Combining CNN and MIL to Assist Hotspot Segmentation in Bone Scintigraphy

Shijie Geng, Shaoyong Jia, Yu Qiao*, Jie Yang, and Zhenhong Jia

[1] Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, Shanghai, China.
[2] School of Information Science and Engineering, Xinjiang University, Urumqi, China.
{jeykigung,jiashaoyong,qiaoyu,jieyang}@sjtu.edu.cn,jzhh@xju.edu.cn

Abstract. Bone scintigraphy is widely used to diagnose tumor metastases. It is of great importance to accurately locate and segment hotspots from bone scintigraphy. Previous computer-aided diagnosis methods mainly focus on locating abnormalities instead of accurately segmenting them. In this paper, we propose a new framework that accomplish the two tasks at the same time. We first use sparse autoencoder and convolution neural network (CNN) to train an image-level classifier that label input image as normal or suspected. For suspected images, multiple instance learning (MIL) is applied to train a patch-level classifier. Then we use this classifier to produce a probability map of hotspots. Finally, level set segmentation is performed with the probability map as initial condition. The experimental results demonstrate that our method is more accurate and robust than other methods.

Keywords: Hotspot segmentation, bone scintigraphy, multiple instance learning, CNN, level set method.

1 Introduction

Bone scintigraphy is very effective in diagnosing cancer and tumor metastases [14]. The abnormalities in bone scintigraphy are called "hotspot", which generally appear to be brighter than its surroundings. It is of great clinical importance to accurate detect and segment hotspots from bone scintigraphy. Many computer-aided diagnosis (CAD) systems have been developed to detect and segment hotspots. May Sadik et al. [10] used adaptive threshold of a specific region for hotspot segmentation. Huang et al. [5] uses linear regression model to find regional threshold to extract hotspot. Chang et al. [3] proposed an algorithm utilizes Gaussian function to approximate intensity probability distribution and perform hotspot segmentation via adaptive region growing [4]. However, most of proposed methods focus on detection of hotspots rather than accurate segmentation. Adaptive threshold and region growing [4] are two most commonly used approaches. Obviously, due to weak boundary contrast and low signal

* Corresponding author: Yu Qiao, qiaoyu@sjtu.edu.cn

noise ratio of bone scintigraphy, these two methods will not achieve satisfactory segmentation results.

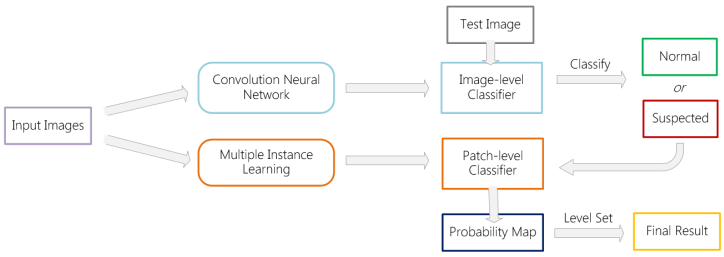

Fig. 1. Overview of our framework for accurate hotspot segmentation. We use convolution neural network to train an image-level classifier and multiple instance learning to train a patch-level classifier respectively. For test image, we first label it as normal or suspected. If the image is suspected, we will produce a probability map and perform segmentation via level set algorithm.

Due to advantages in sub-pixel accuracy and topology variability [11], level set methods [11, 8] are widely used in image segmentation. They evolve a user-specified initial contour in order to minimize a given energy function. Level set methods can be divided into two types: edge-based and region-based. In terms of weak object boundaries, region-based methods have better performance compared with edge-based ones. In this paper, we choose region-based level set with local signed difference energy (LSD) [11] to perform accurate hotspots segmentation. In order to perform a level set segmentation, we need to set an initial contour for the algorithm. Generally, if the initial contour gets closer to the target object, the final segmentation result will be better. A probability map of hotspots is helpful to initialization of level set algorithm. In this paper, we apply convolution neural network (CNN) and multiple instance learning (MIL) to get the probability map.

