IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received September 13, 2020, accepted September 27, 2020, date of publication September 30, 2020, date of current version October 12, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3027982

Multi-Instance Learning Algorithm Based on LSTM for Chinese Painting Image Classification

DAXIANG LI[1,2] AND YUE ZHANG[1]

[1]School of Telecommunication and Information Engineering, Xi'an University of Posts and Telecommunications, Xi'an 710121, China
[2]Ministry of Public Security Key Laboratory of Electronic Information Application Technology for Scene Investigation, Xi'an University of Posts and Telecommunications, Xi'an 710121, China

Corresponding author: Yue Zhang (791846135@qq.com)

This work was supported in part by the Innovation and Entrepreneurship Dual Tutor Foundation of Shaanxi under Grant 2019JM-604, in part by the Postgraduate Co-Cultivation Workstation Construction Project under Grant YJGJ201902, and in part by the Graduate Innovation Fund of Xi'an University of Posts and Telecommunications under Grant CXJJLD2019021.

ABSTRACT Aiming at the problem of weakly supervised learning in traditional Chinese painting image classification, a novel multi-instance learning algorithm based on Long and Short-Term Memory neural network with attention mechanism (ALSTM-MIL) is proposed. Firstly, by using the Pyramid Overlapping Grid Division (POGP), a multi-instance modeling scheme is designed to convert Chinese painting images into multi-instance bag, thereby transforming the problem of Chinese painting image classification into a MIL problem. Secondly, an efficient sequence generator is designed. It selects discriminative instances from the positive bags, construct a discriminative instance set (DIS), and convert multi-instance bags into equal-length ordered sequences. Thirdly, an LSTM network model with an attention mechanism is designed to perform semantic analysis on multi-instance bags to obtain their memory coding features, and then combined with the Softmax classifier to achieve semantic classification of traditional Chinese painting images. Experimental results on the Chinese painting (CP) image set show that the LSTM network built on the visual feature set is feasible, and the performance of the proposed MIL algorithm is also superior to other classification algorithms.

INDEX TERMS Attention mechanism, Chinese painting image classification, discriminative instance set, long and short term memory, multi-instance learning.

I. INTRODUCTION

With the development of the Internet and high-fidelity imaging technology, many art galleries and museums have provided online digital art work viewing services [1], [2]. Among them, Chinese painting, as an important expression of Chinese culture, has attracted more attention and favor from online visitors. As an important part of Chinese traditional culture, the research on the classification of traditional Chinese painting images contributes to better inherit and carry forward traditional culture. Therefore, in order to help the art museum to manage the Chinese painting images efficiently, and to facilitate the visitors to browse, so that the visitors can better understand the connotation of Chinese painting, research on the automatic classification of Chinese painting images has become an important topic in the field of computer vision.

Chinese painting is a painting by a painter dipped in water, ink, and color painting on silk or paper with a brush, and has a unique artistic style. Different from Western painting, Chinese painting focuses on artistic conception. The artist conceives and imagines things by observing things, and outlines with lines, reflecting the subjective grasp of objects and artistic refinement. Western painting focuses on real objects, and usually regards artistic works as true copies of painting goals. Chinese painting can be divided into figures, landscapes, flowers and birds, etc. In terms of painting technique matter, Chinese painting can be divided into Gongbi and Xieyi. The traditional method of Chinese painting classification is to extract and fuse global and local features to the image, and then send it to the classifier. Combining the advantages of contourlet transform and gray-level co-occurrence matrix,

The associate editor coordinating the review of this manuscript and approving it for publication was Jeon Gwanggil.

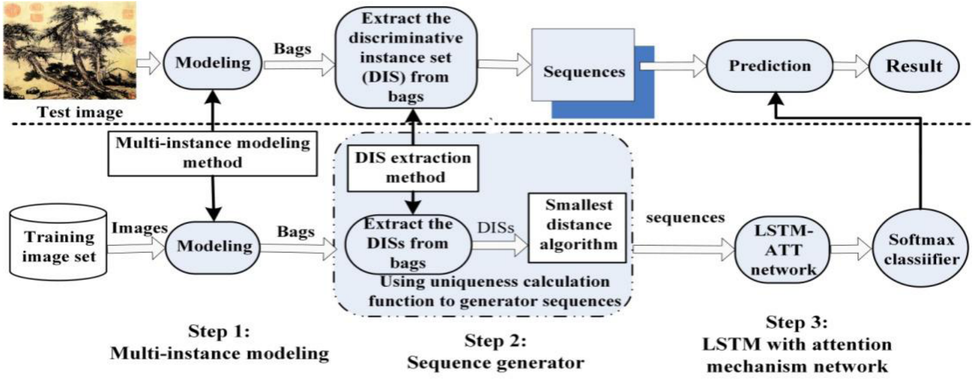

Test image
Modeling
Bags
Extract the discriminative instance set (DIS) from bags
Sequences
Prediction
Result
Multi-instance modeling method
DIS extraction method
Training image set
Images
Modeling
Bags
Extract the DISs from bags
DISs
Smallest distance algorithm
sequences
LSTM-ATT network
Softmax classiifier
Using uniqueness calculation function to generator sequences
Step 1: Multi-instance modeling
Step 2: Sequence generator
Step 3: LSTM with attention mechanism network

FIGURE 1. The framework of the proposed generalized MIL algorithm based on LSTM.

Wang *et al.* [3] proposed a multi-scale and multi-color gamut texture feature extraction method to classify traditional Chinese painting images. In order to find a better image feature representation method in the new signal domain, Sheng [4] proposed a method to obtain the depth information of Chinese painting art in the wavelet domain by using different expressions of artistic styles displayed by image structures of different resolutions and frequency bands. In addition to the traditional support vector machine classifier, with the development of machine learning in recent years, convolutional neural networks (CNN) have been widely used in pattern recognition [5]. As a very good image classification method, CNN has very good classification performance. However, the traditional Chinese painting image has inherent characteristics, that is, the shape, texture, light and darkness and mood of the painted object supplemented by the main line and auxiliary point and surface. Therefore, in the classification, it mainly uses its local painting brush strokes, that is, the combination of lines, points, surfaces, and thickness as features, so under the condition of a limited number of training images, CNN cannot be well applied in the classification of Chinese painting images.

