Learning a Structured Graphical Model with Boosted Top-Down Features for Ultrasound Image Segmentation

Zhihui Hao[1], Qiang Wang[1], Xiaotao Wang[1], Jung Bae Kim[2], Youngkyoo Hwang[2], Baek Hwan Cho[3], Ping Guo[1], and Won Ki Lee[1]

[1] Medical Imaging Group, China Lab
[2] Medical System Lab
[3] Data Analytics Group, Samsung Advanced Institute of Technology

Abstract. A key problem for many medical image segmentation tasks is the combination of different-level knowledge. We propose a novel scheme of embedding detected regions into a superpixel based graphical model, by which we achieve a full leverage on various image cues for ultrasound lesion segmentation. Region features are mapped into a higher-dimensional space via a boosted model to become well controlled. Parameters for regions, superpixels and a new affinity term are learned simultaneously within the framework of structured learning. Experiments on a breast ultrasound image data set confirm the effectiveness of the proposed approach as well as our two novel modules.

1 Introduction

Pathological structure segmentation is one of the core tasks for medical image processing. In recent years, studies on image segmentation are often categorized as one of the two paradigms: top-down or bottom-up ones [1]. The former involves a detection of object bounding-box and a follow-up refining of object contour. A high-precision detector becomes the key to rank the huge amount of sliding windows. The latter is committed to a classification of all image atoms, *i.e.* pixels or superpixels, and a holistic grouping based on their affinities.

The two paradigms suffer their own troubles, however, especially when dealing with medical images. Take breast sonograms for example, the various shapes of lesion make them hardly be handled by box-fashioned detector and the insufficiency of uniform structural features increases the difficulty of detection model training. On the other side, classifying and grouping image atoms often fails without a wide range of inspection, since many pixels in lesion and subcutaneous fat lobules barely have any difference. As discussed in [2], ultrasound image segmentation relies on an organic integration of low- and high-level image knowledge to remedy these problems.

In fact, many efforts have been made towards a combination of top-down and bottom-up segmentation [1,3,4]. We will compare them with ours in Sec. 2. The main contribution of this work is a novel scheme of embedding detection windows into a superpixel based graphical model, which enables an automatic

K. Mori et al. (Eds.): MICCAI 2013, Part I, LNCS 8149, pp. 227–234, 2013.

process of lesion detection and segmentation. Properties of these windows are not simply used for a ranking [5] or heuristically treated as superpixel features [2]. Their impacts to superpixels are well controlled independently after boosted into higher-dimensional space. Furthermore, all parameters for detection window properties, superpixel features and a new superpixel affinity term are trained in the framework of structured learning [6], thereby ensuring a full leverage of them. The simplicity and generality of these features allows a potentially fast execution of our approach, as well as an immediate application of the model to other segmentation problems.

2 Related Works

Conditional random field (CRF) models have been used for solving segmentation problems in both medical images [7,2] and natural images [8,9,3,10]. Structured support vector machine (SSVM) is firstly introduced to CRF by Szummer *et al.* [8] to learn the weights that balance all data terms. As in [9,7], the unary term in [8] is defined based on the posterior probability of graph node. Lucchi *et al.* [10] moves a step forward by co-training the interior coefficients with a linearization of the energy function. Our work belongs to this strand. But different from both, our work is committed to integrating region information into CRF. We put our attention on the design of region model and a corresponding pairwise term.

Many studies have made efforts to combine top-down and bottom-up image cues for object segmentation. Levin and Weiss [1] define a location bias term in CRF, which calculates the cost of aligning pixel-wise segmentation mask with object-part regions. The regions are selected from a large pool and used directly without a discussion of their validities. Ladickỳ *et al.* [3] introduce detection windows into CRF in the form of a higher order potential. The windows could be accepted or rejected based on the harmony with other-level potentials. The work most related to ours is [2], where the authors treat a set of features extracted from detection hypotheses as auxiliary features of superpixel. Our work differs from them in two main aspects: *first*, image regions and atoms have independent features and they interact with, instead of filtering each other; *second*, parameters in the model are freed from individual assignments, but are packaged up and optimized within the structured learning framework.