Deep learning methods have achieve great success in these years. It has been applied to object recognition in ImageNet [7] and feature learning from unlabeled data [9, 6]. Deep learning shows its robustness and discriminative power in feature learning and classification tasks. In this paper,we train an image-level classifier using sparse coding and CNN to classify input image as normal or suspected. If the input image is labeled as normal, it will not be segmented in following steps. MIL technique has been widely used in several scenarios, such as MIL-Boost [2] in object tracking , MCIL [13] in medical diagnosis, MILCUT [12] in natural image segmentation. It has many advantages, such as reducing the efforts in human annotations and automatically exploiting information from data [13]. In this paper, we apply MCIL algorithm to train a patch-level classifier. Given an input image, the classifier can produce a probability map of hotspots.

Fig. 1 gives an illustration of our framework. Experiments in Section.3 demonstrate the effectiveness of our framework in hotspot detection and segmentation over previous methods.

2 Methodology

2.1 Using CNN to train image-level classifier

In this section, we use sparse autoencoder [9] to learn bases from image patches which extracted from bone scintigraphy. To extract feature representation from the hole image, we utilize CNN which consists of convolution and mean pooling layers. Finally, we input the pooled features to SVM and train an image-level classifier. Fig. 2 is a diagram of the overall training process. To increasing discrimination between pixels with similar intensities, all images used in this section are first mapped to RGB color space by applying density slicing. The processed 3-channel images are showed in Fig. 2. Detailed explanations can be found in experiments section.

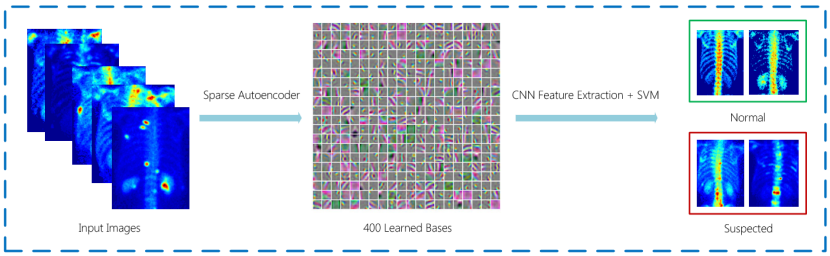

Fig. 2. Diagram of the overall training process: we first use sparse autoencoder to learn 400 bases from image patches. Then we extract feature representation using CNN. Finally we train image-level classifier via SVM.

Sparse autoencoder is an unsupervised learning algorithm that tries to learn higher-level representation of input images. It finds basic elements (bases) of image patches. We first randomly extract a large number of small patches from thoracic bone scintigraphy. Then we convert these $d \times d$ patches to unlabeled vectors $\{x_1, \ldots, x_m\}$, where $x_i \in \mathbb{R}^s$, $s = d \times d$. We denote the activation of basis ϕ_j for input x_i as $a_{i,j}$. Our task is to solve the following optimization problem:

$$\min_{a,\phi} \sum_{i=1}^{m} \left\| x_i - \sum_{j=1}^{k} a_{i,j}\phi_j \right\|_2^2 + \lambda \sum_{i=1}^{m} \sum_{j=1}^{k} \|a_{i,j}\|_1 \tag{1}$$

After solving the above problem, we get all bases vector $\Phi = \{\phi_1, \ldots, \phi_k\}$, where $\phi_j \in \mathbb{R}^s$. In this paper, we set k to 400, the middle of Fig. 2 shows the 400

learned bases via sparse coding. Since the size of input images are not uniform, as a preprocessing step, we resize them to uniform $n * n$ size. Then we apply convolution and mean pooling for these resized input images.

To get the convolved features, for every $d \times d$ region of the $n \times n$ image, we convolve it with the learned bases to get the feature activations. As a result, we will get $k \times (n-d+1) \times (n-d+1)$ array of convolved features. For mean pooling, we divide the convolved image into $l \times l$ equal parts, and calculate the mean for each part, then produce $k \times l \times l$ pooled features. We construct a training set T consists of pooled features and human labels, and train an image-level classifier with SVM algorithm.