Aiming at the problem of Chinese painting image classification, this article proposes a MIL algorithm based on LSTM. First, multi-instance modeling is used to extract the local brush strokes features of the traditional Chinese painting image, and the image is constructed in the form of bag to characterize and describe the local points and line features of the traditional Chinese painting image; Second, a sequence generator is designed which extract the most discriminating brush strokes features set from the multi-instance bags to reduce redundancy and provide equal-length ordered semantic sequences for the LSTM network. Third, we design a LSTM model with an attention mechanism, and use it to memorize the dependency relationship between the local brush strokes of the image, and finally send it to the Softmax layer for classification. In short, the flowchart of the algorithm proposed in this article is shown in Figure 1, and its main contributions are as follows:

(1) A pseudo-sequence generator is designed to convert the disordered and unequal number of visual feature sets (bags) into equal-length sequence signals.

(2) A novel MIL algorithm based on LSTM with attention mechanism (ALSTM-MIL) for Chinese painting classification is proposed. To the best of our knowledge, this is the first work to introduce LSTM network to the MIL problem, compared with other bag embedding methods, the LSTM is more able to capture the causal relationship among instances (brush strokes features) to achieve semantic parsing of a bag (image).

(3) In order to demonstrate the promising performance of the proposed algorithm on the classification of Chinese Painting (CP) image sets, we conducted sensitivity test and experiments on classification accuracy.

The remaining paper is organized as follows: In Section II, we introduce related works of Chinese painting image classification and MIL algorithm. In Section III, details of the multi-instance modeling are described. Section IV provides the details of sequence generator. In Section V, we proposed LSTM with attention mechanism model. The experimental results are presented in Section VI, Section VII discusses the future works, and Section VIII concludes the paper.

II. RELATD WORK

A. CHINESE PAINTING IMAGE CLASSIFICATION ALGORITHM

Retrieval-oriented image classification has always been a hot research topic in multimedia, and has achieved a lot of effective results. However, there is less research work specifically for Chinese painting image classification. At present, the researches on Chinese painting mainly include the authentication of images, classification based on author's style, painting technique, and content. In the authenticity analysis of artwork, Buchana *et al.* [6] was able to discern low-quality digital representations between the original and the fake by training a comprehensive discriminant function (OTSDF) filter on the coarsely segmented image of difference and location of the original painting. Polatkan *et al.* [7] used hidden Markov trees to model the wavelet coefficients of painting images, and used the parameters on the model as input features to perform supervised machine learning to distinguish the copy from the original work. In the classification of paintings based on the author's style, Sun *et al.* [8] used the Monte Carlo convex hull feature selection model to integrate basic feature descriptors, and then used support vector machines to classify the works of different artists. Li and Wang [9] Designed a general framework for the classification of Chinese paintings, and used a hybrid two-dimensional multi-resolution Markov model (MHMM) to represent the stroke attributes of different artists to achieve classification. Sheng and Jiang [10] proposed to extract local features based on histograms to express the style of Chinese painting images, and designed a fusion scheme of window and entropy balance to optimize the classification results. Chinese painting can be divided into two categories according to the painting technique, namely Xieyi painting and Gongbi painting. In this research work, Gao *et al.* [11] Integrated SIFT feature detectors and edge detection to obtain key areas of Chinese painting, described the visual characteristics of key areas and internal area differences to obtain image features, and used different dimensional features to cascade classification strategies to classify Gongbi and freehand painting. Jiang *et al.* [12] combined Discrete Cosine Transform (DCT) and Convolutional Neural Network (CNN) to propose a classification model. Chinese painting images can be classified from the perspective of content, named ancient tree paintings, figure paintings, flower and bird paintings, Jiangnan water villages, etc. Bao *et al.* [13] classified themes by extracting the semantic information of Chinese painting image scripts and adopting a multi-task joint sparse method.

B. MIL RELATED ALGORITHMS

In image classification, each image has its own region of interest, which may be one or several regions of the image, and the remaining regions are uninteresting regions. This assumption is similar to multi-instance learning [14]. Therefore, multi-instance learning is widely used in image classification tasks [15]. Assuming that any image is a multi-instance bag and the image is divided into multiple sub-blocks, the region of interest of the sub-block is taken as an instance.

After MIL was proposed, many classic MIL algorithms have been presented [16], including axis parallel hyper-rectangles, Citation-kNN, Diverse Density (DD), DD with Expectation Maximization (EM-DD) and MI/mi-SVM algorithms [17]. Due to the training instances of the MIL are bags, i.e. some unordered sets composed of different low-level visual features of images, convert each bag into a single representation vector, and then use standard single instance learning (SIL) methods (i.e. SVM) to solve the MIL problem, is a very effective MIl algorithm. For example, DD-SVM [18], Multi-Instance Learning via Embedded Instance Selection (MILES) [19], LSA-MIL [20], MI-J-SC [21], EC-SVM [22], MILDM [23] and miFV/miVLAD [24], etc. Unfortunately, few existing feature representation methods are effective to describe the sematic of images (bags), so it is difficult to adapt some well-known SIL methods to solve the MIL problems. In recent years, due to the excellent performance of the CNN in image classification problems, some MIL algorithms combined with CNN have also been proposed. For example, Xu *et al.* [25] used deep learning method to obtain the CNN features of images, and then trained MIL classifiers based on these CNN features for medical image analysis; Wu *et al.* [26] focused on the image classification and image annotation, a pre-trained deep learning model was used to predict the label of instance in the MIL framework, and then the labels of all the instances in the bag were synthesized to predict the final label of the bag. He *et al.* [27] in order to address the problem of medical image classification, based on prototype learning and bag feature transformation function, a multi-instance convolutional neural network algorithm is designed. Tang *et al.* [28] in order to find an effective and efficient representation for image classification, traditional MIL methods is extended to explicitly learn more than one multi-instance deep discriminative patterns (MiDDP) in positive class by stochastic gradient decent method, and proposed a novel MIL algorithm named MiDDP. Ajjaji *et al.* [29] in order to solve the problem of remote sensing scene image classification, each scene image is divided into multiple sub-images (four corners and center image) to generate instances, and SequeezeNet is used to extract highly descriptive features from each instance. The feature is sent to a deep neural network to learn the appropriate weights for each instance feature, and the feature is fused using a weighted average method to obtain the final example representation. Kausik *et al.* [30] proposed a deep convolution multi-instance algorithm. First, the instances in the bag are sent to VGGNet to obtain advanced features. Then, the multi-instance pool layer (MIP) is introduced after the eighth layer of the network is fully connected. It consists of two branches, the multi-instance maximum pooling layer and the example-level layer. The two branches generate packet-type MIPs and example-level features, respectively, and map them to the decision layer to generate decisions. In image classification, each image has its own region of interest, which may be one or several regions of the image, and the remaining regions are uninteresting regions. This assumption is similar to multi-instance learning. Therefore, multi-instance learning is widely used in image classification tasks. Assuming that any image is a multi-instance bag and the image is divided into multiple sub-blocks, the region of interest of the sub-block is taken as an instance.

