

# Deep learning based feature representation for automated skin histopathological image annotation

Gang Zhang<sup>1</sup> · Ching-Hsien Robert Hsu<sup>2</sup> ·  
Huadong Lai<sup>1</sup> · Xianghan Zheng<sup>3</sup>

Received: 21 June 2016 / Revised: 4 November 2016 / Accepted: 1 May 2017 /  
Published online: 23 May 2017  
© Springer Science+Business Media New York 2017

**Abstract** Automated annotation of skin biopsy histopathological images provides valuable information and supports for diagnosis, especially for the discrimination between malignant and benign lesions. Currently, computer-aid analysis of skin biopsy images mostly relied on some human-designed features, which requires expensive human efforts and experiences in problem domains. In this study, we propose an annotation framework for automated skin biopsy image analysis which makes use of a deep model for image feature representation. A convolutional neural network (CNN) is designed for local regions of skin biopsy images which learns potential high-level features automatically from input raw pixels. The annotation model is constructed in the multiple-instance multiple-label (MIML) learning framework with the features learned through the network. We achieve significant improvement of the model performance on a real world clinical skin biopsy image dataset and a benchmark dataset. Moreover, our study indicates that deep learning based model could achieve better performance than human designed features.

---

✉ Xianghan Zheng  
xianghan.zheng@fzu.edu.cn

Gang Zhang  
ipx@gdut.edu.cn

Ching-Hsien Robert Hsu  
chh@chu.edu.tw

Huadong Lai  
961649205@qq.com

<sup>1</sup> School of Automation, Guangdong University of Technology, Guangzhou, China

<sup>2</sup> Department of Computer Science and Information Engineering, Chung Hua University, Hsinchu, 300, Taiwan

<sup>3</sup> College of Mathematics and Computer Science, Fuzhou University, Fuzhou, China

**Keywords** Skin biopsy histopathological image annotation · Deep learning · Convolutional neural network · Multiple-instance multiple-label learning · Unsupervised feature learning

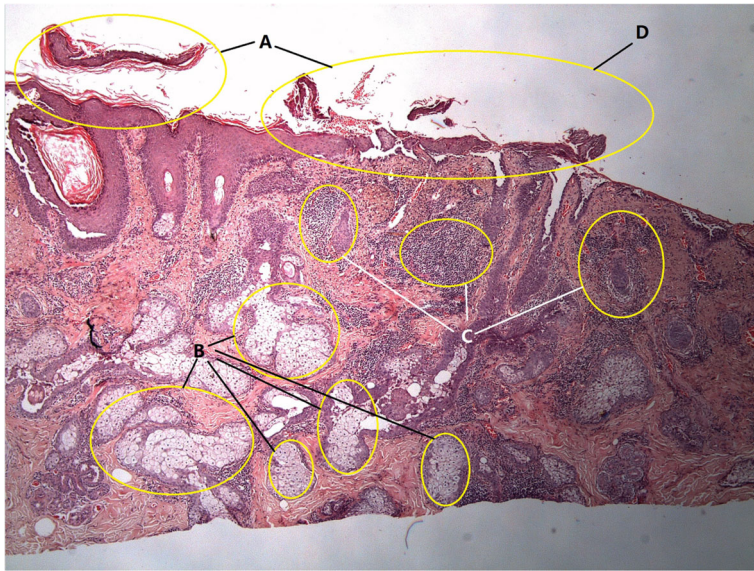
## 1 Introduction

It is well-known that skin diseases are very common in life. There are many kinds of skin diseases in which many of them are not harmful to our health while some of them would lead to serious problems if not been treated properly in their initial stages. Malignant melanoma is a fatal skin cancer but it would look just like some harmless nevus in some cases. Pemphigus characterizes in most cases by the development of blisters on skin is a rare skin disorder which would lead to severe tissue infection. Many of the skin diseases can be diagnosed easily through clinical symptoms, physical examination or laboratory examination. However, some skin diseases, especially the fatal ones, cannot be diagnosed correctly only by simple examinations. To facilitate medical diagnosis for some serious skin disorders, doctor prefers to do a biopsy and analyze histopathological images of lesion tissues. In fact, skin biopsy histopathological images are widely used in dermatological department and critical to dermatologists. It is well accepted they are the gold standard in diagnosis of skin cancers. In many clinical departments related to dermatological department, especially department of Surgery, Obstetrics and Gynecology, and Dermatology need biopsy examination for their medical decisions.

The important role of biopsy histopathological analysis in dermatological department poses the pressing need of an effective computer aid diagnosis (CAD) system on either medical or machine learning researchers. Automated skin biopsy histopathological analysis can release the burden of doctors from common but frequent-occurring histopathological characteristics and make them focus on rare and obscure cases. A CAD system of skin biopsy images can also give some suggestions of what histopathological features an image indicates [24]. But there are some significant challenges in constructing a CAD system for automated annotation of skin histopathological images. First of all, a single skin histopathological characteristic is only associated with some local regions in an image, while an image may have several histopathological characteristics, as shown in Fig. 1.

In Fig. 1, letters A, B, C and D stand for epidermis erosion, blistering, acantholysis and infiltration of lymphocytes, respectively, which are 4 common histopathological characteristics in dermatology. The correspondences between local regions and a certain histopathological characteristics have been manually labeled in Fig. 1. But the concept of local region is subjective and fuzzy. When examining a skin biopsy image, a doctor probably will focus on the regions with conspicuous features at first glance. And then search some regions to find whether there are histopathological features that would confirm his potential diagnosis. But the local regions would not exist after the diagnosis is drawn, neither would the above correspondences. In the database of diagnosis records, histopathological features are associated with a whole image, which is addressed as annotation terms ambiguity in multiple-instance studies [36, 37].

Meanwhile, a skin biopsy histopathological image has complicated features, such as color, light, texture or inner structures, making it difficult to be modeled mathematically or statistically [3]. Most machine learning models require good feature representations of data samples either for training or predicting. For the problem of skin biopsy histopathological image annotation, it is expensive for a doctor to explicitly express some key features



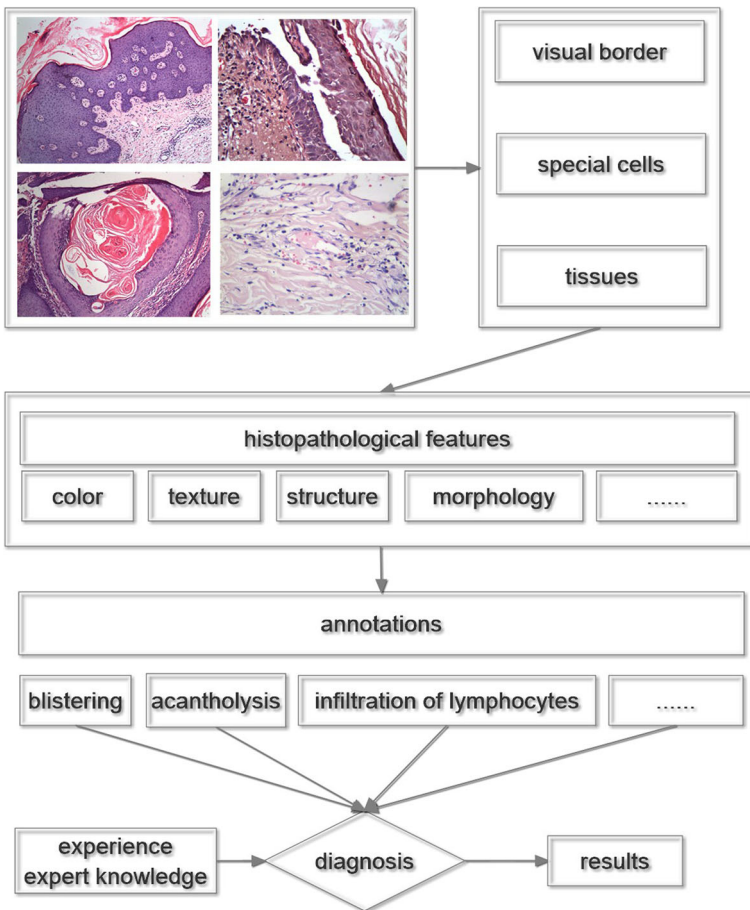
**Fig. 1** Histopathological characteristics and their corresponding local regions

