Object Localization in Medical Images based on Graphical Model with Contrast and Interest-Region Terms

Yang Song[1], Weidong Cai[1], Heng Huang[2], Yue Wang[3], David Dagan Feng[1]
[1]BMIT Research Group, School of IT, University of Sydney, Australia
[2]Computer Science and Engineering, University of Texas at Arlington
[3]Bradley Department of Electrical and Computer Engineering,
Virginia Polytechnic Institute and State University
{ysong,tomc,feng}@it.usyd.edu.au, heng@uta.edu, yuewang@vt.edu

Abstract

In this paper, we propose a novel method for object localization, generally applicable to medical images in which the objects can be distinguished from the background mainly based on feature differences. We design a new CRF model with additional contrast and interest-region potentials, which encode the higher-order contextual information between regions, on the global and structural levels. We also propose a sparse-coding based classification approach for the interest-region detection with discriminative dictionaries, to serve as a second opinion for more accurate region labeling. We evaluate our object localization method on two medical imaging applications: lesion dissimilarity on thoracic PET-CT images, and cell segmentation on microscopic images. Our evaluations show higher performance when comparing to recently reported approaches.

1. Introduction

A wide variety of medical applications comprise object localization as an important step for discovering the anatomical or pathological information from images. We consider object localization as a generalization of both detection and segmentation, with both automatic identification of ROI, and a good approximation of the boundary.

We focus on medical imaging problems in which objects can be localized based on local-level features and feature differences between the objects and background. For example, in positron emission tomography – computed tomography (PET-CT) images, abnormalities typically show higher uptakes than normal tissues. In fluorescence microscopic images, the cell nuclei normally depict darker colors then the other cell structures and the background.

Local features are usually not sufficient for a good localization, because of large inter-subject variations causing same anatomical structures appearing quite differently across images. The problem is further complicated due to low feature differences between different tissue types and especially for the boundary areas between the objects and background. In addition, pathologies often lead to larger imaging variations, and an accurate object localization is thus more challenging.

For such imaging problems, while lots of work have been reported [18, 14, 3], they are mostly designed to be domain specific; and often rely on sophisticated feature sets, which can be computational-intensive and difficult to adapt to other imaging problems. Furthermore, because such features are usually designed based on domain knowledge and empirical studies, their effectiveness may be restricted to the limited scenarios available in the datasets.

Therefore, we propose an object localization method that can be generally applicable, requires simpler feature sets, and addresses low feature differences and large inter-subject variations. In summary, our main contributions are the following: (i) we enhance the discriminative capability of the basic conditional random field (CRF) with a contrast potential and interest-region potential, to encode the global contrast information and region-based feature similarities, for improving the boundary delineations; (ii) a sparse-coding classification method is proposed for interest-region detection, with improved discriminative power of the learned dictionaries; (iii) our design is kept general, and local feature sets are configurable according to the specific application; and (iv) we evaluate the proposed method with both lesion dissimilarity on thoracic PET-CT images and cell segmentation on microscopic images.

1.1. Related Work

We focus our review on CRF-based localization methods in both medical and general imaging domains. As an undirected graphical model, CRF is now one of the most suc-

cessful trends in object class image segmentation [5]. The basic and most commonly used formulation is to have local features represented as graph nodes and consistency constraints between neighboring regions as edge connections [13]. However, comparing to the non-graphical discriminative approach, generally such models add advantages little more than spatial smoothing of labelings [18].

Higher-level features are often acknowledged as important discriminative factors [5, 3]. In particular, relationship information on a larger scale, such as those across image slices [7], relating to reference objects [2], or between distant image regions [6], can be modeled as pairwise connections to encourage labeling consistency or enhance the discriminative power of local features. Such ideas of connecting beyond immediate neighbors are inspiring; however, choosing the related pairs and describing their interactions are rather application specific. At a more structural level, object detectors with bounding box outputs have been incorporated into CRFs as consistency constraints [10, 5]. Although the idea is sound, such methods are normally built based on well-established object detectors and thus require only simple interaction modeling; but both assumptions are not suitable for our problem domain.

2. Object Localization

Given an image I, we first oversegmenting it into a set of regions $\{r_p\}$, using quick-shift clustering [17], to incorporate superpixel-level information around the pixels. The objective of object localization is then to derive a binary mask $L = \{l_p\}$, with each $l_p \in \{0, 1\}$ indicating whether the region r_p belongs to the object.