III. MULTI-INSTANCE MODELING

In the multi-instance learning algorithm, each bag contains specific regional features, that is, positive instances. It is not necessary to know the specific location. As long as the positive instances are included, the bag type can be judged. The performance of Chinese painters when painting is unique, and the characteristics of strokes are difficult to capture. Besides, most of the Chinese paintings focus on the "ideal", and the subjectivity is strong. The objects depicted change with the

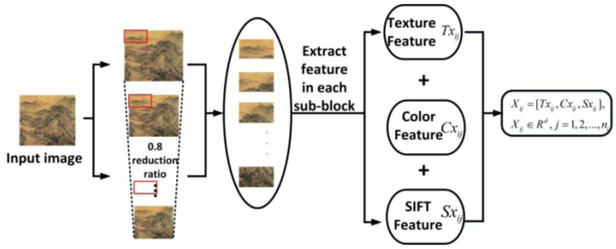


FIGURE 2. Schematic diagram of multi-instance modeling process.

author's emotions which are more abstract. Therefore, the multi-instance algorithm is very suitable for weakly labeled and more ambiguous Chinese painting images. In the modeling of multiple instances, in order to capture the global and local characteristics of Chinese painting images, that is, the description of the overall style and the details of local brush strokes, this article designed a block method of "Pyramid Overlapping Grid Division (POGP)" to achieve multi-bag multi-instance modeling. As shown in Figure 2, it is a multi-bag multi-instance modeling diagram. In the subsequent experiments, the block size of the POGP method is set to pixels, and the block moves from left to right and from top to bottom by 30 pixels. The image reduction ratio is 0.8, when the width or height of image is reduced to 244 pixels or less, the blocking stops. Then extract the brush strokes features for each sub-block, expressed as instances, and multiple instances as bags.

For each instance, texture features, color features and SIFT features are extracted separately. In order to pay attention to the spatial location information of the local Chinese painting image, use bior4.4 wavelet to carry out 4-layer wavelet decomposition for each instance, and then extract the energy distribution mean and variance of each sub-band on each decomposition layer as texture features [31], each instance can extract 26-dimensional texture feature vectors. Use color moments to extract color features for each instance, set the pixel on the i-th HSV color channel to $\{p(t)|t = 1, 2, \ldots, N\}$, and extract first-order (mean), second-order (variance), and third-order (slope) color-moment features for each channel. Therefore, each instance can extract 9-dimensional color feature vectors.

In view of the long origin of traditional Chinese painting images, when scanning into digital images, it is inevitable that there is a defect in the state of traditional Chinese painting, and the poor preservation environment of future generations leads to the problem of low image registration rate. Therefore, this article also chose the scale-invariant feature [32] transformation to describe instance. The scale-invariant feature transformation was first proposed by Lowe [33], also known as the image local feature descriptor. It has invariance to image rotation, scale scaling, and brightness changes. It finds extreme points in the scale space and performs Filter to find feature points, and extract 128-dimensional feature vectors with constant position, scale and rotation around the feature points.

Let $L = \{(B_1, y_1), (B_2, y_2), \ldots, (B_{|L|}, y_{|L|})\}$ be the labeled image set, where $|L|$ represents the total number of labeled images, $y_i \in \{+1, -1\}, i = 1, 2, \ldots N$. $+1$ means that the image of the object p is included, otherwise it is not. $U = \{B_1, B_2, \ldots, B_{|U|}\}$ is an unlabeled image set. The three forms of brush strokes feature with color, texture and SIFT are represented as $BC_i = \{Cx_{ij}\, |j = 1, \ldots, n_i\}$, $BT_i = \{Tx_{ij}|j = 1, \ldots, n_i\}$, $BS_i = \{Sx_{ij}\, |j = 1, \ldots, n_i\}$ respectively. Connect them in series, then the visual feature vector corresponding to each area is denoted as $X_{ij} = [Tx_{ij}, Cx_{ij}, Sx_{ij}], X_{ij} \in R^d, j = 1, 2, \ldots, n_i, d$ represents the dimension of the visual feature vector, and $B_i = \{X_{ij}\big| j = 1, 2, \ldots, n_i\}$ is called a multi-instance bag, where X_{ij} is an instance, and all instances in L are grouped together, called instance set, denoted by $IntSet = \{X_q\big| q = 1, 2, \ldots, Q\}$, $Q = \sum_{i=1}^{N} n_i$ is the total number of instances. After image feature extraction, each instance is represented by 163-dimensional features.

IV. SEQUENCE GENERATOR

Because the training samples of the MIL are bags, which contain varying numbers of instances, a key module in the proposed algorithm is the "***sequence generato*r**", whose function is to transform each bag into equal-length pseudo-sequence signals.

According to the definition of the MIL-based image classification, B_i corresponds to an image, and X_{ij} corresponds to the visual feature of a local region of an image. It is not difficult to see that in the instance feature space, if some images all contain the same semantics, they must contain some unique instances, which can reflect the essential commonality of this semantics, and have strong discriminating ability compared to the other images.

Therefore, to transform each bag into a pseudo-sequence signal, inspired by the DD function [23], this letter defines a new criterion function to pick out some unique instances from training bags in the training set, and call them as "***discriminative instance set*(DIS)**", to guide the construction of pseudo-sequence signals.

Specifically, each category of image is regarded as positive bag in turn, and the other categories of images are regarded as negative bags, as a result, the multi-instance training bags L can rewritten as $L = \{(B_i, y_i) : i = 1, 2, \ldots, N\}$, and here $y_i \in \{-1. + 1\}$ is the label.

In the multi-instance training bags L, let X denotes an instance from any positive bag, and we define its "uniqueness" calculation function as follows:

$$\begin{aligned}\mathrm{U}(X) &= \prod_{i=1}^{N} P(X|BC_i)\\ &= \prod_{i=1}^{N}\left[y_i^* + (-1)^{y_i^*} * P(X|X_i^*)\right]\\ &= \prod_{i=1}^{N}\left[y_i^* + (-1)^{y_i^*} * \exp(-\left\|X_i^* - X\right\|_2)\right] \qquad (1)\end{aligned}$$

where $y_i^* = (1 - y_i)/2$ (i.e. for any positive bag its $y^* = 0$, for any negative bag its $y^* = 1$, X_i^* denotes the instance closest to X in the i-th bag B_i, and $\left\|X_i^* - X\right\|_2$ denotes the Euclidean distance between X_i^* and X. The geometric meaning of the Equation (1) is as follows: For one instance X in any positive bag, if at least one instance in all positive bags is closer to it and all the instances in all negative bags are farther away from it, then the uniqueness of X is larger, and the more it should be selected as a discriminative instance.