2.2 Using MIL to train patch-level classifier

In this section, we use MIL method to train patch-level classifier which can produce probability map of hotspots. To learn detail information from input data, we need to set patches to a relatively small size. However, when the size of patches get smaller, it will be harder for human labeling. In multiple instance learning, instances are not directly labeled, it only needs bag labels. This property will largely reduce the efforts in human annotations. For this reason, we choose MIL to train patch-level classifier.

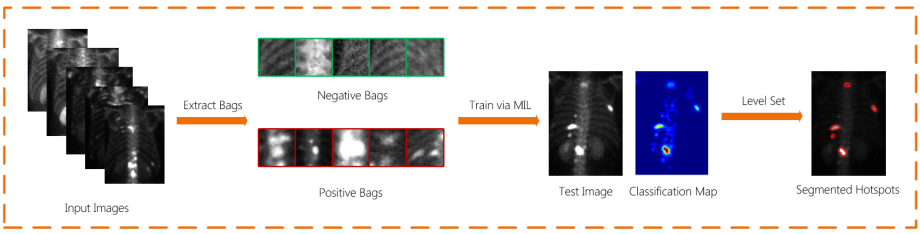

Fig. 3. The flow chart of our hotspot detection and segmentation process.

MCIL algorithm As shows in Fig. 3, we extract positive and negative bags consist of 4×4 instances from input images to construct training datasets. We denote a bag as $X_i = \{x_{i1}, \ldots, x_{in}\}$, in which i is the bag index and j is the instance index. The label of bag is denoted as $y_i \in \{-1, 1\}$, the label of instance $y_{ij} \in \{-1, 1\}$ is latent during training. During the training process, MCIL [13] tries to minimize the following negative log-likelihood loss function:

$$\mathcal{L}(H) = -\sum_{i=1}^{n} (\mathbf{1}(y_i = 1) \log H(X_i) + \mathbf{1}(y_i = -1) \log(1 - H(X_i))) \qquad (2)$$

Here, $H(X_i)$ is the probability of bag i, $h(x_{ij})$ is the probability of instance j in bag i. The relationship between $H(X_i)$ and $h(x_{ij})$ is $H(X_i) = \max_j(h(x_{ij}))$.

MCIL compute new weights as follows: $w_{ij} = -\frac{\partial \mathcal{L}(H)}{\partial h(x_{ij})}$, then use these weights to select best weak classifier h_t. Line search is then performed to find the coefficient α_t of the selected weak classifier. After training given number of weak classifiers, we get the final strong classifier : $h(x_{ij}) = \sum_{t=1}^{T} \alpha_t h_t(x_{ij})$.

Feature Extraction for Bags and Instances In our work, we select relatively large regions from bone scan images and treat them as bags. For each bag, we and densely sample 4×4 patches as instances from it. The overlap step size is 2 pixels. We then extract a 29-dimension feature vector for each instance, including 11-dimension intensity features and 18-dimension texture features. Intensity features consist of statistical histogram, 4-neighborhood contrast and symmetric contrast. We use weighted difference to compute 4-neighborhood contrast: $Neighbor\ Contrast(M, N) = \frac{1}{10} \sum_{i=1}^{n} 2^i (M_i - N_i)$, where M and N are histograms of center patch and 4-neighborhood patch respectively. For symmetric contrast, we first find symmetric patch by calculating body central line as [3] does, then we compute symmetric contrast using modified chi-square distance: $Symmetry\ Contrast(M, S) = \sum_{i=1}^{n} \frac{(M_i - S_i)^2}{M_i + S_i + 1}$, where M and S are histograms of center patch and symmetric patch respectively. For texture feature, we use a variant of local binary patterns (LBP) [1], in which the threshold used is the mean intensity of the whole image. We collect 18947 instances to form 39 positive bags and 33 negative bags. Finally we use MCIL to train the patch-level classifier.

2.3 Using level set to segment hotspot

After getting the probability map of hotspot, we use local signed difference (LSD) [11] level set algorithm to segment hotspots from input image. LSD is able to deal with intensity inhomogeneity and weak object boundaries. It also consider global information – the order of local clusters, thus leads to robust segmentation performance.