2.1. The Proposed CRF Model

We formulate the object localization problem as a binary labeling task using a new CRF model, with the following energy function:

$$E(L|I) = \underbrace{\sum_{p} \phi_L(l_p)}_{\text{local}} + \underbrace{\sum_{(p,q)\in N_S} \psi_S(l_p, l_q)}_{\text{smooth}} + \underbrace{\sum_{(p,c)\in N_C} \psi_C(l_p, l_c)}_{\text{contrast}} + \underbrace{\sum_{(p,i)\in N_R} \psi_R(l_p, l_i)}_{\text{interest-region}} \quad (1)$$

where the set of random variables or nodes of the graph is denoted by $L = \{\{l_p\} \cup \{l_c\} \cup \{l_i\}\}$, including the new auxiliary nodes from the contrast (l_c) and interest-region (l_i) potentials. The probability of a certain configuration is a conditional distribution on the energy function $E(L|I)$, and the optimal labeling is derived by minimizing the total energy using the graph cut [9].

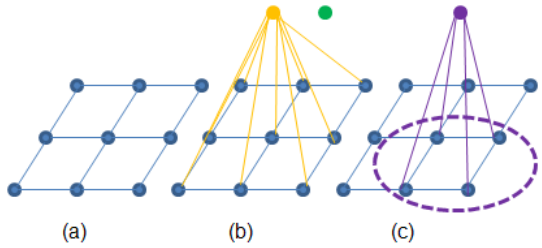


Figure 1. The proposed CRF model. (a) The standard CRF construct, with nodes representing the image regions and edges linking the neighboring regions. (b) Introducing two auxiliary nodes (object and background) for the contrast potential, with edges linking the image regions and the auxiliary nodes (showing only one set of edges for easier viewing). (c) Based on the detected interest region (purple circle), an auxiliary node for the interest-region potential is added, with edges linking all image regions in the interest-region and the added node.

The local potential $\phi_L(l_p)$ describes the cost of r_p labeled as 0 or 1:

$$\phi_L(l_p) = 1 - P(r_p = l_p|f_p) \quad (2)$$

where f_p is the local feature vector of r_p, and $P(.)$ is the probability estimate of labeling obtained using a binary support vector machine (SVM).

The smooth potential $\psi_S(r_p, r_q)$ penalizes the differences in labeling of the neighboring regions r_p and r_q based on their feature distances with a Potts model:

$$\psi_S(l_p, l_q) = \exp(-\frac{\| f_p - f_q \|^2}{2\beta_S})\mathbf{1}(l_p \neq l_q) \quad (3)$$

where β_S is the normalization factor as the average of all L2 distances between neighboring feature vectors in I. The regions r_p and r_q are considered neighbors if they share some common border in I, and the set of all neighboring pairs is denoted by N_S.

While the first two potentials follow the standard CRF constructs (Figure 1a), we describe the contrast and interest-region potentials (ψ_C, ψ_R, N_C and N_R) in the following.

2.2. Contrast Potential

To improve the labeling accuracy, we want to explore the contrast information in the image I, with the following motivations. Across different images, there are often large inter-subject variations, causing overlaps between the feature ranges and hence misclassifications. Nevertheless, within one image, there is always a certain degree of contrast between the objects and background; and the contrast information helps to discriminate between the two types.

To encode the contrast information, two additional nodes corresponding to the object and background, namely the contrast nodes l_c^o and l_c^b, are then added to the graph. A

pairwise connection between the image region l_p and each of the two nodes is also established (Figure 1b), and N_C denotes the set of all such pairwise connections. With such a construct, we expect to encourage the same labelings between the image region and contrast nodes if they exhibit similar features, and also different labelings otherwise.

To do this, we first define the unary potentials of the two contrast nodes:

$$\phi_C(l_c^{o/b}) = \left\{ \begin{array}{ll} 0 & \text{if } l_c^{o/b} = 1/0 \\ C & \text{otherwise} \end{array} \right\} \quad (4)$$

where C is a large constant, so that large costs are assigned to $l_c^o \neq 1$ and $l_c^b \neq 0$ and 0 costs otherwise, to effectively fix the labelings of the two nodes in the inference results.