to identify a certain histopathological feature based on his medical experience [14]. This representation problem is also essential in solving many problems of machine learning and artificial intelligence, which becomes a gap between human knowledge and machine intelligence [4, 10].

Many problem-oriented medical image feature representation methods have been proposed in current literatures, and some of them have reported the success of building disease-oriented analysis models [25, 40]. Bunte et al. [7] proposed a feature learning method based on metric learning [6] for skin surface image classification. Their model learns a vector of weights with a training set by performing a weighted combination of 8 color space feature vectors (e.g. RGB, LUV, HSV, etc.) together. The combined color feature vector is the best feature representation given the training set. Their work shows a good sample of feature learning, in which the feature is learned through the combination of previous designed features. However, their method can only work with color-based features having the same dimensions. Heterogeneous features cannot fit in their method directly. Moreover, basic features have to be designed or chosen manually and the learning task has to be performed in a supervised manner. Another study on medical image feature representation is the Bag-Of-Features (BOF) method [9, 15], which builds a matrix (codebook) containing patches gathered from a training histopathological image set by a clustering-based method. It expresses each training or test image as a histogram, indicating which and how many patches in the codebook the image contains. The histogram is then converted into a feature vector for training and test. The similarities between patches in the image and codebook are measured by some distance function, e.g. Euclidean distance. Current studies show that low level elements can build high level concepts through multiple layer network structures [19], but histogram-based methods adopting simple statistical operations seem not able to generate high level concepts.

Some region-based feature extraction methods for histopathological image analysis have been proposed recently [30, 35, 37]. In these methods, a skin biopsy image is segmented into several visual disjoint regions based on textures or visual edges. Then a region-based feature extraction procedure is applied to generate hybrid features based on color, texture, structure and morphology. However, the region-based features are designed by experts and most of them are previously successful in different application domains. In our previous work [37], each region is first re-expressed in LUV color space, and then a 2D wavelet transformation is applied to the region. The feature is composed of the means of  $L$ ,  $U$ ,  $V$ , wavelet transformation coefficients and some morphological factors of a region. Figure 2 briefly summaries the above methods.

Ali et al. [1] proposed a structure-based feature representation method for micro-computed tomography images analysis. In their work, a region is treated as a set of pixels. The method transforms an image into a graph in which each node is the centroid of a cluster of pixels belonging to the same type of tissue. Then several numerical graph properties are calculated as features of the original image. Though their method is an unsupervised

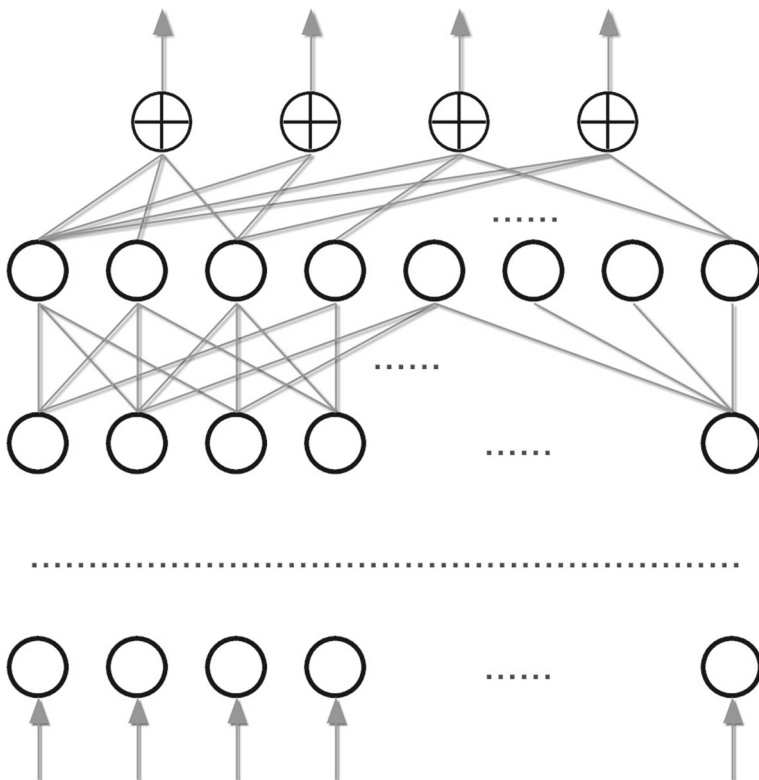


**Fig. 2** Main steps of region-based histopathological image diagnosis

one, it requires the number of different tissue types before running which cannot be directly applied to different application domains.

To tackle the problems of current studies on skin biopsy image analysis, in this study, we propose an unsupervised learning method for region-based feature extraction. The method firstly cuts a histopathological image into disjoint regions with a self-adaptive strategy, so as not to explicitly set the number of regions to be generated. Then a convolutional neural network (CNN) is applied to each generated region to learn the feature representation, also called region encoding. We train the CNN model in a supervised manner with a part of the training dataset. Finally we solve the annotation problem with a multiple instance multiple label (MIML) algorithm. The motivation of this study is twofold. On the one hand, a deep model trained by a supervised manner can extract representative high-level features instead of human-design shallow features. And on the other hand, to model the correspondence between local regions and histopathological characteristics, a MIML method is applied for annotation in which a skin biopsy histopathological image can be regarded as a set of local regions and histopathological characteristics can be regarded as multiple labels attached to a whole image. Though the ground truth correspondence between them may be unknown, a MIML model can also work well as an annotator for a test skin biopsy image.

Deep learning models have been studied in many literatures recently [5], whose main idea is to build network with multiple layers to represent original inputs as some high



**Fig. 3** A deep network model

level concepts. In some cases deep learning model can provide an End-to-End model which means no additional human efforts are required when building the model. Figure 3 shows a deep network model.

In Fig. 3, each node is a nonlinear activation function. An edge between a pair of nodes is associated with a weight. When a data sample passes through the network, each element is multiplied by the edge's weight before going into nodes in the next layer. The feature learning of a deep network can be regarded as feature re-expressing, meaning that the transformed features can be restored to original inputs with least information loss. This is the main idea of most deep learning models, e.g. deep belief network [20], stacked auto-encoders [19], deep neural network and convolutional neural network (CNN) [33]. Currently deep learning becomes a hot topic in machine learning research. Ooi et al. [27] proposed a distributed deep learning platform named SINGA that provides a fast implementation of many deep models. Gao et al. [17] proposed a deep learning method for multimedia data retrieval. Their method is scalable and a deep learning hashing algorithm is designed for effective feature representation. Li et al. [22] proposed a method for joint embeddings of shapes and images based on a convolutional neural network (CNN), which has been proved to be a powerful model for image classification and annotation [26, 41].