Let C denote the contour evolving in image domain Ω, $f_1(x)$ and $f_2(x)$ are local clusters inside and outside contour C. The LSD energy is expressed as follows:

$$\mathcal{E}(C) = \int_{\Omega} \operatorname{sgn}(f_2(x) - f_1(x)) \left| f_2(x) - f_1(x) \right| dx + \mu Length(C) \tag{3}$$

According to level set formulation, we use a Lipschitz function $\phi : \Omega \to R$ to represent LSD energy. To initialize the level set function ϕ_0, we utilize the probability map of hotspot $\mathcal{P}(x)$:

$$\phi_0 = \phi(x, t = 0) = \begin{cases} -c, & x \in \{x | \mathcal{P}(x) > \rho\} \\ 0, & x \in \{x | \mathcal{P}(x) = \rho\} \\ c, & x \in \{x | \mathcal{P}(x) < \rho\} \end{cases} \tag{4}$$

We set ρ to 0.5 in this paper. Then we apply gradient descent method to minimize the LSD energy according to the equation: $\frac{\partial \phi}{\partial t} = -\frac{\mathcal{E}(\phi)}{\partial \phi}$.

3 Experiments

In this section, we conduct three experiments to demonstrate the effectiveness of our framework. We first give the reason why we use mapped color images to train image-level classifiers. In the second experiment, we make a comparison between multiple instance learning and supervised learning methods in terms of classification accuracy. At last, we show that our framework gets better segmentation results compared with previous methods. The bone scintigraphy images used in training are collected from Department of Nuclear Medicine, Shanghai Renji Hospital. Image labels used in training image-level classifier come from the gold standard of radiologist diagnosis.

Experiment I In this experiment, we select 1030 images and resize them to 100×100. We extract 100000 patches with size of 11×11 to train sparse autoencoder. We respectively use mapped color images, original grayscale images as input images, and test recognition accuracy with 5-fold cross validation. Table 1 gives the tesing results. It shows that features learned from mapped color images have better discriminative ability compared with grayscale images.

Table 1. Cross validation accuracy of different input dataset

Input Dataset	Color Images	Grayscale Images
Cross Validation Accuracy	0.947	0.895

Experiment II In this experiment, we respectively train patch-level classifier with MCIL [13] and SVM. The training dataset consists of 18947 instances from 39 positive bags and 33 negative bags. For SVM, bag labels are directly regarded as instance labels. We also construct a testing dataset consists of 1050 labeled instances, each instance forms a bag. Table 2 test the classification accuracy of two algorithms. It shows that multiple instance learning can exploit information from given data, and the classification accuracy is enough to get an approximate probability map of hotspot locations.

Experiment III In this experiment, we compare our hotspot segmentation framework with two previous methods – adaptive region growing [3] and regional threshold [5]. We select 68 suspected thoracic images consists of 572 hotspots as testing dataset and invited expert radiologist to annotate the ground truths for these images. In this paper, Jaccard (J) and Dice (D) index are used to

Table 2. Accuracy comparison between the two algorithms

Algorithm	Semi-supervised (MCIL)	Supervised(SVM)
Classification Accuracy	0.905	0.783

evaluate the performance of hotspot segmentation. They are defined as: $J = \frac{|S \cap G|}{|S \cup G|}$, $D = \frac{2|S \cap G|}{|S|+|G|}$. Table 3 shows the performance of the three methods in hotspot segmentation:

Table 3. The performance comparison on suspected thoracic images

Method	Our method	Region growing [3]	Threshold [5]
Jaccard (J)	0.8051	0.6505	0.5183
Dice (D)	0.8887	0.7613	0.6574

From Table 3, we can conclude that our framework gets more accurate segmentation results compared to the other two methods. In Fig. 4, we give a illustration of different segmentation results and human labeled ground truths. We can see that our method also get better segmentation results in terms of appearance. Our results are with smooth boundaries and very close to ground truths.