We then define the pairwise potentials for the edges (l_p, l_c) with the following. First, based on the labeling outputs with local features only (Eq. (2)), we obtain the initial estimation of the objects and background areas, and two feature vectors f_c^o and f_c^b are then derived for the estimated objects and background (details of feature derivation in Sec 4). Next, we compute the contrast features between r_p and the objects and background as $g_p = \{f_p, f_p/f_c^o, f_p/f_c^b\}$, and classify the feature g_p to two classes – likely or unlikely to represent the object, denoted as $likely(1)$ and $unlikely(1)$ – using a binary SVM. Then, based on the probability estimates γ_p of class $likely(1)$, the pairwise costs are computed as:

$$\psi_C(l_p, l_c^o) = \left\{ \begin{array}{ll} 0 & \text{if } l_p = 1, \text{ and } likely(1) \\ 1-\gamma_p & \text{if } l_p = 1, \text{ and } unlikely(1) \\ \gamma_p & \text{if } l_p = 0 \end{array} \right\} \quad (5)$$

$$\psi_C(l_p, l_c^b) = \left\{ \begin{array}{ll} 0 & \text{if } l_p = 0, \text{ and } unlikely(1) \\ \gamma_p & \text{if } l_p = 0, \text{ and } likely(1) \\ 1-\gamma_p & \text{if } l_p = 1 \end{array} \right\} \quad (6)$$

Note that because of the $likely$ and $unlikely$ terms, the above pairwise potentials no longer follow the Potts model, and penalize labeling consistency if the features of the image regions and the contrast nodes are actually dissimilar. The total energy of the contrast potential can however, be rewritten in the following format, to keep it submodular (binary and with pairwise term encouraging consistency) for efficient graph-cut energy minimization:

$$\begin{aligned} \sum_{(p,c)\in N_C} \psi_C(l_p, l_c) = \sum_c \phi_C(l_c) + \\ \sum_p \alpha_p \mathbf{1}(unlikely(l_p)) + \sum_{(p,c)\in N_C} \alpha_p \mathbf{1}(l_p \neq l_c) \end{aligned} \quad (7)$$

where $\alpha_p = \gamma_p$ if $l_p = 0$, and $\alpha_p = 1 - \gamma_p$ otherwise.

2.3. Interest Region Potential

Although the contrast nodes represent the object and background regions of an image I on a global scale, the structural information between image regions are not explored. An obviously important structural information is that, regions that are likely parts of the same anatomical or pathological structure should take the same labelings.

In our formulation, the hypothesis is that, if we can detect a set of structures, i.e. interest regions R_i, the comprising regions $r_p \in R_i$ should preferably be assigned to the same category, but also depending on their individual suitability of such an labeling. The advantage of such an approach is that, we can employ a totally different method to detect the interest regions (e.g. non-CRF and different features), so the generated regions can serve as a second opinion to refine the object localization.

Assume a set of interest regions R_i are detected from an image I (details in Sec 3), and each interest region is characterized by its feature f_i, most probable label $l_i^* \in \{0, 1\}$ and a set of image regions r_p covered. Note that r_p might partially overlap with R_i especially around the boundary areas of R_i, and hence not all r_p covered by R_i should have the same label as l_i^*. To determine the the probability of $l_p = l_i^*$, we first compute the following feature vector:

$$v_p = \{\cap(r_p, R_i)/r_p, \parallel f_p - f_i \parallel, f_{i-p}/f_i\} \quad (8)$$

which represents the degrees of area overlap and feature homogeneity between r_p and R_i, with f_{i-p} denoting the feature of R_i excluding r_p. Then a binary SVM is trained to classify v_p into $same$ or $diff$ categories, indicating if $l_p = l_i^*$ or otherwise, and the probability estimate of $l_p = l_i^*$ is denoted by $\theta_{p,i}$.

Next, to integrate the interest-region detection hypothesis into the CRF formulation, for each R_i detected, a node l_i is added to the graph, with the unary potential $\phi_R(l_i)$ defined similarly to Eq. (4). An edge is then connected between each pair of (l_p, l_i) for all $r_p \in R_i$ (Figure 1c) with N_R denoting all such edges for image I, and we define the pairwise potential as:

$$\psi_R(l_p, l_i) = \theta_{p,i} \mathbf{1}(l_p \neq l_i) \quad (9)$$

Since $r_p \in R_i$ is quite likely to exhibit the same labeling as R_i, we choose to use the Potts model to encourage such consistency. The cost of different labelings is directly related to the probability of $l_p = l_i^*$, and hence we use $\theta_{p,i}$ as the pairwise cost. If r_p is less likely to be labeled as l_i^*, the use of $\theta_{p,i}$ is also able to lessen the consistency constraint.