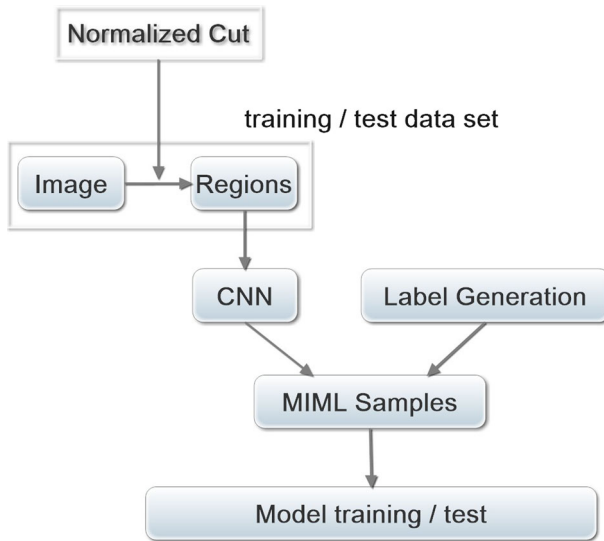
The remainder of this paper is organized as follows. In Section 2 we present the main methods, including the self-adaptive region cutting method for skin biopsy images, convolutional neural network and MIML model. In Section 3 we present the settings of evaluation of the proposed method, and report the evaluation results on a real world clinical histopathological image dataset and a benchmark dataset. In Section 4 we discuss some important issues. And finally we conclude the paper in Section 5.

## 2 Methods

### 2.1 Multiple-instance multiple-label learning

Since most histopathological features are only associated with local regions, as presented in the previous section, feature extraction methods designed for the whole image cannot precisely capture features to build an effective annotation model. We adopt multiple-instance multiple-label learning (MIML) as the main annotation model. Figure 4 sketches the main steps of the proposed model.

There are two key aspects when expressing an image as a MIML sample. The first is multiple-instance decomposition. We proposed a method based on a famous image segmentation algorithm, i.e. normalized cut [29], to generate local regions with a self-adaptive strategy. It is worth noting that normalized cut performs a binary segmentation each time, and the number of generated regions has to be set before running the algorithm. We introduce a strategy to guide the cutting procedure, in which two issues are considered. The first is the lower bound of region size measured by pixels, denoted as  $p$ . According to diagnosis experiences, too small regions may not have significant medical meaning or indicate a histopathological feature. The second is how to choose a region to be further segmented, since normalized cut performs binary segmentation in each round. From a general perspective, it is preferable to choose a large region for further cutting. However, a small region containing complicated pixels may also require further cutting. Hence variances of all pixels in all candidate regions are calculated and the largest one is chosen for further cutting. Algorithm 1 shows the above steps in detail.



**Fig. 4** Main steps of the proposed model

---

**Algorithm 1** Self-adaptive region cutting

---

**Require:**

- skin biopsy image  $I$  in RGB color space;
- lower bound of region size  $p$  measured by pixels;
- maximal number of generated regions  $R$ ;

**Ensure:**

label matrix  $M_I$ : pixels belonging to the same region are marked as the same label;

- 1:  $r = 1$  //number of currently generated regions
  - 2: initialize priority queue  $Q$ ,  $L = 1$
  - 3:  $Enqueue(Q, I, var(I))$ ;
  - 4: **for**  $r < R$  **do**
  - 5:     **if**  $Q$  is empty **then**
  - 6:         return  $M_I$ ;
  - 7:     **end if**
  - 8:      $t = Dequeue(Q)$
  - 9:     **if**  $pixel(t) < p$  **then**
  - 10:         continue;
  - 11:     **end if**
  - 12:      $(t_1, t_2) = NCut(t)$
  - 13:      $Enqueue(Q, t_1, var(t_1))$ ;
  - 14:      $Enqueue(Q, t_2, var(t_2))$ ;
  - 15:      $mark(t_1, M_I, L++)$ ;
  - 16:      $mark(t_2, M_I, L++)$ ;
  - 17: **end for**
  - 18: **return**  $M_I$ ;
- 

In Algorithm 1, a priority queue  $Q$  stores all of the generated regions ordered by their variances. Initially the whole image is enqueued and each time a region is drawn from  $Q$

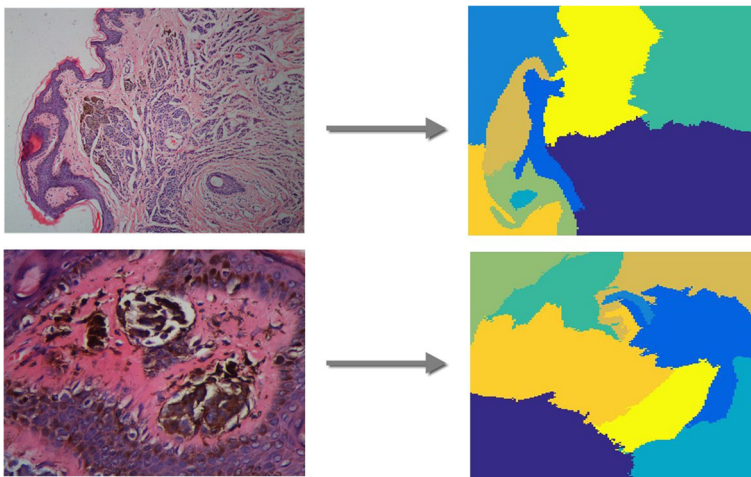
and then two newly generated regions are enqueued. *NCut* is an implementation of the normalized cut method proposed by Shi and Malik [29], which performs a binary segmentation in each round. Line 4 and 9 perform cutting control with preset parameters. With Algorithm 1, an image can be expressed as a set of disjoint regions, which can be regarded as a multiple-instance sample. Figure 5 shows the cutting result of a sample skin histopathological image.

Multiple-label is another important aspect when constructing MIML samples. Since an image is associated with a paragraph of histopathological diagnosis in plain text containing several standard terms, an image can be viewed as a multiple-label sample whose labels are the standard terms appeared in the diagnosis text. After removing the linked words and high-frequency words, a simple text-match method is applied to the plain text associated with each image to find the existence of standard terms. We record the match results in a binary vector. Then we got MIML samples and they can be fed to the proposed model for feature extraction. Figure 6 gives an example.

