# Stacked Multiscale Feature Learning for Domain Independent Medical Image Segmentation

Ryan Kiros, Karteek Popuri, Dana Cobzas, and Martin Jagersand

University of Alberta, Canada

**Abstract.** In this work we propose a feature-based segmentation approach that is domain independent. While most existing approaches are based on application-specific hand-crafted features, we propose a framework for learning features from data itself at multiple scales and depth. Our features can be easily integrated into classifiers or energy-based segmentation algorithms. We test the performance of our proposed method on two MICCAI grand challenges, obtaining the top score on VESSEL12 and competitive performance on BRATS2012.

## 1 Introduction

The choice of image representation plays a crucial role in the success of medical image segmentation algorithms. Most existing methods utilize hand-crafted features incorporated into an energy-based segmentation method or into a machine learning classifier. Commonly, energy-based methods utilize engineered features such as Gabor filters for texture-based segmentation [1], while machine learning approaches use many more simple features like Haar or steerable filters leaving the classification method to disambiguate the ones that are significant for the segmentation task. Popular examples of machine learning methods are ones based on decision trees [2] or random forests [3]. Some methods use very specialized filters designed for a particular task, such as extracting linear structures based on eigenvalues of the image Hessian matrix [4].

Recently there has been much interest within the machine learning and computer vision communities to automatically learn feature representations from scratch. Feature learning methods are general, while hand-crafted features require a certain insight and understanding of the given image data to be analyzed, thus they are often not optimal when applied to a new dataset. Moreover, feature learning algorithms can benefit from many unlabeled examples, even those that may come from a different distribution than the target data [5]. Features can be learned either in an unsupervised setting or in a joint end-to-end system trained with supervision. Successful applications have included object recognition [6] [7], scene parsing and segmentation [8], annotation and retrieval [9], multimodal applications [10] and large-scale learning [11]. What these methods have in common is the emphasis on learning hierarchical representations as opposed to single-layer algorithms such as sparse coding.

Unfortunately, most of the above methods are not directly applicable to medical imaging tasks as they often assume the use of natural images and require a

Table 1: A comparison of different feature learning architectures for application to medical image segmentation: Y is yes, N is no and S is sometimes. Multi-scale and multi-depth methods can often improve performance while patch-based and stagewise learning improve speed. Here sparse coding eefers to any method that aims to learn a filter bank with a sparsity cost.

| **Method** | Patch-based | Multi-scale | Multi-depth | Stagewise |
|---|---|---|---|---|
| Sparse coding | Y | N | N | Y |
| Convolutional sparse coding | N | N | N | Y |
| Convolutional networks | N | S | Y | N |
| Proposed approach | Y | Y | Y | Y |

large number of labeled examples to be effective. There exist few feature learning methods applied to medical data like the segmentation of linear [12] and curvilinear [13] structures; segmentation of electron microscopy (EM) images [14] and a recent work on MS lesions segmentation [15]. In this paper we propose a framework that is domain independent and utilizes features learned from multiple scales and depth. The key features that make our method fast and thus suitable for medical data are detailed below and can be summarized as: (1) patch-based, (2) stage-based system and a (3) fast dictionary learning method.

Table 1 summarizes and distinguishes four types of feature learning architectures. Simpler single layer sparse coding methods like [15] also use patches but with no scales or depth. Convolutional sparse coding algorithms, such as those used by [12],[16] and [13], differ from standard sparse coding methods as convolution is incorporated into the optimization procedure. The third architecture describes convolutional networks, used by [14] for EM segmentation, which are learned jointly with supervision. While convolutional networks are often very effective, jointly training the whole model can be time consuming. Furthermore, convolutional networks require many labeled examples in order to avoid overfitting. The last architectures illustrates our proposed framework. Features are learned one stage at a time using patch-based learning at multiple scales. Since the model does not require joint learning, features can be learned efficiently and quickly. Our framework is the first to utilize the "encoding versus training" principle of [17] in the context of image segmentation. The emphasis of this work is the importance of the feature encoding as opposed to the filter learning algorithm itself. Due to this, we suggest that more expensive convolutional filter learning is unnecessary, so long as a proper encoding is performed after learning.