The necessity of learning a specific pairwise term has also been observed by [11,7]. The novelty of our work is, instead of training an affinity function, we optimize the parametric pairwise term together with the unary term, thus obtain a model with potentially better compatibility.

3 The Problem and Energy Function

Our work is based on the bottom-up strategy. The task of lesion segmentation is treated as that of labeling variables in a conditional random field. Let $X = \{\mathbf{x}_i\}_i$ be the set of random variables that correspond to image atoms (superpixels here) and $Y = \{y_i\}_i$ be one of many possible labelings. The CRF model finds the

optimal Y^* by minimizing an energy function, which consists of a unary term $D(y_i, \mathbf{x}_i; \mathbf{w}^A)$ for all node variables and a pairwise term $V(y_i, y_j, \mathbf{x}_i, \mathbf{x}_j; \mathbf{w}^V)$ for all neighboring pairs (i, j). Let us abbreviate the two terms by $D_i^A(\mathbf{w}^A)$ and $V_{ij}(\mathbf{w}^V)$. In this paper, we will learn a graphical model with the energy function of

$$\mathcal{E}(Y, X, R; \mathbf{w}) = \sum_i \left(D_i^A(\mathbf{w}^A) + D_i^R(y_i; \mathbf{w}^R, \mathbf{x}^R) \right) + \sum_{(i,j)} V_{ij}(\mathbf{w}^V). \quad (1)$$

This equation reveals the problem to be addressed. The energy function contains a new unary term (with a superscript of R), which is introduced by image regions in contrast to the one by image atoms (with superscript A). The regions are generated from some top-down object locating paradigm, *e.g.* from sliding-window detector [12] in this paper. We preserve about 20 detection hypotheses for each image. The impact of this new term D^R is controlled by an unknown parameter $\mathbf{w}^R$ as well as the state of the regions, $\mathbf{x}^R$.

The main goal of this work is to efficiently define the data terms and simultaneously learn their parameters, thereby building a graphical model with structured outputs for image atoms and regions. We simply call it a structured graphical model (SGM) in the paper. With such a model, we can combine different-level image cues and leverage them effectively for the final cut of CRF.

The image regions acceptable to the energy function in Eq. (1) are not limited to detection windows. In medical images with strong ambiguity, different-scale fragments can provide different perspectives on the object of interest, for example, superpixels and maximally stable extremal regions (MSER) [13]. The proposed model allows an alternation of the region finding method and since the energy function is additive, it also allows the overlapping of multiple sets of regions. These regions are possibly from several different methods, or even from different imaging modalities.

4 Learning a Structured Graphic Model

4.1 The Unary Terms for Image Atoms and Regions

The unary term of D_i^A measures the cost of assigning label y_i to the feature vector $\mathbf{x}_i$. As in [10], we define it as a sum of inner products of $\mathbf{x}_i$ with class-dependent parameters:

$$D_i^A(\mathbf{w}^A) = \delta(y_i = 1) D_i^{A+} + \delta(y_i = -1) D_i^{A-} = \delta^+ \langle \mathbf{w}^{A+}, \mathbf{x}_i \rangle + \delta^- \langle \mathbf{w}^{A-}, \mathbf{x}_i \rangle, \quad (2)$$

where $\mathbf{w}^A = [(\mathbf{w}^{A+})^T, (\mathbf{w}^{A-})^T]^T$. Although expressed with binary symbols for clarity, it can be extended to multi-class cases without much effort. Next we use similar notations to define the region term D_i^R.