With the above definitions, the total energy term of the interest-region potential is thus rewritten as the following:

$$\sum_{(p,i)\in N_R} \psi_R(l_p, l_i) = \sum_i \phi_R(l_i) + \sum_{(p,i)\in N_R} \theta_{p,i} \mathbf{1}(l_p \neq l_i) \quad (10)$$

2.4. Graph Inference

All energy terms are given equal weights (based on our empirical evaluation), and piecewise learnings of the prob-

ability estimates used in the local, contrast and interest-region potentials are conducted first. The binary inference problem $L^* = \text{argmin } E(L|I)$ is then solved efficiently using the graph cut.

3. Detection for Interest Region Potential

Due to our motivation of detecting the interest regions in a totally different way from the graph-based approach to support the interest-region potential (Sec 2.3), we choose to design a sparse-coding based classification method for interest-region detection. Besides its popularity and widely demonstrated effectiveness [12], we believe sparse coding can be particularly suitable for our problem because of the large variations in the dataset.

3.1. Sparse Coding for Classification

Let Y be a set of n-dimensional data samples $Y = \{y_j : j = 1, ..., J\}$ and $Y \in R^{n \times J}$. A representative dictionary for Y with K atoms is denoted as $D = \{d_k : k = 1, ..., K\} \in R^{n \times K}$. Each y_j can then be represented as a linear combination of a few (i.e. $\leq T$) atoms in D with minimum reconstruction error, and the corresponding coefficient vector x_j is the sparse code. Denoting the set of sparse codes of the data samples Y as $X = \{x_j : j = 1, ..., J\} \in R^{K \times J}$, both the dictionary D and sparse coding X can be learned with K-SVD [1] by solving the following problem:

$$\langle D, X \rangle = \underset{D,X}{\text{argmin}} \ \|Y - DX\|_2^2 \quad s.t. \forall j, \|x_j\|_0 \leq T \quad (11)$$

where $\|Y - DX\|_2^2$ represents the reconstruction error.

Once the dictionary D is learned, a given data sample y can then be represented as a sparse code x by solving the following using the OMP algorithm [16]:

$$x = \underset{x}{\text{argmin}} \ \|y - Dx\|_2^2 \quad s.t. \ \|x\|_0 \leq T \quad (12)$$

A classifier (e.g. SVM) can then be trained based on a set of such sparse codes, so that x and hence y can be classified.

In our context, an image I is divided into grid-based patches, and each image patch is represented by its feature descriptor y. The dictionary D is generated with a training set Y, and each image patch is then classified as interest region or otherwise ($h \in \{1, 0\}$) based on its sparse code x.

3.2. Discriminative Sparse Learning

A shortcoming with the previously described approach is the separation between the dictionary learning and the classifier training. There is no guarantee that the learned dictionary will produce discriminative sparse codes for the classification. Several approaches have thus been proposed to integrate the two steps of learnings as [8]:

$$\langle D, X, W \rangle = \underset{D,X,W}{\text{argmin}} \ \|Y - DX\|_2^2 + \|W\|^2 + \sum_j \mathcal{L}\{h_j, f(x_j, W)\} \quad s.t. \ \forall j, \|x_j\|_0 \leq T \quad (13)$$

where $\mathcal{L}(.)$ is the loss function, and the learned weights W for the classifier hypothesis $f(.)$ is used to produce the classification result with $l = Wx$.

However, it can be seen from the Eq. (13) that although part of the objective is to minimize the difference between the predicted hypothesis $f(.)$ and the ground truth h_j, the final result is still largely affected by the reconstruction term, which may then reduce the discriminative power of W. Furthermore, the loss function normally resembles a regression goal, and hence limiting the classification performance. Therefore, we suggest that while such an integrated approach has the advantage of generating a more discriminative dictionary D, the weights W produced are not discriminative enough for a direct classification, and hence should not totally replace the separate classifier training.