Experimentally we demonstrate that the same algorithm can be used to obtain strong performance on two completely different medical segmentation tasks. We report superior results on the vessel segmentation of the lung (VESSEL12) challenge data and competitive performance on multimodal brain tumor segmentation (BRATS2012) data. Furthermore, our system is able to learn features in under ten minutes on both challenges. Code for our approach will be released upon publication.

## 2 Method

We assume we are given $m$ volumes with $s$ modalities $\{\{V^{(j)}\}_{j=1}^{s}\}_{i=1}^{m}$, where each $V_i^{(j)} \in \mathbb{R}^{n_V \times n_H \times n}$. $n_V \times n_H$ is the spatial dimension of a slice and $n$ is the number of slices. For simplicity, we assume that each volume $V_i^{(j)}$ has dimensionality $n_V \times n_H \times n$ although this is not needed. As a specific example, brain tumor segmentation tasks can use $s = 4$ modalites consisting of FLAIR, T1, T2 and post-Gadolinium T1. The general outline of our feature learning framework is as follows:

- Extract multimodal patches at multiple scales using a Gaussian pyramid.
- Learn a filter bank using orthogonal matching pursuit.
- Convolutionally extract feature maps using the learned filters as kernels.
- Repeat the above steps, using the computed features maps as input to another layer. The number of feature maps (next layer modalities) corresponds to the number of filters.

In each of the following subsections, we describe the above operations in detail.

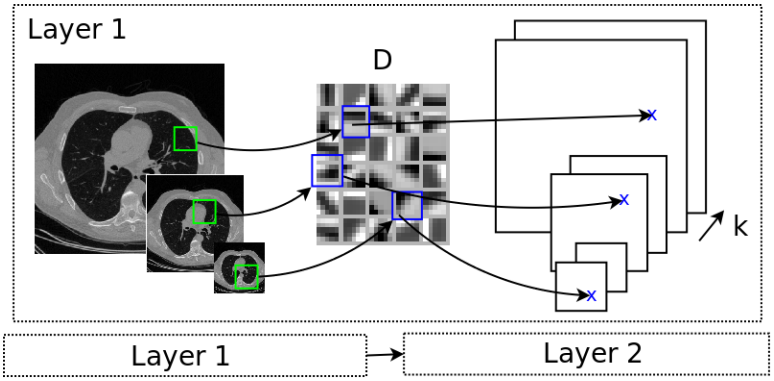


Fig. 1: Visualization of our feature learning approach. Each volume slice is scaled using a Gaussian pyramid. Patches are extracted at each scale to learn a dictionary $D$ using OMP. Convolution is performed over all scales with the dictionary filters, resulting in $\Gamma k$ feature maps. After training the first layer, the feature maps can then be used as input to a second layer.

### 2.1 Pre-processing and dictionary learning

Given a volume $V$, a Gaussian pyramid with $\Gamma$ scales is applied to each modality of each slice. Let $\{p^{(1)}, \ldots, p^{(m_P)}\}$ denote a set of $m_P$ patches randomly extracted from the scaled volumes. Each patch $p^{(l)}$ is of spatial dimension $r \times c \times s$ where $r \times c$ is the receptive field size. These patches are then flattened into column

vectors. Per patch contrast normalization and patch-wise mean subtraction is performed. For dictionary learning we use orthogonal matching pursuit (OMP). OMP aims to solve the following optimization problem:

$$\begin{aligned} \underset{D,x^{(i)}}{\text{minimize}} \quad & \sum_{i=1}^{m_P} ||Dx^{(i)} - p^{(i)}||_2^2 \\ \text{subject to} \quad & ||D^{(l)}||_2^2 = 1, \forall l \\ & ||x^{(i)}||_0 \leq q, \forall i \end{aligned} \tag{1}$$

where $D \in \mathbb{R}^{n_P \times k}$ and $D^{(l)}$ is the $l$-th column of $D$. Optimization is done using alternation over the dictionary $D$ and codes $x$. For all our experiments we set $q = 1$, which reduces to a form of gain-shape vector quantization. In particular, given a dictionary $D$, an index $\kappa$ is chosen as

$$\kappa = \underset{l}{\text{argmax}} |{D^{(l)}}^T p^{(i)}| \tag{2}$$

for which the $\kappa$-th index of $x^{(i)}$ is set as $x_\kappa^{(i)} = {D^{(\kappa)}}^T p^{(i)}$ with all other indices left as zero in order to satisfy the constraint $\|x^{(i)}\|_0 \leq 1$ for all $i$. Given the one-hot codes $X$, the dictionary is easily updated by first solving the unconstrained problem, followed by re-normalization to satisfy the constraint $\|D^{(l)}\|_2^2 = 1$ for all $l$.