Suppose there is a detection window R that covers superpixel i with a confidence of x^R. We define the data term from R to i as a non-linear step function: $D_i^{R+}(w^R, x^R) = w^{R+} \delta(x^R \gtrless \tau)$, where $\gtrless$ takes one from *greater-than* and *less-than* signs. By adjusting the gain w^{R+} and the threshold τ, we can easily control

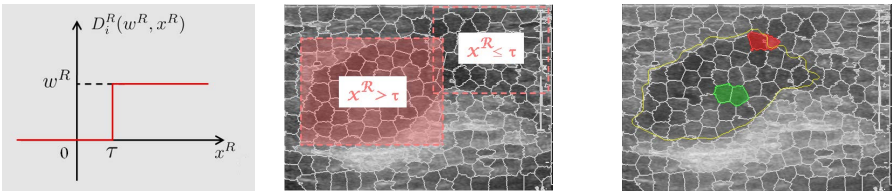

Fig. 1. The *left* and *middle* figures illustrate the simplified region model and an application example, where the detection windows will release energies of w^R and 0 respectively to superpixels they cover. The *right* figure shows a positive (red) and a negative (green) pair of superpixels for training the pairwise term.

and discriminate the impacts of different region instances that come from a same paradigm. The D_i^{R-} is defined in the same way, with the parameter of w^{R-}. The region model is shown in Fig. 1.

Moreover, the regions could have multiple properties. For example, the *objectness* measures in [5] describe the saliency of the region, the color contrast with the surrounding, and *etc.* We can then derive a more complex version of the region term:

$$D_i^{R+}(\mathbf{w}^{R+}, \mathbf{x}^R) = \sum_{(k)} w_{(k)}^{R+} \delta(x_{i(k)}^R \gtrless \tau_{(k)}) \triangleq \langle \mathbf{w}^{R+}, \phi(\mathbf{x}_i^R) \rangle. \tag{3}$$

Note that the region states could behave very different when projected on different image atoms, for example, on those they cover and miss, thereby represented by $\mathbf{x}_i^R$ in practical calculating. As indicated by Eq. (3), the set of the step functions can be regarded as a mapping function, which transforms the state $\mathbf{x}^R$ into a potentially higher dimensional feature vector $\phi(\mathbf{x}^R)$. Therefore, the whole term can be expressed as an inner product as in Eq. (2). The mapping function can be trained with $(\mathbf{x}_i^R, y_i)_i$ by off-the-shelf boosting learning algorithm [14], which returns an ensemble model composed of organized decision stumps and their associated weights. We preserve the first K decision stumps as our step functions. The weight parameters $\mathbf{w}^R$ will be re-trained as a part of our structured model.

From another perspective, the unary term equals to a weighted sum of image atom features and boosted region features, rather than with raw region features. This is the foundation that our method can outperform previous ones. We will show the related comparisons in the experiment section.

4.2 The Parametric Pairwise Term

The pairwise term measures the affinity of any superpixel pair in the graph. A general-purpose pairwise term in CRF involves a product of several image factors (such as the color similarity between superpixels and their shared boundary length [9]) with a single, scalar parameter. In our case, however, the intervention of region hypotheses creates many artificial edges in the probability map along the region boundaries, which brings much trouble for tuning the pairwise term.

Table 1. Features extracted for the data terms in our structured graphical model

Unary Terms		Parametric Pairwise Term
From Image Atoms (Superpixels)	*Intensity histogram*; *Centroid Location (x,y).*	*Intensity contrast*; *Length of shared boundary (LSB)*; *Edge strength*; *Edge strength averaged by LSB*; *Localized version*† *of above features.*
From Regions* (Detection Windows [12])	*Detection confidence*; *Objectness measures* [5]; *Localized version*† *of above*;	

*For any atom covered by multiple regions, take their maximum features as the atom features.
†A feature in image I is localized by $\tilde{x} = (x - \mu)/\sigma$, where the mean μ and standard deviation σ are estimated from the feature instances in I.

We need a new definition with higher degrees of freedom to accommodate our new unary term.