Finally, following the above "uniqueness" calculation function, we can calculate the uniqueness of each instance in all training bags, and then collect the top-T instances with the largest uniqueness as the final DIS, which is recorded as $\Omega = \{V_1, V_2, \ldots, V_T\}$. As a result, under the guidance of Ω, any bag B_i can be transformed into a pseudo-sequence composed of T signals with attention, which is formulated as:

$$\phi(B_i) = \{(X_t, w_t) : t = 1, 2, \ldots, T\} \tag{2}$$

where $X_t \in \mathbb{R}^d$ and w_t represent the t-th signal and its attention in the pseudo-sequence respectively. Specifically, we calculate the distance between each instance in B_i and V_t. Assuming that the distance between instance $F_{i,J}$ and V_t is the smallest, then:

$$\begin{cases} X_t &= F_{i,J} \\ w_t &= 1/\left\|F_{i,J} - V_t\right\|_2 \end{cases} \tag{3}$$

where X_t denotes the instance of B_i closest to V_t $(t = 1, 2, \ldots, T)$ according to Euclidean distance, and w_t is the inverse of this minimum distance (i.e. the smaller the distance, the greater the attention). Finally, the detailed steps for constructing pseudo-sequence signals from multi-instance bags are summarized as follows:

```
Algorithm 1 Sequence Generator
    Input: Training bags: $D = \{(B_i, L_i) : i = 1, 2, \ldots, N\}$; $T$: size of DIS
    Output: DIS $\Omega$ and. pseudo-sequence $\phi(B_i)$
Step 1: Regarded each category of bags in D as positive bags in turn, and the other categories of bags as negative bags, then calculate each instance's uniqueness via Equation (1);
Step 2: Select top-T instances with the largest uniqueness as discriminative instances, and add them to $\Omega$ regard as DIS $\Omega = \{V_1, V_2, \ldots, V_T\}$;
Step 3: For $\forall B_i \in D$, calculate its pseudo-sequence and attention $\phi(B_i) = \{(X_t, w_t) : t = 1, 2, \ldots, T\}$ via Equation (2);
Step 4: Return $\Omega$ and $\phi(B_i)$, $i = 1, 2, \ldots, N$
```

V. LSTM WITH ATTENTION MECHANISM MODEL

A. LONG SHORT TERM MEMERY NETWORK

Recurrent neural network is a commonly used method to deal with sequence problems. It has been extensively studied in the fields of text classification, translation, behavior recognition [34]–[36], etc. With the increase of sequence length or the number of network layers, RNN is used in information synthesis. The loss in the process is large, often focusing on the content of the last stage of learning information, unable to achieve long-term memory. LSTM is an improved structure of RNN. It overcomes the problems of gradient disappearance and gradient explosion owned by RNN, changes the hidden layer of RNN, adds control gate and memory unit $c^{<t>}$; Memory unit is the key to the network, it stores and transmits useful sequence information. The control gate is composed of an input gate $i^{<t>}$, a forget gate $f^{<t>}$ and an output gate $o^{<t>}$, which respectively control whether the current state of the input, forget and output. When a new input comes, the input gate works to store new information. At the same time, the forget gate determines how much information is forgotten before that moment, and the output gate further controls how much cell state c information is transmitted to the final state $h^{<t>}$.The internal structure of a single-layer LSTM network is shown in Figure 3. The LSTM activations are calculated as follows:

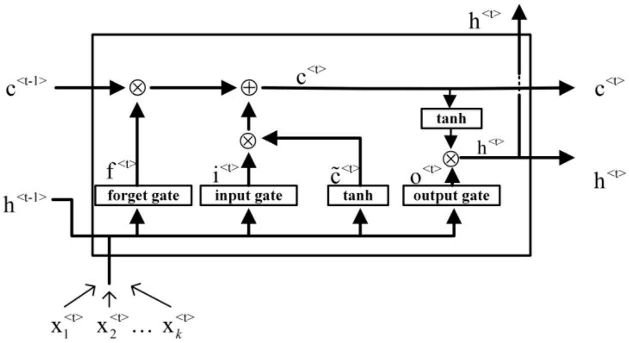


FIGURE 3. The internal structure of a single-layer LSTM network.

$$\begin{aligned}
\tilde{c}^{<t>} &= \tanh(W_{\tilde{c}x}x^{<t>} + W_{\tilde{c}h}h^{<t-1>} + b_c) \\
i^{<t>} &= \sigma(W_{ix}x^{<t>} + W_{ih}h^{<t-1>} + b_i) \\
f^{<t>} &= \sigma(W_{fx}x^{<t>} + W_{fh}h^{<t-1>} + b_f) \\
o^{<t>} &= \sigma(W_{ox}x^{<t>} + W_{oh}h^{<t-1>} + b_o) \\
c^{<t>} &= i^{<t>} \circ \tilde{c}^{<t>} + f^{<t>} \circ \tilde{c}^{<t-1>} \\
h^{<t>} &= o^{<t>} \circ c^{<t>}
\end{aligned} \tag{4}$$

where W_{cx},$W_{ch}W_{ix}$,W_{ih},W_{fx},$W_{fh,}W_{ox}$,W_{oh} are the weight matrix of memory cell, input gate, forget gate, and output gate. $b_{\tilde{c}}$,b_i,b_f,b_o are the corresponding bias.

B. MULTILAYER LSTM CLASSIFICATION MODEL WITH ATTENTION MECHANISM

In the algorithm proposed in this article, the Chinese painting images are divided into blocks and treated as a set of examples to provide more training data. Then, clustering algorithms are used to convert examples with similar semantics into semantic sequences. For label correlation, a multi-level LSTM with attention mechanism is used to train the classification model, as shown in Figure 4.