### 2.2 Convolutional feature extraction

Let $T_j^\gamma$ denote a volume slice of modality $j$ and scale $\gamma$. Each $r \times c \times s$ patch in $T_j^\gamma$ is pre-processed by contrast normalization and mean subtraction. Let $D_j^{(l)} \in \mathbb{R}^{r \times c}$ denote the $l$-th basis for modality $j$ of $D$. We will define the feature encoding for basis $l$ as:

$$f_l^\gamma = \sum_{j=1}^{s} T_j^\gamma * D_j^{(l)} \tag{3}$$

where * denotes convolution. The resulting feature maps $\{f_l^\gamma\}_{l=1}^k$ are of the same spatial dimensions as $T_j^\gamma$. The feature maps are finally upsampled to the original $n_V \times n_H$ spatial dimension. Figure 1 illustrates our approach.

### 2.3 Stacking multiple layers

Our described setup for feature learning has involved scaling, dictionary learning and convolutional extraction. Just as the volumes slices were inputs to a first layer with $s$ modalities, the upsampled output feature maps $\{\{f_l^\gamma\}_{\gamma=1}^\Gamma\}_{l=1}^k$ may be seen as inputs to a second layer but with $\Gamma k$ modalities. The same described operations are applied a second time resulting in additional second layer output feature maps. These groups of feature maps can be concatenated together resulting in a total number of $\Gamma_1 k_1 + \Gamma_2 k_2$ feature maps, where $\Gamma_1, k_1$ are the number

of first layer scales and filters while $\Gamma_2, k_2$ are the number of second layer scales and filters. Thus each pixel in a volume slice can be represented as a $\Gamma_1 k_1 + \Gamma_2 k_2$ dimensional feature vector.

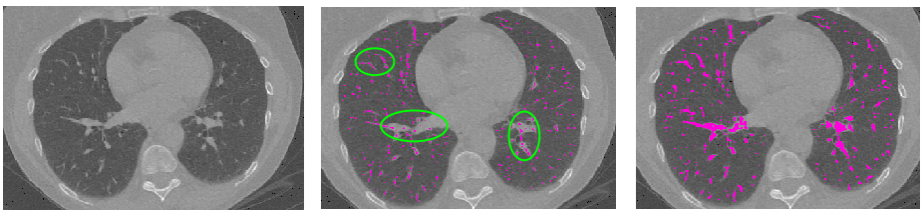

Fig. 2: Visualizing the importance of scale and depth for vessel segmentation.

## 3 Experiments

We perform experimental evaluation using data from two MICCAI grad challenges: vessel segmentation of the lung [1] and multimodal brain tumor segmentation [2].

### 3.1 Vessel segmentation

The vessel segmentation challenge consists of 20 volumes of CT scans to segment with 3 additional volumes that include 882 labeled pixels based on the agreement of at least 3 experts. Each slice is of size $512 \times 512$ with each volume containing a few hundred slices. We performed feature learning with 2 depths, 6 scales, a receptive field size of $5 \times 5$, 32 first layer filters and 64 second layer filters. The final feature vector is thus of size $6 \times (32+64) = 576$. In order to perform segmentation, we extracted features for the existing labeled pixels and trained a L2-regularized logistic regression classifier, using 10-fold cross validation in order to tune the L2 hyperparameter. Each pixel of a new slice is then classified, resulting in a probability of whether or not the pixel is a vessel. For our submission to the challenge, the probabilities are scaled and rounded to unsigned 8-bit integers as requested.

Figure 2 illustrates the importance of adding depth and scale to segmentation. The first image is the original CT scan. The second image shows segmentation when neither depth nor scale is added while the third image shows segmentation with added depth and scale. Without scale, larger vessels are less likely to be segmented while without depth, segmentation is much more scattered and less contiguous. For visualization purposes, a pixel is labeled as being a vessel if the probability of a vessel given the pixel features is greater than 0.5.