## 2.2 Region-based unsupervised feature learning

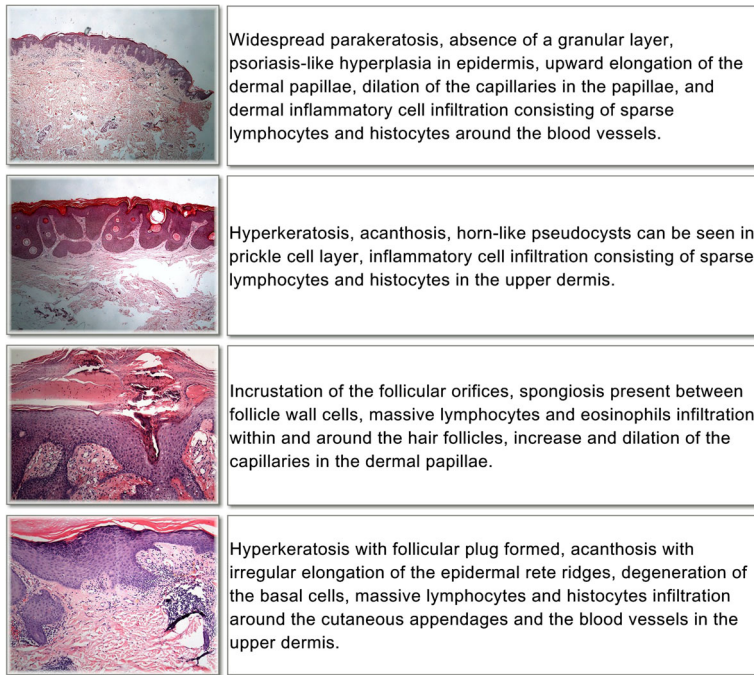
Since several feature representation methods have been reported successful in literatures, it is worth developing a general method with less human efforts, whose essential idea is almost the same as that proposed in [7]. However, different from previous study, our model does not make weighted combination of existing features, but considers all pixels in a region as input instead, and let the model learn potential concepts in a supervised manner. There are some considerable benefits. First of all, our model accepts pixels as input, instead of human designed features, which simulates the structure and mechanism of human brain. Since it directly processes the pixels, least information loss can be achieved if the model is well trained. Secondly, by applying nonlinear transformation in each node, original inputs are encoded into high level features, which provides a powerful way to express arbitrary complex functions and abstract concepts [5].

We propose to use a convolutional neural network (CNN) as the main model. It performs unsupervised feature learning such that the network outputs should be equal to the inputs after processing by hidden layers. The traditional CNN model has two types of layers. The first type is convolutional layer (C-layer) and the second is sub-sampling layer (S-layer).



**Fig. 5** Cutting results of two skin biopsy histopathological images





**Fig. 6** Histopathological images and their corresponding diagnosis in plain text

In a C-layer a  $m * m$  convolutional kernel matrix is sliding over the input image and the convolution operation is performed at each position. Hence an input image with dimension  $p \times q$  would be transformed into a  $(p - m + 1) \times (q - m + 1)$  matrix. In a S-layer,  $s \times s$  pixels are summarized as a single value by applying a average weighted summation, adding a bias and then a nonlinear transformation (Relu function in this work). For the output of the final C-layer or S-layer, some full-connected layers are attached in which each node performs a weighted softmax function to generate a vectorial output. The model parameters  $w$  are weights of connections between pairs of nodes belonging to the layers next to each other. The training method for determining optimal  $w$  is almost the same as the famous BP algorithm [34]. If the numbers of layers and channels are large, however, the model training is time consuming. The detailed structure of the CNN model adopted in this work is shown in Table 1.

The column **Layer Type** indicates the type of each layer. The input layer is a direct-pass pipe between the input image and the network. The adaptive sub-sampling layer is for image scaling, as mentioned above. It calculates the ratio between the size of a region and the model input. And then adaptively set the sub-sampling size, i.e. the size of rectangle for sub-sampling. The column **Channels** indicates the how many kernels are used in the layer for convolution or sub-sampling. Note that in a certain sub-sampling layer, the kernels that applied to feature maps belonging to different channels are identical. The column **Kernel** stands for the size of the kernel used in each layer. The size of a kernel is identical for each channel. In line 7, the C-layer-5 has a full connection of all features maps of S-layer-4.

As an encoding network, CNN requires the inputs having the same sizes since the number of nodes in input layer should be determined before training. But the regions generated through normalized cut are of different sizes. We add an adaptive sub-sampling layer before the first C-layer to perform image scaling, as shown in Table 1. The goal is to scale the input

**Table 1** The structure of CNN model

No.	Layer Type	Kernel	Channel
1	input	N/A	N/A
2	adaptive sub-sampling	N/A	N/A
3	C-layer-1	$35 \times 25$ - stripe 1	64
4	S-layer-2	$4 \times 4$ - stripe 2	64
5	C-layer-3	$35 \times 25$ - stripe 1	128
6	S-layer-4	$4 \times 4$ - stripe 2, pad [0101]	128
7	C-layer-5	$10 \times 8$ - stripe 1 - stripe 2	512
9	S-layer-6	$4 \times 3$ - stripe 2	512
10	C-layer-7	$6 \times 5$ - stripe 1	512
11	C-layer-8	$1 \times 1$ - full	512
12	C-layer-9	$1 \times 1$ - full	1204
13	output	$1 \times 1$	N/A

region to a preset size  $200 \times 150$ . Since the regions are not rectangles, for conveniently processing, we use Melkman algorithm to find its convex hull before scaling, i.e. minimum bounding box with padding pixels.

In most study of deep learning, there is an additional supervised weights fine-tuned which will lead to better performance of the target encoding model [12, 20, 21]. The fine-tuned is guided by the concept labels of the training data samples. To do this, a softmax [23] or linear SVM [31] is added at the top of the deep model to classify or predict. The loss between the model output and the true value is measured and passed backward according to the gradient of the network. A prerequisite is that function at the top layer should be differentiable. Current study showed that either softmax or linear SVM meets the requirement. In these cases supervised fine-tuned can be performed. However, in our study the learning framework is different from those methods mentioned above. The data sample is multiple-instance, meaning that there are one or more instances in a data sample, leading to the so-called label ambiguity [37, 42]. As a result, it makes it difficult to measure the loss between model output and the ground truth label. Hence in this study, we do not combine the encoding model and the MIML model together and only perform single-instance training for the CNN model. Algorithm 2 summaries the main steps of training the encoding network.

---

**Algorithm 2** CNN training for skin biopsy image region encoding
 

---

**Require:**

$X = \{x_1, x_2, \dots, x_n\}$ , regions generated by normalized cut

$Y = \{y_1, y_2, \dots, y_n\}$ , associated labels

**Ensure:**

$W$ : trained model parameters

1:  $W = \mathbf{initCNN}()$

2: **for**  $i = 1$  to  $n$  **do**

3:      $xs = \mathbf{mbb}(x)$  //minimum bound box algorithm

4:     adaptive sub-sampling  $xs \rightarrow x'$

5:      $f = \mathbf{evalCNN}(w, x')$

6:      $y' = \mathbf{softmax}(f)$

7:      $W = \mathbf{backprop}(x', y', y, W)$

8: **end for**

9: **return**  $W$

---

In Algorithm 2, line 1 initializes the network parameters  $W$  through the procedure `initCNN`. The procedure `mbb` Line 3 processes the minimum bound box algorithm to regularize a region. Line 4 performs an adaptive sub-sampling to scale the region and save the result in matrix  $xs$ . Line 5 to 7 performs CNN output evaluation and error back-propagation. We use a softmax function in Line 6 to generate the evaluation result from the features encoded by CNN. The label set  $Y$  has to be constructed manually since the label ambiguity of MIML sample cannot be back-propagated through the CNN.