We propose a different method as follows. First, for the data samples $Y \in R^{n \times J}$, we create a corresponding labeling vector $H = \{h_j\} \in \{-1, 1\}^{1 \times J}$, with 1 for interest region. Based on linear-kernel SVM, the optimization objective of the weight vector $w \in R^{1 \times K}$ is:

$$\begin{aligned} \underset{w,\xi,b}{\text{argmin}} \quad & \frac{1}{2}\|w\|^2 + C\sum_j \xi_j \\ s.t. \quad & \forall j, h_j(w * x_j + b) \geq 1 - \xi_j, \ \xi_j \geq 0 \end{aligned} \quad (14)$$

Combining Eq. (11) and (14), we get:

$$\langle D, X, w, b \rangle = \underset{D,X,w,b}{\text{argmin}} \ \|Y - DX\|_2^2 + \frac{1}{2}\|w\|^2 + C\sum_j \xi_j$$
$$s.t. \ \forall j, \|x_j\|_0 \leq T, \ h_j(w * x_j + b) \geq 1 - \xi_j, \ \xi_j \geq 0 \quad (15)$$

which follows the general form as Eq. (13). To simplify the complexities caused by the inequality constraints on ξ_j and the signed h_j, we relax the formulation based on least squares SVM (LS-SVM) [15] as:

$$\langle D, X, w \rangle = \underset{D,X,w}{\text{argmin}} \ \|Y - DX\|_2^2 + \|w\|^2 + \sum_j \xi_j^2$$
$$s.t. \ \forall j, \|x_j\|_0 \leq T, \ h_j(w * x_j + b) = 1 - \xi_j \quad (16)$$

in which the constants are omitted for clarity. By combining w and b, and substituting ξ_j, the problem is then equivalent to the following:

$$\langle D', X', w' \rangle = \underset{D',X',w'}{\text{argmin}} \ \|Y - D'X'\|_2^2 + \|w'\|^2 + \|H - w'X'\|_2^2 \quad s.t. \ \forall j, \|x'_j\|_0 \leq T \quad (17)$$

where $w' = [w, b] \in R^{1\times(K+1)}$ and $X' \in R^{(K+1)\times J}$ appended an addition dimension with constant value 1 to absorb b, and $D' \in R^{n\times(K+1)}$ with an additional atom to be dimensionally compatible with X'.

To solve Eq. (17), an alternative approach is used:

Step 1. The objective is to find D' and X', with constraints specified by H for a more discriminative dictionary. The function is rewritten as the following:

$$\begin{aligned}\langle D', X', w'\rangle &= \underset{D',X',w'}{\operatorname{argmin}} \|Y - D'X'\|_2^2 + \|H - w'X'\|_2^2 \\ &= \| \begin{pmatrix} Y \\ H \end{pmatrix} - \begin{pmatrix} D' \\ w' \end{pmatrix} X'\|_2^2 \\ s.t. \quad &\forall j, \|x'_j\|_0 \le T \end{aligned} \tag{18}$$

which resembles the standard K-SVD formulation. However, since the $(K+1)$th dimension of X' should be 1, we modify the K-SVD algorithm by enforcing such fixed values during the alternating sparse coding steps for dictionary learning. Note that H is rescaled to the same range as Y, and the $\|w'\|^2$ term in Eq. (17) is no longer necessary because of the column-wise normalization in K-SVD.

Step 2. The computed D' and X' are input to Eq. (14) to derive w' using SVM. Although the direct solution is to use LS-SVM as in Eq. (17), we opt for the standard SVM to encourage larger positive or negative predications.

Step 3. D', and w' are then used to initialize Eq. (18) for another round of K-SVD optimization.

Several iterations of the above steps are executed to derive the final sets of D', X' and w'. To detect the interest regions, the feature vector y of an image patch is transformed to a sparse code x as a linear combination of D', which is then classified using SVM based on the final w'.

4. Implementation Details

To evaluate the proposed object localization method, we apply it to two medical image analysis tasks, and we describe the application specific details in this section.

4.1. Lesion Dissimilarity

Measuring lesion similarity is important in many medical applications, such as content-based image retrieval for diagnosis referencing. In the case of lesions visible on thoracic PET-CT images, their size and spatial extents are critical for lung cancer staging, which are hence the main criteria for dissimilarity measure. In our approach, first, lesions (i.e. lung tumors and abnormal lymph nodes) are localized in each image slice with the proposed method. Second, their textural (Gabor filters) and spatial features (circular-histogram) are extracted in 3D. Lastly, a weighted histogram-intersection is used to compute the distance, with the feature weights learned using the triplet method [4].