[1] `http://vessel12.grand-challenge.org/`
[2] `http://www2.imm.dtu.dk/projects/BRATS2012/`

Table 2: The top 5 results from the VESSEL12 challenge leaderboard.

| **Team** | Method type | score |
|---|---|---|
| our method | feature learning + classification | **0.986** |
| LKEBChina | Krissian-inspired vesselness | 0.984 |
| FME_LungVessels | Frangi vesselness + region growing | 0.984 |
| LKEBChina | Krissian-inspired vesselness with bi-Gaussian kernel | 0.981 |
| FME_LungVessels | Frangi vesselness + region growing (raw) | 0.981 |

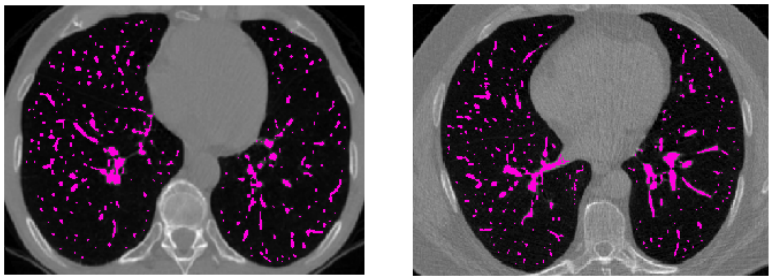

Fig. 3: Sample vessel segmentation results.

Table 2 shows the top 5 performing methods on the VESSEL12 challenge. Our proposed method tops all existing approaches. The top performing methods in the competition are largely based on the use of Frangi [4] and Krissian vesselness [18] all of which derive structural properties from the eigenvalues of the Hessian.

### 3.2 Brain tumor segmentation

To emphasize that the proposed method is domain independent, we evaluated it on the BRATS2012 multimodal brain tumor segmentation challenge, a dataset that has totally different properties and segmentation task than the vessel data. Due to BRATS2012 site maintenance, the test volume labels were unavailable at the time we did our experiments. Instead we perform evaluation using leave-one-out cross validation on the training set. Two types of tumour data are evaluated: high-grade and low-grade. Each volume voxel is labeled as being one of three classes: tumor, edema and other. We utilized our approach with one scale and two depths, with 16 bases in each depth for a total of 32 features. A 2 hidden layer network with dropout [19] is used to make predictions. Within each training fold, 10-fold cross validation is used to select the dropout parameters.

Table 3 shows our results in comparison to the top 2 methods in the competition. We note again that out comparison is not on the same held-out data. None the less, our results are competitive with the top performing methods.

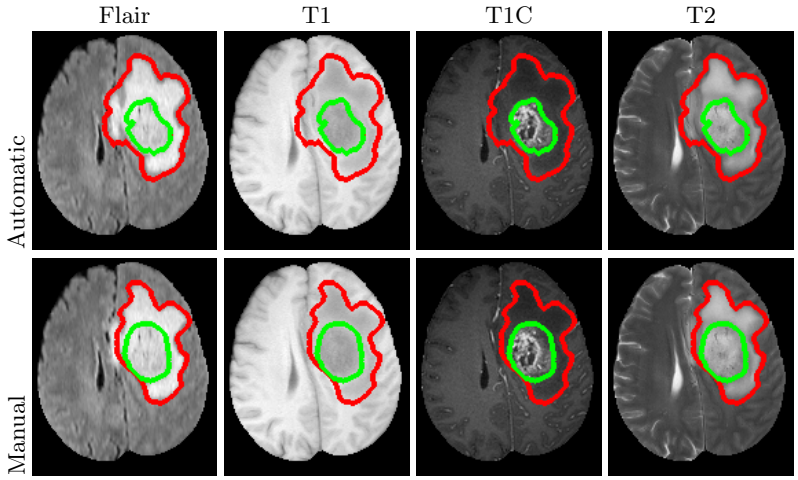


Fig. 4: Sample brain tumor segmentation results.