Attention mechanism is introduced in this architecture. It assigns different weights to the upper LSTM input through

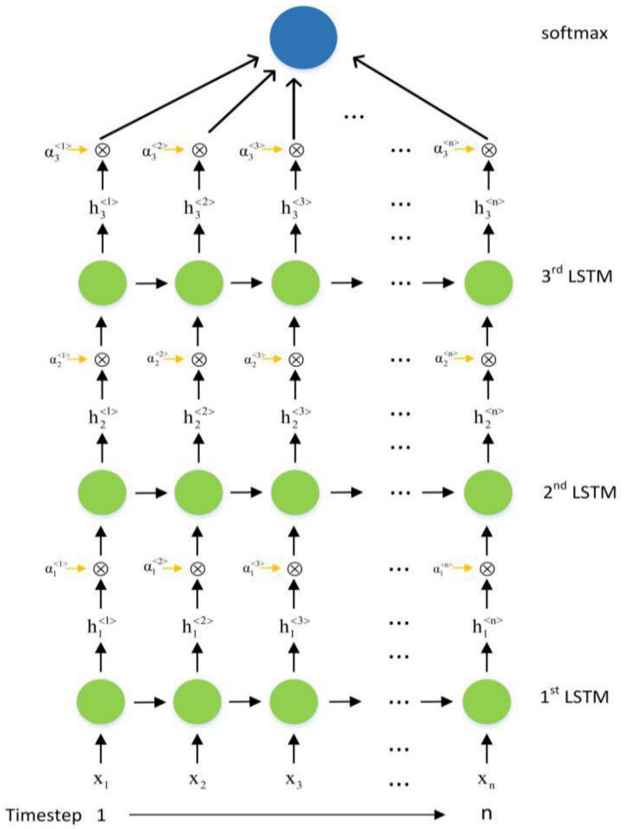

softmax
3rd LSTM
2nd LSTM
1st LSTM
$\alpha_3^{<1>}$ $\alpha_3^{<2>}$ $\alpha_3^{<3>}$ $\alpha_3^{<n>}$
$h_3^{<1>}$ $h_3^{<2>}$ $h_3^{<3>}$ $h_3^{<n>}$
$\alpha_2^{<1>}$ $\alpha_2^{<2>}$ $\alpha_2^{<3>}$ $\alpha_2^{<n>}$
$h_2^{<1>}$ $h_2^{<2>}$ $h_2^{<3>}$ $h_2^{<n>}$
$\alpha_1^{<1>}$ $\alpha_1^{<2>}$ $\alpha_1^{<3>}$ $\alpha_1^{<n>}$
$h_1^{<1>}$ $h_1^{<2>}$ $h_1^{<3>}$ $h_1^{<n>}$
x_1 x_2 x_3 x_n
Timestep 1
n

FIGURE 4. The structure of multi-level LSTM network with attention mechanism, where $h_i^{<j>}$ is the output state of the j-th timestep in the i-th LSTM. α_i is the attention vector of the i-th layer.

the learning network, and can dynamically capture the key parts of the input. For inputs that are helpful for classification, the network assigns higher weights to retain useful information; for inputs that are not good for judgment, assign low or no assignments to discard information. Each input vector weight is calculated by the attention model function. Let $H \in R^{d\times n}$ be an input sequence $\{h^{<1>}, h^{<2>}, \ldots, h^{<n>}\}$ in which each element is a input vector, d is the size of hidden layers whereas n is the length of sequence. The attention mechanism can be computed as follows:

$$e^{<t>} = att(\mathrm{h}) = V_a^T \tanh(W_a * \mathrm{h} + b_a) \tag{5}$$

The attention vector of the i-th layer is expressed by the following formula:

$$\alpha^{<t>} = \frac{\exp(e^{<t>})}{\sum_{t=1}^{n} \exp(e^{<t>})} \tag{6}$$

The network architecture consists of three layers of LSTM and attention mechanism layer and a softmax classification layer. The input $x^{<t>}$ of the first layer LSTM is a semantic sequence with time sequence. The input of the upper LSTM is the output of the lower LSTM, that is, the final status $h^{<t>}$ of this layer. Multi-layer LSTM can fully capture the long-term dependencies of input sequences. In order to provide a confidence score at the last time step n, a softmax layer is added to the highest LSTM layer.

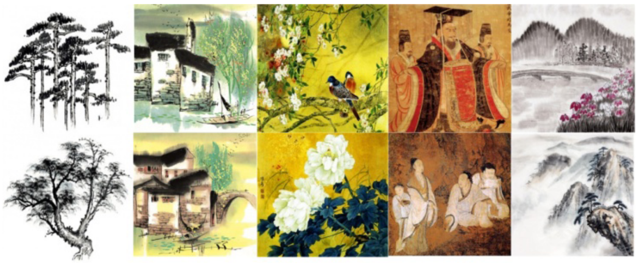


FIGURE 5. Sample images of the CP. From left to right are ancient trees, Jiangnan water-bound town, flowers&birds, people and ink painting.

In the first layer structure, the input is $x_i, i = 1, 2, \ldots n$. After processing by the LSTM unit and the attention mechanism, the output state h_1 and the attention vector α_1 are obtained, and the product of them is used as the input of the next layer. By analogy, during softmax classification, the input becomes the product of the attention vector and the output state at all time steps of the third LSTM layer. The choice of activation function affects the nonlinear expression ability of the network model. In this article, tanh is used as the activation function. The reason for choosing it is that the function is symmetric about the origin, which makes the input data training effect better and the convergence speed faster. In addition, the fault tolerance and performance of the tanh function are better. Use the tanh activation function for the output of the third LSTM layer, which is publicly expressed as:

$$\mathrm{z} = \tanh(\mathrm{W}(\sum_{t=1}^{n} \alpha_3^{<t>} \mathrm{h}_3^{<t>}) + \mathrm{b}) \tag{7}$$

After through the softmax layer, the probability score that the sequence X belongs to the category $C_i, i = 1, 2, \ldots |L|$ as follows:

$$p(C_i\,|X) = \frac{\exp(z_i)}{\sum_{j=1}^{|L|} \exp(z_j)} \tag{8}$$

VI. EXPERIMENTS AND RESULTS

A. IMAGE SET AND EXPERTMENTAL SETUP

In order to verify the performance of the proposed ALSTM-MIL algorithm in Chinese painting image classification, we applied it on the CP image set. The CP image set contains 2,000 images in JPEG format with size 256 × 384 or 384 × 256. There are all-together five different categories, each containing 400 images, the names are ancient trees, people, flowers&birds, Jiangnan water-bound town, and ink paintings. Figure 5 shows a sample of the CP image data set. In the experiment, we also chose the Corel image library [37] for comparison, which is divided into 20 different categories, which are Africa, Beach, Building, Buses, Dinosaurs, Elephants, Flowers, Horses, Mountains, Food, Dogs, Lizards, Fashion models, Sunset scenes, Cars, Waterfalls, Antique