### 2.3 MIML annotation model

After feature extraction of regions, MIML samples can be expressed as sets of feature vectors. Though there are a lot of MIML classifiers which have been proposed and reported successfully in various tasks, we adopt a current proposed MIML annotation model called S-MIMLGP, proposed in our previous work [38]. The motivation is twofold. Firstly it is designed for skin biopsy image classification working with some traditional features, e.g. color, texture, sub-structure. It has been proved to be effective in the classification of certain histopathological characteristics. Secondly, the method works under a probabilistic foundation and it is able to give the posterior distribution for each annotation terms, which indicates the confidence of annotating a term to an skin image. Probabilistic models are preferable for medical decision support. Another famous MIML classification model, MIMLBoost is also implemented for comparison.

## 3 Evaluations

### 3.1 Data set and settings

We evaluate the proposed model on two datasets. The first is a clinical dataset from the department of dermatology of a large hospital (denoted as  $D1$ ) and the second is a skin tissue image dataset (denoted as  $D2$ ). The dataset  $D1$  contains skin biopsy images and their diagnosis descriptions in plain text. There are 12600 images in  $D1$ , each of which is taken from lesion tissues of a patient and imaged under an electronic microscope. The image size is  $2048 \times 1536 \times 24b$ . We follow the above-mentioned processing method to transform each image into MIML sample. In dataset  $D1$ , there are 15 standard annotation terms to be taken into account. Table 2 shows the details of the histopathological features as well as their occurrence rates in  $D1$ .

Dataset  $D2$  is a skin tissue image dataset which was firstly analyzed by Angel et al. in [2]. It has 2828 images belonging to 4 different skin tissues. The image size of  $D2$  is  $720 \times 480 \times 24b$ . Figure 7 shows sample images of the two datasets.

The parameters of the adaptive region cutting method are set as following: the maximal number of regions  $R = 13$  and the lower bound of total pixels in a region  $p = 1800$ . These two parameters are determined by medical experience as our previous work [37] did. We use S-MIMLGP, MIMLRBF and MIMLboost as the models for annotation.

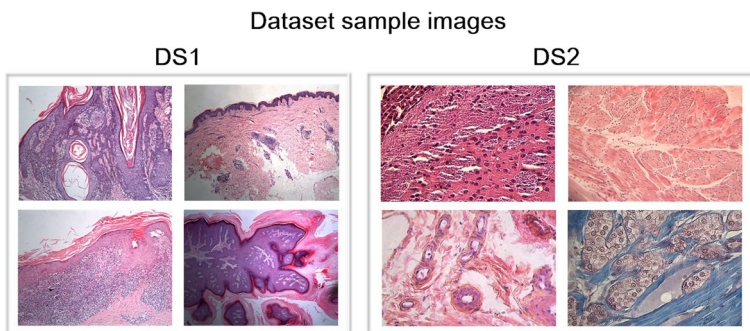
To show the effective of the proposed model, 4 previous successful methods with different intuitions are implemented for comparison. For brevity, we denote them as  $F_1$  to  $F_4$ . Method  $F_1$  and  $F_2$  come from our recent work.  $F_1$  is a multiple-instance learning method [35], in which the annotation problem is divided into several binary classification problems according to the annotation terms.  $F_1$  utilizes Citation-KNN as the main model and region-based wavelet transformation algorithm to extract features.  $F_2$  makes use of a MIML

**Table 2** 15 annotation terms and their occurrence rates in *D1*

No.	Name	Rate
T1	retraction space	28.65%
T2	papillomatosis	22.71%
T3	follicular plug	1.8%
T4	hypergranulosis	32.15%
T5	horn cyst	4.14%
T6	basal cell liquefaction degeneration	6.48%
T7	thin prickle cell layer	2.61%
T8	infiltration of lymphocytes	9.12%
T9	hyperpigmentation of Basal cell layer	36.99%
T10	nevocytic nests	18.56%
T11	Munro microabscess	7.72%
T12	acanthosis	19.05%
T13	absent granular cell layer	23.24%
T14	parakeratosis	6.81%
T15	hyperkeratosis	11.30%

learning model based on sparse Bayesian learning [37], it combined several probabilistic MIML model with a relevant vector machine (RVM) [32] to find the optimal combination weights.  $F_3$  is a bag-of-features (BOF) data sample representation method with kernel learner [8]. It constructed a codebook composing of image patches as basic features and expressed each training or test example as a histogram based on the codebook. Method  $F_3$  had been used to classify biopsy tissues images. Method  $F_4$  comes from the famous DD-SVM [11], which applied a clustering-based algorithm to cluster similar blocks together and thus form regions. It represented an image as a multiple-instance sample and used DD-SVM to classify it. Table 3 shows the references and the corresponding data representations of these methods.

In the third column of Table 3, the data representation methods are provided.  $R_0$  stands for the CNN based feature extraction method proposed in this work.  $R_1$  stands for the combining feature of LUV color space and wavelet transformation.  $R_2$  stands for the Bag-Of-Features (BOF) feature representation.  $R_3$  is a multiple-instance representation of a

**Fig. 7** Sample images of the two datasets

**Table 3** Data representation and their consistent methods

Method	Reference	Data representation
<i>CNN</i> : Our method	this work	$R_0$
$F_1$ : Citation	Zhang et al. [35]	$R_1$
$F_2$ : SBL-MIML	Zhang et al. [37]	$R_1$
$F_3$ : BOF	Caicedo et al. [8]	$R_2$
$F_4$ : DD-SVM	Chen et al. [11]	$R_3$

histopathological image, but it regards pixel clusters as regions, which may not be contiguous. Hence it may not be able to obtain visual disjoint regions. To construct the evaluation data set, we randomly divide the whole data set into training part and test part at size ratio 1 : 4. Then apply the feature extraction methods listed in Table 3 to construct the data set.

### 3.2 Evaluation criteria

We use two well-known multiple-label learning criteria [28] to measure the performance of the methods. The first criteria is accuracy, denoted as *acc*, measuring the general performance of an annotation model. It does not consider the relation between annotation terms. However, it only calculates the mean accuracy of each label evaluated by a zero-one loss function. The second criteria is *hloss* which measures the number of misclassified label pairs. We also use the false positive rate (FPR) and false negative rate (FNR) of each label to measure the performance of the proposed model. FPR measures the ratio of wrong annotation per sample per label. FNR measures the ratio of missing annotation per sample per label by the model.

### 3.3 Evaluation results

We report the overall accuracy of the proposed method and the methods for comparison, as shown in Table 4. Since the dataset is multiple-label, the overall accuracy is the mean value of all accurate rates for all terms. The best result in the table has been highlighted. The proposed method achieves the best result among all evaluated methods. The column **Method** shows the analysis methods for evaluation. Note that some of these methods are composed of both feature extraction and annotation parts. For a comprehensive evaluation, we record the performance of the original implementation which has been marked with a superscript star. And we also use the feature extraction methods in our MIML annotation model and record the performance, e.g. line 4 and 5 of Table 4.

In line 9, the method  $F_3$  only works with a multiple-class SVM classifier. This is because it is a single-instance learning method. The S-MIMLGP and MIMLBoost cannot be applied to single-instance samples generated by  $F_3$ .

$D2$  is in fact a single-label dataset because every image is associated with only one of the four categories. However, current study on MIML indicated that multiple-label classifiers also work well on single-label datasets. Note that in this case the criteria *hloss* is of no use. Figure 8 shows the overall accuracy of  $D2$ . It can be seen that the proposed CNN feature representation method outperforms the other methods.