Again, we enforce the non-linear mapping function in Eq. (3) to boost the raw pair-features. Our pairwise term is defined as follows:

$$V_{ij}(\mathbf{w}^V) = \delta(y_i = y_j) f(\mathbf{x}^{ij}, \mathbf{w}^V) = \delta(y_i = y_j) \sum_k w_k^V \delta(x_k^{ij} \gtrless \tau_k). \tag{4}$$

Training samples are collected around the groundtruth of lesion to learn the mapping function. As shown in Fig. 1(c), pairs that cross the lesion boundary are regarded as positive samples with edge labels of 1 and pairs inside the lesion are negatives. The other cases are ignored.

The feature $\mathbf{x}^{ij}$ characterizing a neighboring atom pair is extracted from different aspects of image cues. We summarize the feature types in Tab. 1. The edge strength in the table is an accumulation of Canny edge within a narrow band along the shared boundary. Note that we also introduce a set of localized version of these features by calculating their standard scores. The localized features can focus the mapping function on local contrast especially during training to highlight ambiguous lesion boundaries. Besides those in Tab. 1, we believe that a concatenation of the individual features $\mathbf{x}_i$ and $\mathbf{x}_j$ [7] could be a valid supplement.

4.3 Learning the Parameters

The parameter $\mathbf{w}$ consists of $\mathbf{w}^A$, $\mathbf{w}^R$ and $\mathbf{w}^V$. After simple transformations, we can see that the energy in Eq. (1) becomes linearly expressible in $\mathbf{w}$. In order to recover the segmentation by minimizing this $\mathcal{E}_\mathbf{w}$, the following inequality should hold for any image X with a groundtruth labeling G.

$$\mathcal{E}(Y) - \mathcal{E}(G) = \mathbf{w}^T \Psi(Y) - \mathbf{w}^T \Psi(G) \geq 0, \ \forall Y \neq G. \tag{5}$$

The optimal $\mathbf{w}$ can be learned in the framework of structured support vector machine [6]. Given a set of training images and their groundtruth labelings, the SSVM optimizes the parameter by enlarging the margins between $\mathcal{E}(G)$ and any other $\mathcal{E}(Y)$:

$$\min_{\boldsymbol{\xi},\mathbf{w}} \sum_{n=1}^{N} \xi^{(n)},\ \text{s.t.}\ \ \boldsymbol{\xi} \succeq \mathbf{0},\ \mathbf{w} \succeq \mathbf{0},\ \|\mathbf{w}\|_1 = 1, \tag{6}$$

$$\mathbf{w}^T \Psi(Y) - \mathbf{w}^T \Psi(G^{(n)}) \geq \Delta(Y, G^{(n)}) - \xi^{(n)},\ \forall Y \neq G^{(n)},\ \forall n$$

We constrain the 1-norm of $\mathbf{w}$ to encourage its sparsity. The term of $\Delta(Y, G^{(n)})$ applies re-scaled margins [6] to different cases of Y. Instead of using the hamming loss, it has a slightly different definition in our work:

$$\Delta(Y,\ G^{(n)}) = \sum_{n} \delta(y_i = 0,\ g_i^{(n)} = 1), \tag{7}$$

which means that we only punish the cases where lesion superpixels are falsely classified into backgrounds. This is very important to our problem considering in most cases breast lesion possesses only a small percentage of area. The biased penalty prevents them from being overshadowed by redundant negatives in the model learning phase.

The number of possible constraints could be nearly infinite. The SSVM solves this problem by employing the column-generation technique. At each iteration, only the most violated constraint for each training sample is added into the working set. The corresponding labeling can be found using the standard graph cut method [15] as in inference. We refer the reader to [6,10] for similar details. We stop the learning when $\mathbf{w}$ converges. With a tolerance of $1e-5$ on $\|\mathbf{w}^t - \mathbf{w}^{t-1}\|_2$, the learning process usually takes about 30 iterations in our experiments.

5 Experiments

We evaluate the proposed approach on a 2D breast ultrasound image data set, which contains 469 B-mode ultrasound images with about 52% benign cases and 48% malignant cases. The image size is about 780×540 with a spatial resolution of 0.23mm/pixel. The malignant cases have been confirmed with biopsy, and the benign cases have been followed up for at least 3 years. Lesion boundaries are delineated manually by experienced radiologists. We have provided them a special assisting software for gently refining the contours.

