patch intensity values $P(\mathbf{x})$. This design is motivated by the human visual system, where spatial acuity peaks at the central fovea and diminishes with distance. In this study, we consider only two scales; however, this feature has a natural extension to additional scale levels.

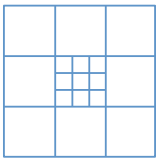

Image Filtering: Additional image features are obtained by filtering the images and extracting patches from the filtered images. We consider mean, median, variance, Sobel, Gaussian, Laplacian and Gabor wavelet [8, 10] filters. The mean, median and Gaussian filters we apply have masks of size $9 \times 9 \times 5$.

Entropy Image: Entropy images were proposed for the multi-modal registration of images in [15]. The information content of a patch is measured with the Shannon entropy which is computed and stored at the center voxel of the patch. Repeating this calculation for all voxels in the image yields the entropy image, which represents the structural information in the image. Entropy image features have similarities to gradient magnitude features. However, entropy image features provide a representation less dependent on actual intensity values.

Histogram of Oriented Gradients: To compute histogram of oriented gradients (HoG) features, we calculate 3D image gradients in each patch of the image [5]. These gradients are used to produce a histogram over gradient orientations, where the contribution of each gradient to the histogram is equal to its magnitude. Gradients created from image noise therefore have lower impact than strong gradients at image boundaries. The histograms produced have 8 bins corresponding to the 8 octants that the 3D vector can lie in.

Multi-scale Probability of Boundary: We compute the multi-scale probability of boundary (mPb) as described in [1]. In the first step, we estimate image and texture gradients per slice with the oriented gradient signal. This method calculates the χ^2 distance between the histograms of two half-discs at each location for various orientations and at multiple scales. Textons are calculated to quantify the texture by convolving the image with 17 Gaussian derivative and center-surround filters and subsequently clustering with k-means into 64 classes [1]. Image and texture gradients of multiple scales are added to yield the multi-scale probability of boundary.

Local Binary Patterns: Local binary patterns (LBP) measure the relations between a voxel and its neighbors, encoding these relations into a binary word and quantifying the texture in a local region. LBP is primarily used for 2D images. We compared a 2D implementation applied on all slices in the volume with a 3D extension of LBP. The 3D extension was motivated by spatio-temporal 2D+t video analysis and was implemented to use three orthogonal planes (LBP-TOP). The concurrence statistics for these three planes are concatenated. We obtained a better performance from the 2D implementation and report only its results.

Haar-like Features: Haar-like features are computed by considering adjacent rectangular regions at a specific location in a detection window, summing the pixel intensities in each region and calculating the difference between these sums. The key advantage of Haar-like features over most other features is their low computation time. Integral images permit the rapid calculation of these features at many scales. Haar-like features bear a certain similarity to Haar basis functions but also consider patterns that are more complex than Haar filters.

3 Experiments

To evaluate each of the methods described in section 2 and the effect of augmenting with each of the features in section 2.4, we test on a data set of 18 CT scans of patients with head and neck cancer. Each image was labeled by a trained anatomist for treatment planning. The images contain between 80 and 200 axial slices with a slice thickness of 2.5mm. All images were resampled to an in-plane resolution of 0.976mm. All 18 images have the left parotid labeled. The right parotid gland was consumed by a tumor in one patient. Three of the 18 patients have dental artifacts that obscure the image intensity values in a region around the parotid gland. We segment the left and right parotid glands in each of the 18 images with a leave-one-out procedure, using the remaining 17 images to generate the atlas. To limit the number of patches, we only consider every second patch in the training set in a way similar to [14]. We measure segmentation quality by calculating the dice volume overlap score and modified Hausdorff distance between the automatic and manual segmentations.

The standard configuration for our experiments uses $9 \times 9 \times 5$ patches, computes $k = 10$ nearest neighbors and includes location information. We also consider omitting location information, varying the parameter k, and patches of sizes ranging from $3 \times 3 \times 1$ to $9 \times 9 \times 5$. We work with anisotropic patches to account for anisotropy in the CT data. We threshold the image at -100 and 150 Hounsfield units to lessen the effects of dental artifacts and image noise on the computed distances between descriptors.

3.1 Evaluation of Location and Label Propagation Methods

First we evaluate the inclusion of location information in the descriptor using a patch size of $9 \times 9 \times 5$. We also compare point-wise and multi-point label propagation methods. Figure 1 shows a statistical analysis of the segmentation results for these methods applied to

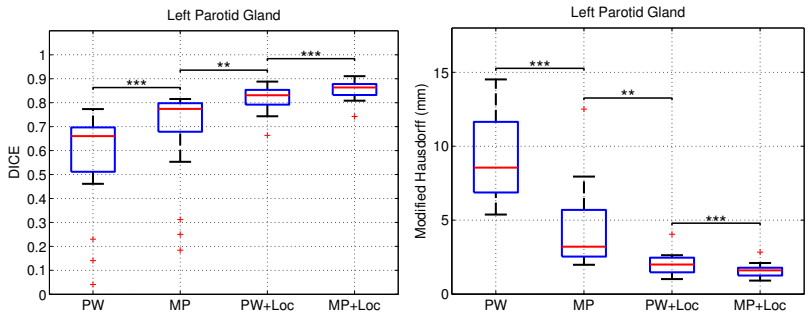

Fig. 1. Comparison of dice volume overlap and modified Hausdorff distances (MHD) for point-wise (PW), multipoint (MP) and the inclusion of location information (+Loc) for the left and right parotid glands. In the plots, the red line indicates the median, the boxes extend to the 25th and 75th percentiles, and the whiskers reach the most extreme values not considered outliers. *, ** and *** indicate significance levels at 0.05, 0.01 and 0.001, respectively.