For the first lesion localization step, the average CT and PET intensities are used as the local features of an image region r_p. An initial SVM classification is first performed to categorize the image regions into normal lung field (LF), mediastinum (MS) and lesion areas (LA). The identified LF is accurate enough due to its distinct features, but not the other two groups, especially the MS regions could contain lower-intensity lesions and the boundary areas of lesions. The CRF formulation is thus used to relabel the (MS+LA) regions to lesion and mediastinum. The object feature f_c^o is computed as the average intensities of LA regions, and the background feature f_c^b from the largest connected component of the MS regions that approximates the mediastinum. For the interest-region detection, the image slices are divided into 4×4 patches, and the mean, standard deviation, minimum and maximum intensities computed from both CT and PET are used as the patch feature descriptor y, which are then trained for sparse coding and classification of interest regions (i.e. lesions).

4.2. Cell Segmentation

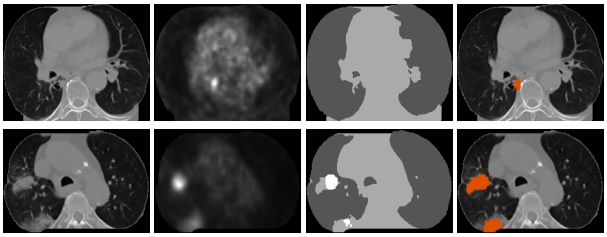
Cell nucleus segmentation is one of the most important tasks in analyzing and quantifying fluorescence microscopic images. In our approach, the cell nucleus is localized using the proposed method; and since the localization results also tend to delineate the nucleus boundaries closely, such an approach can be directly used for segmentation.

A 10-dimensional feature vector is used to describe the image region r_p: average RGB/LUV/RGB after Gabor filtering, and average grayscale intensity. An initial SVM classification is performed to categorize the image regions into nucleus (NL), cytoplasm (CP) and background (BG), to filter BG from further processing. The intensity values of the NL and CP regions are quite similar and often vary largely between different images, hence causing mislabelings in the initial classification. The CRF formulation is then used to relabel the NL and CP regions. The object features f_c^o and f_c^b are computed as the average intensities of the NL and CP regions, respectively. For the interest-region detection, the images are divided into 8×8 patches, and each patch is described by its mean, standard deviation, minimum and maximum values in RGB and grayscale spaces as y for sparse coding. And interest regions representing both NL and CP are detected, to have a good separation between the two types for an accurate segmentation.

5. Experimental Results

5.1. Results on Lesion Dissimilarity

The datasets comprise of 40 thoracic PET-CT 3D image sets from non-small cell lung cancer studies. Each image set has on average 25 transaxial PET-CT slice pairs. A total of 64 lesions including lung tumors and abnormal lymph



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

Figure 2. Two example localization outputs. (a) Transaxial CT image slices (showing the thorax after preprocessing). (b) Co-registered PET image slices. (c) The labeling outputs using standard CRF, with dark gray for lung field, light gray for mediastinum and white for lesion. (d) Our localization outputs with the two additional potentials, with lesions highlighted in orange.

Table 1. The localization performances comparing our proposed method with standard CRF.

	Recall (%)	Precision (%)	F-score (%)
Ours	97.0	95.4	96.2
CRF	76.6	94.2	84.5

nodes are annotated, and the similarity/dissimilarity relationships between each pair of 3D image sets are marked as the ground truth. An automatic preprocessing is performed on each image slice to remove the soft tissues outside of the lung field and mediastinum using thresholding and connected component analysis. Three image sets showing typical thoracic characteristics are selected for training, and testing is performed on all image sets.

Figure 2 shows examples of the lesion localization. The first example illustrates the benefits of the contrast potential, in which the lesion is initially not detected with standard CRF, due to the relatively low PET intensities. The interest-region potential is particularly useful in refining the lesion boundaries, which tend to be underestimated with the standard CRF, as shown in the second example.