The dimension of boosted features in our graphical model, *i.e.* the number of step functions, is set to 100 for both the unary term and the pairwise term. About 200 SLIC superpixels [16] are generated in each image (also, for competitors in the following experiments). We randomly select 75% images for model training and the other 25% images for testing. Fig. 2 has shown some results of lesion segmentation. In the 4-th column, detection windows with $D^{R+} > D^{R-}$ are drawn. Note that our approach can handle the cases that contain multiple lesions, whereas detection windows with highest confidence (shown as valid rectangles) often miss one of the targets and lead to incomplete results.

Next, we compare the proposed approach with some baselines to highlight the contribution of each module. Considering the primary purpose of this work, we mainly focus on these three aspects: 1) Boosted Features (BF) vs. Raw Features (RF) provided by regions; 2) Parametric Pairwise-term (PP) vs. Ordinary

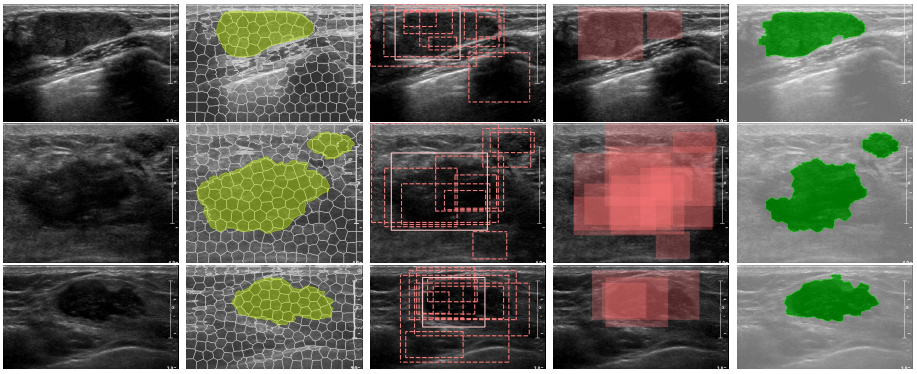

Fig. 2. Experiment results of lesion segmentation in breast ultrasound images. Each column shows 1) the input images, 2) superpixels with outlined groundtruth, 3) top-10 detection windows, 4) windows selected by our region model that have positive impacts to superpixels, and 5) the final segmentations.

Table 2. Experiment results of 5-time random subsampling cross-validations

Approaches	Average Jaccard	Hausdorff Distance	Average Distance
SGM(BF+PP)	0.693 ± 0.012	38.7 ± 3.1	15.6 ± 1.3
SGM(BF+OP)	0.681 ± 0.025	50.1 ± 3.8	17.8 ± 1.8
SGM(RF+OP)	0.669 ± 0.028	52.9 ± 5.4	20.8 ± 2.0
Hao12 [2]	0.654 ± 0.028	53.0 ± 4.8	25.4 ± 5.1
Fulkerson09 [9]	0.580 ± 0.052	69.5 ± 6.7	36.3 ± 7.2

Pairwise-term (OP); 3) the proposed vs. other segmentation approaches. Specifically, the following methods are compared:

- SGM(BF+PP), where all modules proposed in this paper are included;
- SGM(BF+OP), where the pairwise term is defined as in [9];
- SGM(RF+OP), where raw region features are treated as atom features;
- Hao12 [2], similar to SGM(RF+OP) but with more powerful features and without structured learning;
- Fulkerson09 [9], a typical CRF based segmentation approach without regions involved.

The comparison tests are repeated 5 times with a 3:1 random train/test split. A quantitative result is reported in Tab. 2. The measurements are the average overlapping ratio (*i.e.* the *Jaccard*), the maximum (*i.e.* the *hausdorff*) and the average contour-to-contour distances. Details of the latter two measurements can be found in [17]. The mean and standard derivation of 5 tests are reported. We can see that the proposed method with boosted features and parametric pairwise term outperforms the variants, and also obtains a better result than the previous methods.