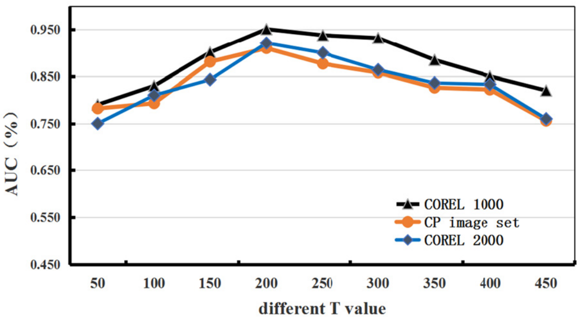


FIGURE 6. Different T values impact to the classification accuracies of the proposed algorithms.

furniture, Battle ships, Skiing and Desserts, each category contains 100 color images, a total of 2000 images. We refer to 1,000 images of the top 10 categories names COREL 1000 [38], [39], and the entire 2000 images are COREL 2000 image set.

During subsequent experiments, 60% of images are randomly selected from one category to form training set, and all the remaining images to form test set. In the ALSTM-MIL algorithms, the number of LSTM network neurons is set to 1024, initialized using uniform random numbers between the -0:05 and 0:05. As one of the input parameters of the algorithm, the learning rate is set to 0.001. For each training, the batch size is set to 64, Adam [40] is adapted to automatically adjust the learning rate, and the training process was run for 500 epochs. All experiments were performed under win10 + python3.6 environment with Intel i7 2.8GHZ CPU and 16GB RAM memory.

B. SENSITIVITY TEST

While we use Algorithm 1 to construct DIS and pseudo-sequence, one important parameter T must be predefined. In order to confirm the influence of T value to the proposed algorithm, we chose T from 200 to 450 with step size 50. Over ten rounds repeated training and testing, the average classification accuracies in CP image set, COREL 1000, COREL 2000 are shown in Figure 6.

As can be seen from Figure 6, the T values have effect to the performance of the proposed algorithm. The reason is that when T is too small, the pseudo-sequence constructed by them is not complete, which affects the expression ability of memory coding features. On the contrary, when T is too large, it only increases the redundancy of the pseudo-sequence, which does not help to improve classification accuracy. Therefore, in subsequent comparative experiments, the T value in algorithm was set to 200.

C. ACCURACY AND CONFUSION MATRIX ON PROPOSED ALGORITHM

In order to verify the effectiveness of the algorithm proposed, we performed comparative experiments with various other MIL algorithms, such as MI/mi-SVM [17], DD-SVM [18], MILES [19], and some classic deep learning algorithm, such as CNN [25], MIDDP [28], LSTM [32]. The average AUC values used in the comparison experiments on three data sets are shown in Table 1.

TABLE 1. Comparison of experimental results (%).

Algorithms	COREL 1000	COREL 2000	CP Image sets
ALSTM-MIL	**0.951± .034**	**0.927± .021**	**0.912± .043**
DD-SVM	0.817± .075	0.795± .015	0.797± .067
MILES	0.826± .041	0.804± .033	0.805± .051
MI-SVM	0.747± .014	0.743± .015	0.725± .043
mi-SVM	0.774± .055	0.761± .024	0.756± .062
MiDDP	0.923± .050	0.905± .037	0.889± .014
CNN	0.909± .059	0.882± .019	0.864± .058
LSTM	0.832± .032	0.815± .039	0.811± .042

From the experimental results in Table 1, it can be seen that the classification performance of the proposed algorithm on three datasets is better than the existing classic algorithms.

The main reasons are as follows. First, the sequence generator accurately generates semantic sequences from the instance features and has a strong image semantic representation capability. Then, when the semantic sequence is passed to a multi-layer LSTM network model, the model has a strong long sequence processing capability, which can take into account the weight relationship between the various semantics, making the algorithm adaptive and robust. The confusion matrix of CP image sets and COREL 1000 is shown in Figure 7. We can see that our model performs well on most of image classification.

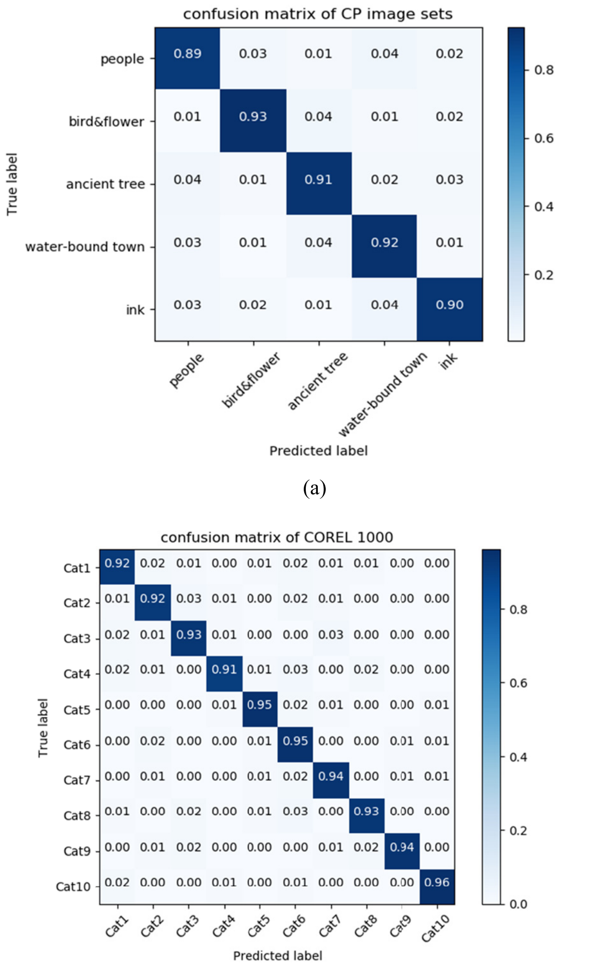
D. COMPARISON MODEL OF ALGORITHM WITH AND WITHOUT ATTENTION MECHANISM

In order to verify the importance of the attention mechanism in the proposed algorithm, we conducted an ablation study. Compare the classification accuracy of the Baseline, model with the attention mechanism and without the attention mechanism respectively on CP image set. Table 2 shows the results of the three models.

TABLE 2. Comparison of experimental results (%).