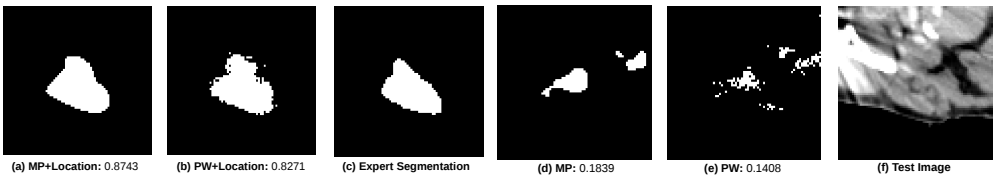

Fig. 2. Comparison of segmentation results for left parotid gland of a patient with dental artifacts. The four segmentation methods evaluated are: (a) multi-point with location; (b) point-wise with location; (d) multi-point; and (e) point-wise. We also show the Dice scores for this subject for each method. The expert segmentation is shown in (c) and the CT slice with dental artifacts in (f).

the left parotid gland quantified with Dice and Modified Hausdorff distance. We measure a significant improvement using multi-point (MP) label propagation over point-wise (PW) label propagation, which is consistent with results in [14]. We further observe a significant improvement for including location information (Loc) in the descriptor. On including location information, multi-point (MP+Loc) still yields a significant improvement over point-wise label propagation (PW+Loc).

As shown in Figure 1, there are three outlying Dice scores in the results of the point-wise and multi-point labeling of the left parotid. These outliers correspond to patients with dental artifacts. Figure 2 shows qualitative segmentation results for one of the subjects with dental artifacts together with the corresponding Dice scores. The CT slice of the test image shown demonstrates the strong impact of the dental artifact on the image. Including location information yields a clear improvement in the generated segmentation as illustrated by Figure 2 and the Dice increase of roughly 0.7. In this case, location information spatially regulates the segmentation, impeding the selection of patches from distant locations in the training images that have similar intensity patches due to the artifacts but correspond to a different anatomical structure. Furthermore, the multi-point approach smoothes the generated segmentation along the boundary of the parotid gland and yields a single connected component.

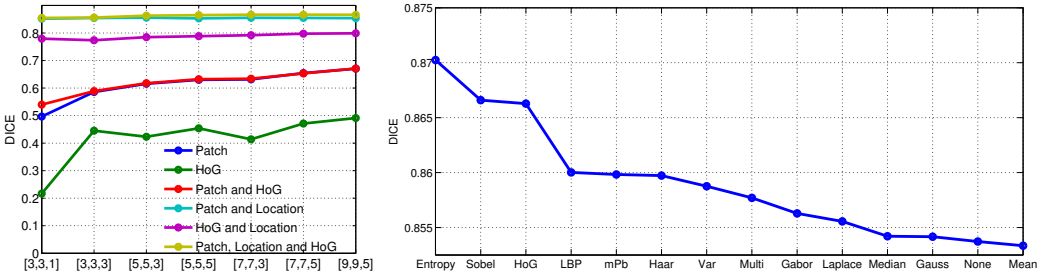

Fig. 3. Left: Mean Dice volume overlap scores for segmentations of the left parotid generated such that $D(\mathbf{x})$ contains: (1) patch intensity values; (2) HoG features; (3) patch intensity values and location information; (4) patch intensity values and HoG features; (5) HoG features and location information; and (6) patch intensity values, location information and HoG features. Right: Comparison of different features used in addition to the intensity patch and the location information. Plot of the mean Dice.

3.2 Comparison of Features

In this section, we evaluate the inclusion of image features into the descriptor, in addition to patch intensity values and location information. Figure 3 plots the mean Dice scores for several different classes of compositions of the descriptor $D(\mathbf{x})$ against patch size, using HoG as a representative feature. Using HoG alone as the descriptor leads to a worse performance than using only patch intensites. However, combining both HoG and patch intensities yields an improvement. Adding location information improves the results of all three of these combinations, resulting in an upward translation of their respective patch-size-Dice curves. The best results are achieved on including patch intensities, image features and location information in the descriptor. Based on these results, we select a patch size of $9 \times 9 \times 5$ and include all three sub-vectors in the descriptor $D(\mathbf{x})$ to test the performance of each feature.