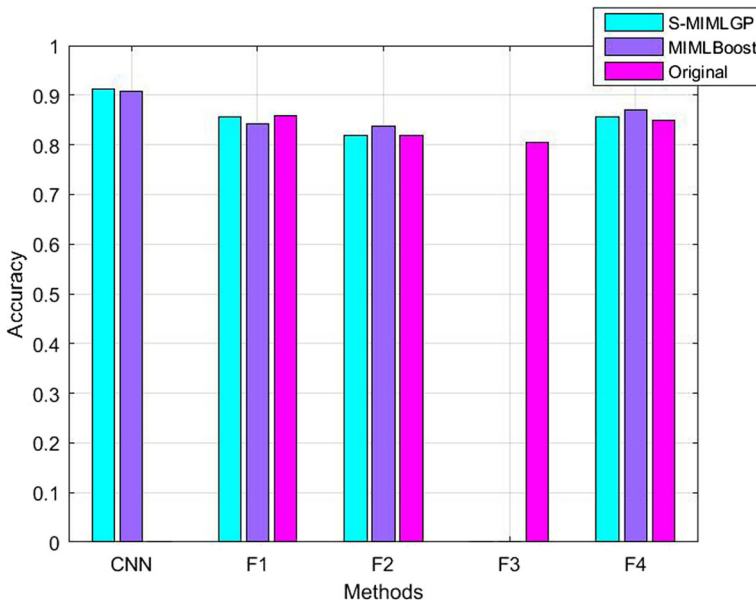
The FPR and FNR are important for the evaluation of effectiveness of the proposed model. For brevity, we report the FPRs and FNRs of method *CNN*,  $F_2$  and  $F_4$  on  $D1$ , which

**Table 4** Overall accuracy of  $D1$  ( $acc$ : %)

No.	Method	Annotation model	acc	hloss
1	<i>CNN</i>	S-MIMLGP	<b>81.75</b>	<b>0.221</b>
2	<i>CNN</i>	MIMLBoost	79.55	0.241
3	$F_1$	Citation*	74.02	0.276
4	$F_1$	S-MIMLGP	75.91	0.289
5	$F_1$	MIMLBoost	77.12	0.250
6	$F_2$	SBL-MIML*	78.88	0.257
7	$F_2$	S-MIMLGP	75.91	0.289
8	$F_2$	MIMLBoost	78.40	0.247
9	$F_3$	SVM*	67.33	0.341
10	$F_4$	DD-SVM*	69.56	0.313
11	$F_4$	S-MIMLGP*	72.05	0.309
12	$F_4$	MIMLBoost*	72.37	0.320

stand for different feature representation methods and learning models. Table 5 shows the FPR and FNR of each annotation term.

It can be seen that our method also achieves the best results in a large body of annotation terms. But the model performs poorly in some terms, such as T2, T3 and T10. We owe this to the unbalanced occurrence of annotation terms in the data set. Note that we make natural segmentation on the whole data set, meaning that we do not guarantee that each label in either training set or test set is distributed equally as that in the whole data set. The model may tend to annotate the frequent terms according to the training set, which



**Fig. 8** the overall accuracy of  $D2$

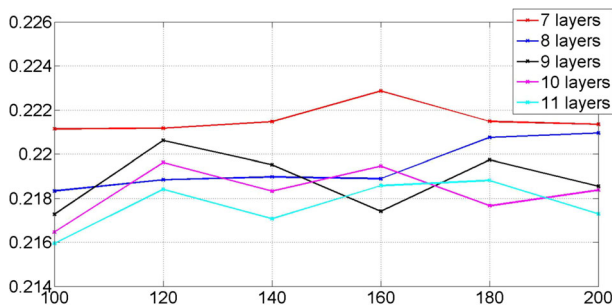
**Table 5** FPRs and FNRs of three methods on  $D1$  (%)

CNN				$F_1 - Original$				$F_4 - Original$			
Terms	FPR	FNR	Total	Terms	FPR	FNR	Total	Terms	FPR	FNR	Total
T1	8.04	10.19	18.23	T1	7.87	11.35	19.21	T1	7.86	12.41	20.27
T2	10.05	10.57	20.62	T2	13.21	13.52	26.64	T2	13.42	10.99	24.42
T3	9.81	12.02	21.83	T3	11.48	13.02	24.51	T3	9.12	13.05	22.17
T4	5.13	5.59	10.72	T4	7.99	8.24	16.23	T4	5.22	6.91	12.13
T5	8.25	8.61	16.86	T5	10.95	7.44	18.40	T5	10.88	6.58	17.46
T6	9.78	7.65	17.43	T6	8.89	12.10	20.99	T6	11.88	11.32	23.20
T7	7.99	10.53	18.52	T7	8.46	12.16	20.62	T7	8.45	12.24	20.69
T8	6.21	7.34	13.55	T8	7.15	7.21	14.37	T8	5.79	10.56	16.34
T9	9.02	9.42	18.44	T9	9.71	13.52	23.23	T9	10.76	10.18	20.94
T10	11.50	12.06	23.56	T10	14.15	13.95	28.10	T10	11.71	12.22	23.93
T11	13.18	7.03	20.21	T11	6.68	14.78	21.46	T11	10.61	11.95	22.57
T12	5.51	6.00	11.51	T12	7.20	8.03	15.23	T12	11.06	5.40	16.45
T13	6.91	10.54	17.45	T13	7.56	12.14	19.70	T13	12.47	8.05	20.52
T14	6.40	8.81	15.21	T14	7.85	9.45	17.30	T14	5.97	13.16	19.13
T15	5.85	6.71	12.56	T15	10.43	7.46	17.89	T15	7.13	7.81	14.94

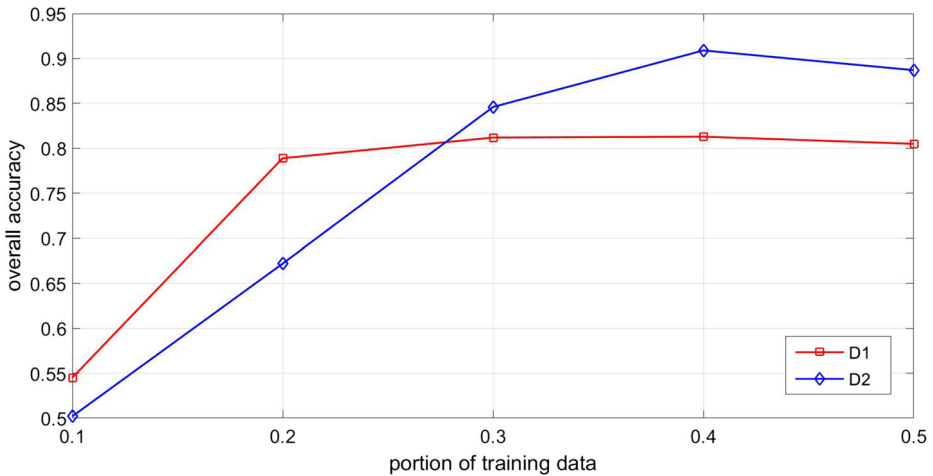
is called model bias in machine literatures [6]. Another reason may be that some of these histopathological features do not have explicit expression in digital images. Our model is ineffective to annotate these kind of features.