Model	Accuracy	F1 score	AUC
Baseline	0.823	0.832	0.825
multi-layer LSTM Without Attention	0.878	0.882	0.873
ALSTM-MIL	**0.915**	**0.925**	**0.912**

It can be seen from the experimental results in Table 2 that the proposed ALSTM-MIL algorithm is 3% higher in classification accuracy than without Attention mechanism and 9% higher than Baseline model. Because the attention mechanism assigns different weights to different instances in the network, increasing the proportion of useful instances, thereby improving the accuracy of Chinese painting image



(a)

(b)

FIGURE 7. (A) confusion matrix of CP image sets. (B) confusion matrix of COREL 1000.

TABLE 3. Comparison of average running time(SEC).

Bag generator	Total time	Time per image
POGD	**130**	**0.13**
DD-SVM	477	0.47
blobworld	159	0.15

classification. At the same time, multi-layer LSTM can learn more memory information than single-layer LSTM.

E. AVERAGE RUNNING TIME COMPARISON

We respectively compared the average running time of the POGD method in this article with blobworld [41] and bag generator of DD-SVM on the CP dataset, as shown in Table 3.

From the results in the table 3, it can be seen that the POGD method is 29s shorter than the blobworld. The reason is that each instance of blobworld needs to extract 230-dimensional features, which is more than the number of dimensions extracted by POGP. In addition, the running time of this method is much less than that of the bag generator in DD-SVM. This because DD-SVM finds the concept point of positive and negative instances based on the diversity density function, which is time-consuming.

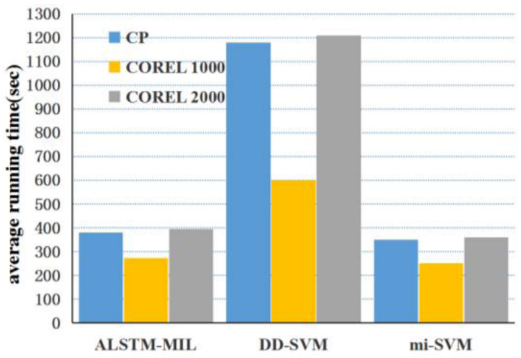


FIGURE 8. Average running time of three algorithms in CP, COREL 1000 and COREL 2000 image sets.

In addition, we have compared the performance of our method and other methods, and it is still interesting to know the average running time of each MIL algorithm. We conduct 20 experiments on these methods and take the average as the running time. Figure 8 illustrates the average running time of different MIL algorithms ALSM-MIL, mi-SVM and DD-SVM executing the COREL 1000, COREL 2000 and CP image sets, respectively.

According to the experimental results, we found that ALSTM-MIL has no obvious advantage in time efficiency. This is because the LSTM model with attention mechanism requires iterative training, and the training batches extend the running time. In addition, the running time of ALSTM-MIL on the COREL 2000 and CP image sets differs very little. The reason is that the data sets are similar in size.

VII. DISCUSSION

There are still some interesting optimizable areas about Chinese painting image classification model. For example, when designing a sequence generator, we can choose more distance metrics, such as Levenshtein distance [42] to verify classification performance.

The application areas of the algorithm proposed in this article are not only in image classification and retrieval, but also in image surface defect detection and target tracking. In the training phase of image defect detection in [43], the algorithm in this article can be used to implement multi-instance modeling on surface images to construct a multi-instance bag. Then, using the uniqueness calculation function mentioned to calculate and select the top-T instances with the largest uniqueness as the discriminative instance set in order to implement the surface defect detection. In addition, the proposed algorithm can also be used for target tracking [44], [45]. Since the target deforms with movement, the multi-instance modeling method should consider the instances near the target as positive bags. Moreover, although the ALSTM-MIL algorithm we discuss is aimed at the classification of Chinese painting images, it is undeniable that when there is other image retrieval or classification problem, the classification accuracy can be improved by only modifying the extracted feature vectors for specific images.

In addition, in the future work, other factors inherent in the classification of traditional Chinese painting images can be considered to further improve performance.

VIII. CONCLUSION

In order to realize the classification of Chinese painting images under the framework of MIL, the work of this article is mainly reflected in three aspects: (1) A new multi-instance modeling scheme is designed, that is, the Pyramid Overlapping Grid Division method is used to convert Chinese painting images into multiple instance bag; (2) A sequence generator is designed to construct a discriminative instance set, and the multi-instance bag is transformed into an ordered and equal-length semantic sequence; (3) A novel MIL algorithm based on LSTM with attention mechanism (ALSTM-MIL) for Chinese painting classification is proposed into perform semantic analysis the discriminative instance set. Combining long and short-term memory neural network with MIL algorithm is proposed for the first time. Capturing and long-term memory to identify the correlation information between instances, and finally combining with the Softmax layer to realize the classification of Chinese painting images, which is innovative in the research of MIL algorithm. The experimental results show that the proposed algorithm is a very effective MIL algorithm, and its classification performance in the COREL image sets and the Chinese painting image set is better than other algorithms.

REFERENCES

[1] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Comput. Vis. Pattern Recognit.*, Boston, MA, USA, Sep. 2014, pp. 1–14.

[2] M. Korytkowski, L. Rutkowski, and R. Scherer, "Fast image classification by boosting fuzzy classifiers," *Inf. Sci.*, vol. 327, pp. 175–182, Jan. 2016, doi: 10.1016/j.ins.2015.08.030.

[3] M. Wang, J. Wang, and Y. S. Wang, "Multi-scale algorithm of texture feature extraction based on gray-level co-occurrence matrix," *Chin. J. Liquid Crystals Displays*, vol. 31, no. 10, pp. 967–972, 2016, doi: 10.3788/YJYXS20163110.0967.

[4] J. C. Sheng, "Automatic categorization of traditional Chinese paintings based on wavelet transform," *Comput. Sci.*, vol. 41, no. 2, pp. 317–319, 2014, doi: 10.3969/j.issn.1002-137X.2014.02.069.

[5] K. A. Jangtjik, T.-T. Ho, M.-C. Yeh, and K.-L. Hua, "A CNN-LSTM framework for authorship classification of paintings," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Beijing, China, Sep. 2017, pp. 2866–2870.

[6] P. Buchana, I. Cazan, M. Diaz-Granados, F. Juefei-Xu, and M. Savvides, "Simultaneous forgery identification and localization in paintings using advanced correlation filters," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Phoenix, AZ, USA, Sep. 2016, pp. 146–150, doi: 10.1109/ICIP.2016.7532336.

[7] G. Polatkan, S. Jafarpour, A. Brasoveanu, S. Hughes, and I. Daubechies, "Detection of forgery in paintings using supervised learning," in *Proc. 16th IEEE Int. Conf. Image Process. (ICIP)*, Cairo, Egypt, Nov. 2009, pp. 7–10, doi: 10.1109/ICIP.2009.5413338.