Table 3: Comparison against the top two performers in the BRATS2012 competition. HG and LG stand for high-grade and low-grade, respectively.

| **Team** | region | mean dice coeff. | region | mean dice coeff. |
|---|---|---|---|---|
| our method | HG edema | 0.485 | LG edema | 0.250 |
| Bauer et al. | HG edema | 0.536 | LG edema | 0.179 |
| Zikic et al. | HG edema | **0.598** | LG edema | **0.324** |
| our method | HG tumor | 0.470 | LG tumor | **0.406** |
| Bauer et al. | HG tumor | **0.512** | LG tumor | 0.332 |
| Zikic et al. | HG tumor | 0.476 | LG tumor | 0.339 |
| our method | HG GTV | 0.720 | LG GTV | 0.494 |

## 4 Conclusion

In this paper we proposed a domain independent approach for segmenting medical images. Our approaches involves learning feature representations at multiple scales and depths which are compatible with existing classification and energy-based segmentation methods. We obtain the best performing result on the VESSEL12 challenge and competitive results on the BRATS2012 multimodal brain tumor segmentation challenge. For future work we intend to further evaluate our approach on additional grand challenge problems. We also intend to study various transfer learning scenarios between domains and modalities.

## References

1. Paragios, N., Deriche, R.: Geodesic active regions and level set methods for supervised texture segmentation. IJCV **50**(3) (2002) 223–247

2. 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 Trans. Medical Imaging **27**(11) (2008) 1668–1681
3. Criminisi, A., Robertson, D., Konukoglu, E., Shotton, J., Pathak, S., White, S., Siddiqui, K.: Regression forests for efficient anatomy detection and localization in computed tomography scans. Medical Image Analysis (2013)
4. Frangi, A., Niessen, W.J., Vincken, K.L., Viergever, M.A.: Multiscale vessel enhancement filtering. In: MICCAI. (1998) 130–137
5. Raina, R., Battle, A., Lee, H., Packer, B., Ng, A.Y.: Self-taught learning: transfer learning from unlabeled data. In: ICML. (2007) 759–766
6. Bo, L., Ren, X., Fox, D.: Unsupervised Feature Learning for RGB-D Based Object Recognition. In: ISER. (June 2012)
7. Krizhevsky, A., Sutskever, I., Hinton, G.: Imagenet classification with deep convolutional neural networks. In: NIPS. (2012) 1106–1114
8. Farabet, C., Couprie, C., Najman, L., LeCun, Y.: Scene parsing with multiscale feature learning, purity trees, and optimal covers. In: ICML. (2012)
9. Kiros, R., Szepesvari, C.: Deep representations and codes for image auto-annotation. In: NIPS. (2012) 917–925
10. Srivastava, N., Salakhutdinov, R.: Multimodal learning with deep boltzmann machines. In: NIPS. (2012) 2231–2239
11. Dean, J., Corrado, G., Monga, R., Chen, K., Devin, M., Le, Q., Mao, M., Ranzato, M., Senior, A., Tucker, P., Yang, K., Ng, A.: Large scale distributed deep networks. In: NIPS. (2012) 1232–1240
12. Rigamonti, R., Lepetit, V.: Accurate and efficient linear structure. segmentation by leveraging ad hoc features with learned filters. In: MICCAI. (2012) 189–197
13. Becker, C., Rigamonti, R., Lepetit, V., Fua, P.: Supervised feature learning for curvilinear structure segmentation. In: MICCAI. (2013) 526–533
14. Ciresan, D., Giusti, A., Schmidhuber, J., et al.: Deep neural networks segment neuronal membranes in electron microscopy images. In: NIPS. (2012) 2852–2860
15. Weiss, N., Rueckert, D., Rao, A.: Multiple sclerosis lesion segmentation using dictionary learning and sparse coding. In: MICCAI. (2013) 735–742
16. Rigamonti, R., Türetken, E., González, G., Fua, P., Lepetit, V.: Filter learning for linear structure segmentation. Technical report, Technical report, EPFL (2011)
17. Coates, A., Ng, A.Y.: The importance of encoding versus training with sparse coding and vector quantization. In: ICML. Volume 8. (2011) 10
18. Krissian, K., Malandain, G., Ayache, N., Vaillant, R., Trousset, Y.: Model-based detection of tubular structures in 3d images. Computer vision and image understanding **80**(2) (2000) 130–171
19. Hinton, G.E., Srivastava, N., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.R.: Improving neural networks by preventing co-adaptation of feature detectors. arXiv preprint arXiv:1207.0580 (2012)