The dimension of the output of CNN and the number of hidden layers may significantly affect the model performance. The output vector of CNN is the input of MIML model. We vary the number of hidden layers and the dimension of output layer of the network, and record the model mean loss of all annotation terms. See Fig. 9.

In Fig. 9, the x axis is the number of dimension of the encoding model output, and the y axis is the accuracy value. We can see that models with more hidden layers achieve better performance. While the dimensions of encoding model outputs have little effects on the model performance. It motivates us to use networks with more hidden layers and low-dimension outputs.



**Fig. 9** Overall loss of different number of hidden layers and output dimension



**Fig. 10** Overall loss of different number of hidden layers and output dimension

Finally, we report how the trained CNN network affects the MIML model, i.e. whether the larger size of manually labeled region dataset means better performance of the target model. We train the CNN model with portions of training data from 10% to 50% by step 10% and model the overall performance. And we keep the settings of MIML annotator the same so as to highlight the quality of feature extraction deep network. Figure 10 shows the relationship between the MIML model performance and the labeled dataset size for CNN training.

## 4 Discussions

We give some discussions related to this study. In the first place, we would like to discuss the relationship between region cutting and feature extraction through deep model. A natural question is why we do not directly feed a whole histopathological image to the CNN model. The reason lies in the high computational cost. The network is not able to process an image of large size. Hence we cut an image into regions, and extract feature per region, so as to cut down the computational cost. There may be information loss in the cutting process. But the cutting method generates visual disjoint regions which are supported by medical knowledge and experience.

The second problem is whether we can perform supervised fine-tune of the model weights. It is a challenging problem, since the encoding network can only accept instance (region). In order to launch a BP like algorithm to adjust the network weights, the directly connection between instances and labels have to be established. However, due to the label ambiguity of MIML learning, the relation cannot be directly expressed. Hence we use manually labeled region for the model training and apply the trained network for feature extraction. Meanwhile, there are studies that establish the connection between multiple-instance samples and target labels. In He et al.'s work [18], a likelihood function establishes the connection between instances in a bag and its labels by introducing a vector of hidden variables. Thus the posterior distribution of labels given a bag can be derived by integrating out all hidden variables. In our model this kind of relation seems not easy to establish.



Since the number of instances in each bag may be different, it is not possible to input them into the model at the same time. To encode a data sample of complicated structure needs to be further studied.

The third problem is how different multiple-instance assumption affects the model performance. As proposed in [16], difference of background information may lead to different multiple-instance assumptions, which in fact define the relation between instances within a bag and the corresponding labels. In our model, we use the original multiple-instance assumption which was proposed by Dietterich et al. in 1997 [13]. It assumes that in binary classification case, a bag is labeled positive if and only if it has a positive instance, and negative otherwise. The assumption is roughly suitable for our skin histopathological image annotation problem. If an image contains a region that should be annotated to a term, the whole image should be annotated to this term as well. Though this assumption does not take the relationship of instances and labels into consideration, the model based on it can achieve good performance even if the problem domain indicates much complicated assumptions [39]. In the literatures of multiple-instance learning, there have been reported powerful models to support different assumption [42], which require additional computational costs to model the quantity or structure information of instances within a bag. We place our study under the standard multiple-instance assumption to simplify the annotation model, so as to focus on the feature representation by the CNN model.

## 5 Conclusions

In this paper we proposed a feature representation method based on deep learning for skin biopsy histopathological image annotation. Different from previous methods that adopt human designed features, we proposed to learn features from low level pixels in a supervised manner. CNN is used as a feature learning model. The proposed method learns abstract features through multiple-layer weighted combination and nonlinear transformation of the original features. Then a supervised MIML learning model is placed at the top of the deep model to generate annotation results. Evaluation results on a real clinical data set and a famous benchmark dataset show the proposed method are superior to recent methods. Though the feature extraction method is region-based and requires manually labeled regions, it can achieve better features than the original ones. The model simulates the structure of human brain and attempts to be trained and work like what the human brain does.

There are some problems yet to be solved. One problem is that the proposed method only performs region-based supervised learning. Due to the label ambiguity of multiple-instance, the loss of model output cannot be propagated through the network, which leads to the failure of supervised fine-tune of the network weights. Another problem to be solved is the design of multiple-instance data sample CNN. The essence of this problem is the question whether we can design a CNN model to encode a multiple-instance sample, instead of an instance (region). These two problems will be studied in our future work.

**Acknowledgements** This work is supported by National Natural Science Foundation of China (No. 61502106, 81373883, 81573827), Natural Science Foundation of Guangdong Province (No. 2016A030310340), the College Student Career and Innovation Training Plan Project of Guangdong Province (xj201511845018, yj201511845038, yj201611845074, yj201611845075, yj201611845366), the Special Fund of Cultivation of Technology Innovation for University Students (pdjh2016b0150), the 2015 Research Project of Guangdong Education Evaluation Association (No. G-11) and Fujian Major Project of Regional Industry (No. 2014H4015).

## Compliance with Ethical Standards

**Competing Interests** The authors declare that there is no conflict of interests regarding the publication of this article.

## References

1. Ali R, Gunduz-Demir C, Szilágyi T, Durkee B, Graves EE (2013) Semi-automatic segmentation of subcutaneous tumours from micro-computed tomography images. *Phys Med Biol* 58(22): 8007
2. Angel CR, Juan CC, Fabio AG (2011) Visual pattern mining in histology image collections using bag of features. *Journal Artificial Intelligence in Medicine* 52(2011):91–106
3. Baldi A, Murace R, Dragonetti E, Manganaro M, Bizzi S (2014) Automated content-based image retrieval: Application on dermoscopic images of pigmented skin lesions. In: *Skin Cancer*. Springer, pp 523–528
4. Bengio Y (2009) Learning deep architectures for ai. *Found Trends Mach Learn* 2(1):1–127
5. Bengio Y (2013) Deep learning of representations: Looking forward. In: *Proceedings of the 1st International Conference on Statistical Language and Speech Processing*, Springer-Verlag, Berlin, Heidelberg, SLSP'13, pp 1–37
6. Bishop CM (2006) *Pattern recognition and machine learning (information science and statistics)*. Springer-Verlag, Secaucus
7. Bunte K, Biehl M, Jonkman MF, Petkov N (2011) Learning effective color features for content based image retrieval in dermatology. *Pattern Recogn* 44(9):1892–1902
8. Caicedo JC, Cruz-Roa A, González FA (2009) Histopathology image classification using bag of features and kernel functions. In: Combi C, Shahar Y, Abu-Hanna A (eds) *Proceedings of AIME, Lecture Notes in Computer Science*, vol 5651, pp 126–135
9. Cerroni L, Argenyi Z, Cerio R, Facchetti F, Kittler H, Kutzner H, Requena L, Sanguenza OP, Smoller B, Wechsler J, Kerl H (2010) Influence of evaluation of clinical pictures on the histopathologic diagnosis of inflammatory skin disorders. *J Am Acad Dermatol* 63(4):647–52
10. Chen J, Zhao F, Cao H (2013) Knowledge acquisition from generalized experts oriented to product innovation. In: *2013 6th international conference on Information management, innovation management and industrial engineering (ICIII)*, vol 2. IEEE, pp 546–549
11. Chen Y, Wang JZ (2004) Image categorization by learning and reasoning with regions. *J Mach Learn Res* 5:913–939
12. Cho Y (2012) *Kernel methods for deep learning*. PhD thesis, La Jolla, CA, USA, aAI3513249
13. Dietterich TG, Lathrop RH, Lozano-Pérez T (1997) Solving the multiple instance problem with axis-parallel rectangles. *Artif Intell* 89(1-2):31–71
14. Dong R (2011) Feature grouping technique to relax sample support requirement for sparse linear feature extraction. In: *2011 seventh international conference on Natural computation (ICNC)*, vol 3. IEEE, pp 1654–1657
15. Ferrara G, Argenyi Z, Argenziano G, Cerio R, Cerroni L, Di Blasi A, Feudale EA, Giorgio CM, Massone C, Nappi O, Tomasini C, Urso C, Zalaudek I, Kittler H, Soyer HP (2009) The influence of clinical information in the histopathologic diagnosis of melanocytic skin neoplasms. *PLoS One* 4(4): e5375
16. Foulds J, Frank E (2010) A review of multi-instance learning assumptions. *Knowl Eng Rev* 25(1): 1–25
17. Gao L, Song J, Zou F, Zhang D, Shao J (2015) Scalable multimedia retrieval by deep learning hashing with relative similarity learning. In: *Proceedings of the 23rd ACM International Conference on Multimedia*, ACM, New York, NY, USA, MM '15, pp 903–906
18. He J, Gu H, Wang Z (2012) Bayesian multi-instance multi-label learning using gaussian process prior. *Mach Learn* 88(1-2):273–295
19. Hinton GE, Salakhutdinov RR (2006) Reducing the dimensionality of data with neural networks. *Science* 313(5786):504–507