Figure 3 shows the segmentation results for the left parotid gland on including the features from section 2.4. We omit the results for the right parotid which are similar to those for the left due to space constraints. Entropy image features perform considerably better than all other image features; and Entropy image, Sobel and HoG features are the three image features with the highest mean Dice scores. The only feature that performs slightly worse than including no additional image features in $D(\mathbf{x})$ is the mean image. The major difference between the results for the left and right parotids is that LBP is one of the better performing features for the left parotid but one of the worse performing features for the right, dropping from 4th to 10th place in relative feature rankings.

Other than these differences, the relative order of the performances of each feature is consistent from the left to right parotid glands. The three worst-performing features for both parotid glands are the median, Gaussian and mean filtered image features, all of which are features extracted from smoothed versions of the original image. The best performing features measure contours in the images (entropy, Sobel, HoG, mPb). It seems reasonable that adding contour information to the descriptor improves performance since this highlights the change from foreground to background in patches. Instead of only matching patches that have overall a similar appearance, this also ensures that they show similar contours. The performance for texture measures such as LBP is not as consistent between the left and right parotids. An interesting direction for future research is to investigate combinations of several features.

4 Conclusions

We introduced a descriptor-based approach for image segmentation, focusing on the identification of parotid glands in head and neck images, and proposed a descriptor containing patch intensity values, image features and location information. We also proposed to use an approximate nearest neighbor search for non-local means segmentation which enabled us to use much larger search windows than previous studies using NLM. Our results demonstrate the importance of location information when working with large search windows and the advantage of applying a soft constraint favoring close locations over a hard cut-off. Furthermore, we found that the inclusion of image features yields a clear improvement in segmentation accuracy.

References

1. Arbelaez, P., Maire, M., Fowlkes, C., Malik, J.: Contour detection and hierarchical image segmentation. IEEE Trans. on Pat. Anal. Mach. Intel. 33(5), 898–916 (2011)
2. Buades, A., Coll, B., Morel, J.M.: A review of image denoising algorithms, with a new one. Multiscale Modeling & Simulation 4(2), 490–530 (2005)
3. Chen, A., Noble, J.H., Niermann, K.J., Deeley, M.A., Dawant, B.M.: Segmentation of parotid glands in head and neck CT images using a constrained active shape model with landmark uncertainty. In: SPIE. vol. 8314, p. 83140P (2012)
4. Coupé, P., Manjón, J.V., Fonov, V., Pruessner, J., Robles, M., Collins, D.L.: Patch-based segmentation using expert priors: Application to hippocampus and ventricle segmentation. NeuroImage 54(2), 940 – 954 (2011)
5. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on. vol. 1, pp. 886–893. IEEE (2005)
6. Fritscher, K.D., Peroni, M., Zaffino, P., Spadea, M.F., Schubert, R., Sharp, G.: Automatic segmentation of head and neck ct images for radiotherapy treatment planning using multiple atlases, statistical appearance models, and geodesic active contours. Medical physics 41(5), 051910 (2014)
7. Han, X., Hibbard, L.S., O'connell, N.P., Willcut, V.: Automatic segmentation of parotids in head and neck CT images using multi-atlas fusion. In: Medical Image Analysis for the Clinic: A Grand Challenge. pp. 297–304 (2010)
8. Jain, A.K., Farrokhnia, F.: Unsupervised texture segmentation using gabor filters. Pattern recognition 24(12), 1167–1186 (1991)
9. Liao, S., Gao, Y., Lian, J., Shen, D.: Sparse patch-based label propagation for accurate prostate localization in ct images. Medical Imaging, IEEE Transactions on 32(2), 419–434 (2013)
10. Liao, S., Gao, Y., Shi, Y., Yousuf, A., Karademir, I., Oto, A., Shen, D.: Automatic prostate mr image segmentation with sparse label propagation and domain-specific manifold regularization. In: IPMI. pp. 511–523 (2013)
11. Muja, M., Lowe, D.G.: Scalable nearest neighbor algorithms for high dimensional data. Pattern Analysis and Machine Intelligence, IEEE Transactions on 36 (2014)
12. Qazi, A.A., Pekar, V., Kim, J., Xie, J., Breen, S.L., Jaffray, D.A.: Auto-segmentation of normal and target structures in head and neck CT images: A feature-driven model-based approach. Medical physics 38, 6160 (2011)
13. Ramus, L., Malandain, G.: Multi-atlas based segmentation: Application to the head and neck region for radiotherapy planning. In: Medical Image Analysis for the Clinic: A Grand Challenge. pp. 281–288 (2010)
14. Rousseau, F., Habas, P.A., Studholme, C.: A supervised patch-based approach for human brain labeling. IEEE Trans. Med. Imaging 30(10), 1852–1862 (2011)
15. Wachinger, C., Navab, N.: Entropy and laplacian images: Structural representations for multi-modal registration. Medical Image Analysis 16(1), 1 – 17 (2012)
16. Wachinger, C., Sharp, G., Golland, P.: Contour-driven regression for label inference in atlas-based segmentation. In: MICCAI 2013. LNCS, Springer, Heidelberg (2013)