It is observed that, the standard CRF tends to produce a large number of either totally undetected or underestimated lesions; and our proposed method is able to enhance the localization performance. Based on the measured 3D object-level true positives (TP), i.e. accurately localized with >50% lesion volume, false positives (FP) and false negatives (FN), we summarize the localization recall, precision and F-score in Table 1. Our proposed method demonstrates much higher recall and F-score comparing to the standard CRF approach.

The localized lesions are then used to retrieval images with similar lesions. The retrieval tests are performed by using each 3D image set as a query image, and the remaining 39 images are ranked accordingly. If an image set contains multiple lesions, the most prominent lesion (normally primary lung tumor) is used as the representing feature. We compare the retrieval performance with three other approaches: (i) the state-of-the-art labeling-based method [14] for thoracic PET-CT image retrieval; (ii) a bag-of-words representation using spatial pyramid matching (SPM), with local intensity features extracted from grid-based image patches; and (iii) a bag-of-words representation using the SIFT descriptor. All methods are trained in the same way to have optimal feature weights for the dissimilarity computation. As shown in Figure 3, our proposed method exhibits the highest retrieval precisions for all recall levels. The SPM approach also demonstrates good performance, suggesting the advantage of integrating spatial information into the image descriptor.

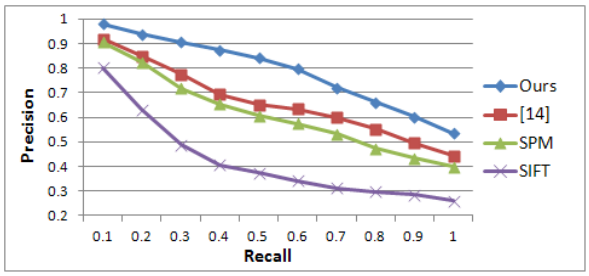


Figure 3. The retrieval precision and recall.

5.2. Results on Cell Segmentation

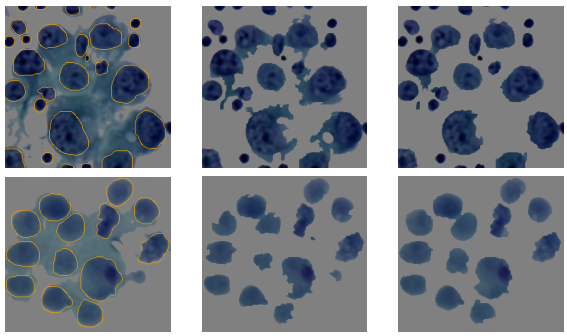
The serous database [11] is used to evaluate the cell segmentation. The database contains 10 microscopic images from serous cytology, each of 512×512 pixels. A total of 254 cell nuclei are present in the images, with ground truth of cell nuclei segmentation provided. The images show isolated or touching cells as well as clustered or overlapping cells, and the color of the nuclei can range from very dark to very pale blue. Same as [3], half of the images are used for training and the others for testing.

To evaluate the segmentation performance, we compute the PASCAL VOC criteria of pixel- and object-level accuracies, both as TP/(TP+FN+FP). We also compare our results with three approaches: (i) L+S, the standard CRF with local and smooth potentials; (ii) L+S+C, with additional contrast potential; (iii) L+S+R, with additional interest-region potential; and (iv) the state-of-the-art discriminative labeling method [3] reported for the same database.

As listed in Table 2, our method achieves the highest pixel- and object-level accuracies. The improvements of having the contrast and potential terms are evident. The method [3] produces the second highest pixel-level accuracy, by classifying superpixel-based appearance, shape and context features. The performance difference between L+S and [3] suggests that if we incorporate the feature set of

Table 2. The segmentation results comparing various methods.

	Ours	L+S	L+S+C	L+S+R	[3]
Pixel Acc (%)	85.6	82.0	83.1	84.6	85.1
Obj Acc (%)	89.3	84.5	86.2	88.7	84.0



(a) (b) (c)

Figure 4. Two example segmentation results. (a) Cropped microscopic images, with orange circles delineating the segmentation ground truth. (b) The segmentation results with L+S. (c) The segmentation results of our proposed method.

[3], the segmentation accuracies would be further improved. However, we use only simple intensity features here to mainly demonstrate our object localization idea.

We also test replacing the interest-region detection with standard sparse-coding classification, to evaluate the usefulness of introducing the discriminative dictionary learning enhancement. It is found that our proposed method exhibits on average 1.1% improvement for both pixel- and object-level measurements with the new approach.