20. Hinton GE, Osindero S, Teh YW (2006) A fast learning algorithm for deep belief nets. *Neural Comput* 18(7):1527–54
21. Huang PS, He X, Gao J, Deng L, Acero A, Heck L (2013) Learning deep structured semantic models for web search using clickthrough data. In: *Proceedings of the 22nd ACM International Conference on Conference on Information & Knowledge Management*, ACM, New York, NY, USA, CIKM '13, pp 2333–2338
22. Li Y, Su H, Qi CR, Fish N, Cohen-Or D, Guibas LJ (2015) Joint embeddings of shapes and images via cnn image purification. *ACM Trans Graph* 34(6):234:1–234:12
23. Lopes N, Ribeiro B (2014) Towards adaptive learning with improved convergence of deep belief networks on graphics processing units. *Pattern Recogn* 47(1):114–127
24. Malik MSA, Sulaiman S (2014) DbA's perspective on use of information visualization in electronic health records
25. Marrugo AG, Millan MS (2011) Retinal image analysis: preprocessing and feature extraction. *Journal of Physics: Conference Series*, IOP Publishing 274:012039
26. Murthy VN, Maji S, Manmatha R (2015) Automatic image annotation using deep learning representations. In: *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*, ACM, New York, NY, USA, ICMR '15, pp 603–606
27. Ooi BC, Tan KL, Wang S, Wang W, Cai Q, Chen G, Gao J, Luo Z, Tung AK, Wang Y, Xie Z, Zhang M, Zheng K (2015) Singa: A distributed deep learning platform. In: *Proceedings of the 23rd ACM International Conference on Multimedia*, ACM, New York, NY, USA, MM '15, pp 685–688
28. Schapire RE, Singer Y (2000) Boostexter: a boosting-based system for text categorization. *Mach Learn* 39(2-3):135–168
29. Shi J, Malik J (2000) Normalized cuts and image segmentation. *IEEE Trans Pattern Anal Mach Intell* 22(8):888–905
30. Stoean C, Stoean R, Sandita A, Mesina C, Gruia CL, Ciobanu D (2015) Evolutionary search for an accurate contour segmentation in histopathological images. In: *Proceedings of the Companion Publication of the 2015 Annual Conference on Genetic and Evolutionary Computation*, ACM, New York, NY, USA, GECCO Companion '15, pp 1491–1492
31. Tang Y (2013) Deep learning using support vector machines. CoRR arXiv:1306.0239
32. Tipping ME (2001) Sparse bayesian learning and the relevance vector machine. *J Mach Learn Res* 1:211–244
33. Trinh HP, Duranton M, Paindavoine M (2015) Efficient data encoding for convolutional neural network application. *ACM Trans Archit Code Optim* 11(4):49:1–49:21
34. Vincent P, Larochelle H, Lajoie I, Bengio Y, Manzagol PA (2010) Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion 11:3371–3408
35. Zhang G, Shu X, Liang Z, Liang Y, Chen S, Yin J (2012a) Multi-instance learning for skin biopsy image features recognition, Philadelphia, PA, United states, pp 83–88
36. Zhang G, Yin J, Li Z, Liang Z, Fu W (2012b) Deep learning for acupuncture point selection patterns based on veteran doctor experience of chinese medicine. In: *Proceedings of the 2012 IEEE International Conference on Bioinformatics and Biomedicine Workshops (BIBMW)*, IEEE Computer Society, Washington, DC, USA, BIBMW '12, pp 396–401
37. Zhang G, Yin J, Li Z, Su X, Li G, Zhang H (2013) Automated skin biopsy histopathological image annotation using multi-instance representation and learning. *BMC Med Genet* 6(3):1–14
38. Zhang G, Yin J, Su XY, Huang YJ, Lao YR, Liang ZH, Ou SX, Zhang HL (2014) Augmenting multi-instance multilabel learning with sparse bayesian models for skin biopsy image analysis. *BioMed Research International* 2014(Article ID 305629):13 pages
39. Zhang ML (2009) Generalized multi-instance learning: Problems, algorithms and data sets
40. Zhang S, Huang Q, Hua G, Jiang S, Gao W, Tian Q (2010) Building contextual visual vocabulary for large-scale image applications. In: *Proceedings of the international conference on Multimedia*. ACM, pp 501–510
41. Zhong R, Tezuka T (2014) Parametric learning of deep convolutional neural network. In: *Proceedings of the 19th International Database Engineering & Applications Symposium*, ACM, New York, NY, USA, IDEAS '15, pp 226–227
42. Zhou ZH, Zhang ML, Huang SJ, Li YF (2012) Multi-instance multi-label learning. *Artif Intell* 176(1):2291–2320



**Gang Zhang** born in 1979, member of China Computer Federation (CCF). He received his doctor's degree in Computer Software and Theory from SUN YAT-SEN University, China. Now he is a lecturer in School of Automation, Guangdong University of Technology. His current research interests include data mining, machine learning, and its applications to bioinformatics and Traditional Chinese Medicine.



**Ching-Hsien Robert Hsu** Professor Ching-Hsien (Robert) Hsu is a professor in department of computer science and information engineering at Chung Hua University, Taiwan; and distinguished chair professor at Tianjin University of Technology, China. His research includes high performance computing, cloud computing, parallel and distributed systems, big data analytics, ubiquitous/pervasive computing and intelligence.



**Huadong Lai** born in 1996, undergraduate of School of Automation, Guangdong University of Technology, China. His current research interests include deep learning and its application on histopathological image analysis.



**Xianghan Zheng** is associate professor in the College of Mathematics and Computer Sciences, Fuzhou University, China. He received his MSc of Distributed System (2007) and Ph.D of Information Communication Technology (2011) from University of Agder, Norway. His current research interests include Big Data Processing, Cloud Computing Services and Applications.