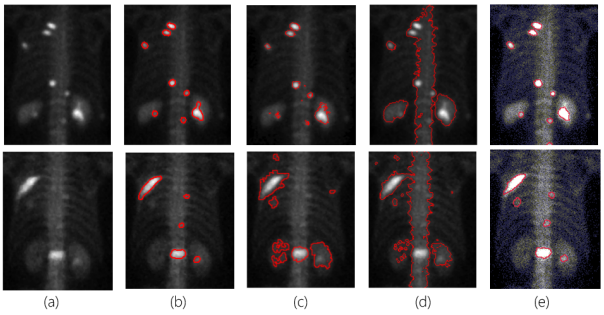

Fig. 4. Segmentation result comparison: (a) input image (b) our method (c) adaptive region growing [3] (d) regional threshold [5] (e) human annotations.

4 Conclusion

In this paper, we propose a novel framework which combine CNN and MIL to assist accurate hotspot segmentation. Experiments show that our method has

advantage over existed methods and gets more accurate segmentation results. In future work, we will extend our framework from thoracic region to other regions in bone scintigraphy.

Acknowledgments. This research is partly supported by NSFC, China (No: 61375048).

References

1. Ahonen, T., Hadid, A., Pietikainen, M.: Face description with local binary patterns: Application to face recognition. Pattern Analysis and Machine Intelligence, IEEE Transactions on 28(12), 2037–2041 (2006)
2. Babenko, B., Dollár, P., Tu, Z., Belongie, S., et al.: Simultaneous learning and alignment: Multi-instance and multi-pose learning. In: Workshop on Faces in'Real-Life'Images: Detection, Alignment, and Recognition (2008)
3. Chang, Q., Wang, Q., Qiao, Y., Zhu, Y., Huang, G., Yang, J.: Adaptive detection of hotspots in thoracic spine from bone scintigraphy. In: Neural Information Processing. pp. 257–264. Springer (2011)
4. Hojjatoleslami, S., Kittler, J.: Region growing: a new approach. IEEE Transactions on Image processing 7(7), 1079–1084 (1998)
5. Huang, J.Y., Kao, P.F., Chen, Y.S.: A set of image processing algorithms for computer-aided diagnosis in nuclear medicine whole body bone scan images. Nuclear Science, IEEE Transactions on 54(3), 514–522 (2007)
6. Kim, M., Wu, G., Shen, D.: Unsupervised deep learning for hippocampus segmentation in 7.0 tesla mr images. In: Machine Learning in Medical Imaging, pp. 1–8. Springer (2013)
7. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems. pp. 1097–1105 (2012)
8. Li, C., Xu, C., Gui, C., Fox, M.D.: Distance regularized level set evolution and its application to image segmentation. Image Processing, IEEE Transactions on 19(12), 3243–3254 (2010)
9. Raina, R., Battle, A., Lee, H., Packer, B., Ng, A.Y.: Self-taught learning: transfer learning from unlabeled data. In: Proceedings of the 24th international conference on Machine learning. pp. 759–766. ACM (2007)
10. Sadik, M., Hamadeh, I., Nordblom, P., Suurkula, M., Höglund, P., Ohlsson, M., Edenbrandt, L.: Computer-assisted interpretation of planar whole-body bone scans. Journal of Nuclear Medicine 49(12), 1958–1965 (2008)
11. Wang, L., Wu, H., Pan, C.: Region-based image segmentation with local signed difference energy. Pattern Recognition Letters 34(6), 637–645 (2013)
12. Wu, J., Zhao, Y., Zhu, J.Y., Luo, S., Tu, Z.: Milcut: A sweeping line multiple instance learning paradigm for interactive image segmentation. In: Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on. IEEE (2014)
13. Xu, Y., Zhu, J.Y., Eric, I., Chang, C., Lai, M., Tu, Z.: Weakly supervised histopathology cancer image segmentation and classification. Medical image analysis 18(3), 591–604 (2014)
14. Yin, T.K., Chiu, N.T.: A computer-aided diagnosis for locating abnormalities in bone scintigraphy by a fuzzy system with a three-step minimization approach. Medical Imaging, IEEE Transactions on 23(5), 639–654 (2004)