The first example shown in Figure 4 indicates that our method is quite effective in removing the cytoplasm areas that connect the cell nuclei. As shown in the second example, lighter intensities of the cell nuclei cause many false negatives with the standard CRF approach; and our result shows more accurate delineations of the actual contours.

6. Conclusion

In this paper, we present a new method for object localization in medical images. A new CRF model with additional contrast and interest-region potentials are designed for effective object localization, addressing large inter-subject variations and low feature differences between the objects and background. A new sparse-coding classification approach is also designed for the interest-region detection, with enhanced discriminative power of the learned dictionaries. We evaluate the proposed method on lesion dissimilarity on thoracic PET-CT images, and cell segmentation on microscopic images, and our method shows higher performance compared to the state-of-the-art techniques.

References

[1] M. Aharon, M. Elad, and A. Bruckstein. K-SVD: an algorithm for designing overcomplete dictionaries for sparse representation. *TSP*, 54(1):4311–4322, 2006. 4

[2] I. Ayed, K. Punithakumar, G. Garvin, W. Romano, and S. Li. Graph cuts with invariant object-interaction priors: application to intervertebral disc segmentation. *In IPMI*, 6801:221–232, 2011. 2

[3] L. Cheng, N. Ye, W. Yu, and A. Cheah. Discriminative segmentation of microscopic cellular images. *In MICCAI*, 6891:637–644, 2011. 1, 2, 6, 7

[4] A. Frome, Y. Singer, F. Sha, and J. Malik. Learning globally-consistent local distance functions for shape-based image retrieval and classification. *In ICCV*, pages 1–8, 2007. 5

[5] J. Gonfaus and X. Boix. Harmony potentials for joint classification and segmentation. *In CVPR*, pages 3280–3287, 2010. 2

[6] R. Guo, Q. Dai, and D. Hoiem. Single-image shadow detection and removal using paired regions. *In CVPR*, pages 2033–2040, 2011. 2

[7] V. Jagadeesh, N. Vu, and B. Manjunath. Multiple structure tracing in 3D electron micrographs. *In MICCAI*, 6891:613–620, 2011. 2

[8] Z. Jiang, Z. Lin, and L. Davis. Learning a discriminative dictionary for sparse coding via label consistent K-SVD. *In CVPR*, pages 1697–1704, 2011. 4

[9] V. Kolmogorov and R. Zabih. What energy functions can be minimized via graph cuts? *TPAMI*, 26(2):147–159, 2004. 2

[10] L. Ladicky, P. Sturgess, K. Alahari, C. Russell, and P. Torr. What, where and how many? Combining object detectors and CRFs. *In ECCV*, 6314:424–437, 2010. 2

[11] O. Lezoray and H. Cardot. Cooperation of color pixel classification schemes and color watershed: a study for microscopical images. *TIP*, 11(7):783–789, 2002. 6

[12] M. Liu, L. Lu, X. Ye, S. Yu, and M. Salganicoff. Sparse classification for computer aided diagnosis using learned dictionaries. *In MICCAI*, 6893:41–48, 2011. 4

[13] J. Shotton, J. Winn, C. Rother, and A. Criminisi. Textonboost: joint appearance, shape and context modeling for multi-class object recognition and segmentation. *In ECCV*, 3951:1–15, 2006. 2

[14] Y. Song, W. Cai, S. Eberl, M. Fulham, and D. Feng. Discriminative pathological context detection in thoracic images based on multi-level inference. *In MICCAI*, 6893:191–198, 2011. 1, 6

[15] J. Suykens and J. Vandewalle. Least squares support vector machine classifiers. *NPL*, 9(3):293–300, 1999. 4

[16] J. Tropp. Greed is good: algorithmic results for sparse approximation. *TIT*, 50:2231–2242, 2004. 4

[17] A. Vedaldi and S. Soatto. Quick shift and kernel methods for mode seeking. *In ECCV*, 5305:705–718, 2008. 2

[18] D. Wu, L. Lu, J. Bi, Y. Shinagawa, K. Boyer, A. Krishnan, and M. Salganicoff. Stratified learning of local anatomical context for lung nodules in CT images. *In CVPR*, pages 2791–2798, 2010. 1, 2