6 Conclusion

We have proposed a structured graphical model to efficiently combine the region-level features and superpixel-level features. Relationships between regions and

superpixels can also be captured by defining new energy terms. Also, a structured labeling and segmentation for multi-class objects (if there are) could be studied based on our method. These are some directions of our future work.

References

1. Levin, A., Weiss, Y.: Learning to combine bottom-up and top-down segmentation. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006. LNCS, vol. 3954, pp. 581–594. Springer, Heidelberg (2006)
2. Hao, Z., Wang, Q., Seong, Y.K., Lee, J.-H., Ren, H., Kim, J.: Combining crf and multi-hypothesis detection for accurate lesion segmentation in breast sonograms. In: Ayache, N., Delingette, H., Golland, P., Mori, K. (eds.) MICCAI 2012, Part I. LNCS, vol. 7510, pp. 504–511. Springer, Heidelberg (2012)
3. Ladický, Ľ., Sturgess, P., Alahari, K., Russell, C., Torr, P.H.S.: What, where and how many? combining object detectors and crfs. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part IV. LNCS, vol. 6314, pp. 424–437. Springer, Heidelberg (2010)
4. Kuettel, D., Ferrari, V.: Figure-ground segmentation by transferring window masks. In: CVPR, pp. 558–565. IEEE (2012)
5. Alexe, B., Deselaers, T., Ferrari, V.: What is an object? In: CVPR, pp. 73–80. IEEE (2010)
6. Tsochantaridis, I., Hofmann, T., Joachims, T., Altun, Y.: Support vector machine learning for interdependent and structured output spaces. In: ICML, p. 104. ACM (2004)
7. Lucchi, A., Smith, K., Achanta, R., Lepetit, V., Fua, P.: A fully automated approach to segmentation of irregularly shaped cellular structures in em images. In: Jiang, T., Navab, N., Pluim, J.P.W., Viergever, M.A. (eds.) MICCAI 2010, Part II. LNCS, vol. 6362, pp. 463–471. Springer, Heidelberg (2010)
8. Szummer, M., Kohli, P., Hoiem, D.: Learning crfs using graph cuts. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part II. LNCS, vol. 5303, pp. 582–595. Springer, Heidelberg (2008)
9. Fulkerson, B., Vedaldi, A., Soatto, S.: Class segmentation and object localization with superpixel neighborhoods. In: CVPR, pp. 670–677. IEEE (2009)
10. Lucchi, A., Li, Y., Smith, K., Fua, P.: Structured image segmentation using kernelized features. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part II. LNCS, vol. 7573, pp. 400–413. Springer, Heidelberg (2012)
11. Batra, D., Sukthankar, R., Chen, T.: Learning class-specific affinities for image labelling. In: CVPR, pp. 1–8. IEEE (2008)
12. Felzenszwalb, P., Girshick, R., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part-based models. PAMI 1627–1645 (2009)
13. Feng, X., Shen, X., Wang, Q., Kim, J., et al.: Learning based ensemble segmentation of anatomical structures in liver ultrasound image. In: SPIE Medical Imaging, Citeseer (2013)
14. Freund, Y., Schapire, R.: A decision-theoretic generalization of on-line learning and an application to boosting. J. Comput. Syst. Sci. 55(1), 119–139 (1997)
15. Boykov, Y., Veksler, O., Zabih, R.: Fast approximate energy minimization via graph cuts. PAMI 23(11), 1222–1239 (2001)
16. Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., Susstrunk, S.: Slic superpixels compared to state-of-the-art superpixel methods. PAMI 34(11), 2274–2282 (2012)
17. Madabhushi, A., Metaxas, D.: Combining low-, high-level and empirical domain knowledge for automated segmentation of ultrasonic breast lesions. IEEE Trans. Med. Imaging 22(2), 155–169 (2003)