[8] M. Sun, D. Zhang, Z. Wang, J. Ren, and J. S. Jin, "Monte Carlo convex hull model for classification of traditional Chinese paintings," *Neurocomputing*, vol. 171, pp. 788–797, Jan. 2016, doi: 10.1016/j.neucom.2015.08.013.

[9] J. Li and J. Z. Wang, "Studying digital imagery of ancient paintings by mixtures of stochastic models," *IEEE Trans. Image Process.*, vol. 13, no. 3, pp. 340–353, Mar. 2004.

[10] J. Sheng and J. Jiang, "Recognition of Chinese artists via windowed and entropy balanced fusion in classification of their authored ink and wash paintings (IWPs)," *Pattern Recognit.*, vol. 47, no. 2, pp. 612–622, Feb. 2014.

[11] F. Gao, J. Nie, L. Huang, and L. Y. Duan, "Traditional Chinese painting classification based on painting techniques," *Chin. J. Comput.*, vol. 40, no. 12, pp. 2871–2882, 2017, doi: 10.11897/SP.J.1016.2017.02871.

[12] W. Jiang, Z. Wang, J. S. Jin, Y. Han, and M. Sun, "DCT–CNN-based classification method for the Gongbi and Xieyi techniques of Chinese ink-wash paintings," *Neurocomputing*, vol. 330, pp. 280–286, Feb. 2019, doi: 10.1016/j.neucom.2018.11.003.

[13] H. Bao, Y. Liang, H.-Z. LIU, and D. Xu, "A novel algorithm for extraction of the scripts part in traditional Chinese painting images," in *Proc. 2nd Int. Conf. Softw. Technol. Eng.*, Puerto Rico, USA, Oct. 2010, pp. 26–30, doi: 10.1109/ICSTE.2010.5608756.

[14] X.-S. Wei and Z.-H. Zhou, "An empirical study on image bag generators for multi-instance learning," *Mach. Learn.*, vol. 105, no. 2, pp. 155–198, Nov. 2016, doi: 10.1007/s10994-016-5560-1.

[15] M. A. Medina-Pérez, M. Á. F. Ballester, M. García-Borroto, O. Loyola-González, A. M. Moreno, and L. Altamirano-Robles, "Latent fingerprint identification using deformable minutiae clustering," *Neurocomputing*, vol. 175, pp. 851–865, Jan. 2016, doi: 10.1016/j.neucom.2015.05.130.

[16] J. Foulds and E. Frank, "A review of multi-instance learning assumptions," *Knowl. Eng. Rev.*, vol. 25, no. 1, pp. 1–25, Mar. 2010, doi: 10.1017/S026988890999035X.

[17] S. Andrews, T. Hofmann, and L. Tsochantaridis, "Multiple instance learning with generalized support vector machines," in *Proc. Int. AAAI*, Edmonton, AB, Canada, Jul. 2002, pp. 943–944.

[18] Y. X. Chen and J. Z. Wang, "Image categorization by learning and reasoning with regions," *J. Mach. Learn. Res.*, vol. 5, no. 8, pp. 913–939, 2004, doi: 10.5555/1005332.1016789.

[19] Y. Chen, J. Bi, and J. Z. Wang, "MILES: Multiple-instance learning via embedded instance selection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 12, pp. 1931–1947, Dec. 2006, doi: 10.1109/TPAMI.2006.248.

[20] D.-X. Li, J.-Y. Peng, Z. Li, and Q. Bu, "LSA based multi-instance learning algorithm for image retrieval," *Signal Process.*, vol. 91, no. 8, pp. 1993–2000, Aug. 2011, doi: 10.1016/j.sigpro.2011.03.004.

[21] W. Hu, X. Ding, B. Li, J. Wang, Y. Gao, F. Wang, and S. Maybank, "Multi-perspective cost-sensitive context-aware multi-instance sparse coding and its application to sensitive video recognition," *IEEE Trans. Multimedia*, vol. 18, no. 1, pp. 76–89, Jan. 2016, doi: 10.1109/TMM.2015.2496372.

[22] W.-J. Li and D.-Y. Yeung, "Localized content-based image retrieval through evidence region identification," in *Proc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Miami, FL, USA, Jun. 2009, pp. 1666–1673, doi: 10.1109/CVPRW.2009.5206796.

[23] J. Wu, S. Pan, X. Zhu, C. Zhang, and X. Wu, "Multi-instance learning with discriminative bag mapping," *IEEE Trans. Knowl. Data Eng.*, vol. 30, no. 6, pp. 1065–1080, Jun. 2018, doi: 10.1109/TKDE.2017.2788430.

[24] X.-S. Wei, J. Wu, and Z.-H. Zhou, "Scalable algorithms for multi-instance learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 4, pp. 975–987, Apr. 2017, doi: 10.1109/TNNLS.2016.2519102.

[25] Y. Xu, T. Mo, Q. W. Feng, P. L. Zhong, M. D. Lai, and E. C. Chang, "Deep learning of feature representation with multiple instance learning for medical image analysis," in *Proc. Int. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Florence, Italy, May 2014, pp. 1626–1630.

[26] J. Wu, Y. Yu, C. Huang, and K. Yu, "Deep multiple instance learning for image classification and auto-annotation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Boston, MA, USA, Jun. 2015, pp. 3460–3469.

[27] K. L. He, Y. H. Shi, Y. Gao, J. Huo, D. Wang, and Y. Zhang, "A prototype learning based multi-instance convolutional neural network," *Computer*, vol. 40, no. 6, pp. 1–12, 2017, doi: 10.11897/SP.J.1016.2017.01265.

[28] P. Tang, X. Wang, B. Feng, and W. Liu, "Learning multi-instance deep discriminative patterns for image classification," *IEEE Trans. Image Process.*, vol. 26, no. 7, pp. 3385–3396, Jul. 2017, doi: 10.1109/TIP.2016.2642781.

[29] D. A. Ajjaji, M. A. Alsaeed, A. S. Alswayed, and H. S. Alhichri, "Multi-instance neural network architecture for scene classification in remote sensing," in *Proc. Int. Conf. Comput. Inf. Sci. (ICCIS)*, Aljouf, Kingdom Saudi Arabia, Apr. 2019, pp. 1–5, doi: 10.1109/ICCISci.2019.8716411.

[30] K. Das, S. Conjeti, A. Guha Roy, J. Chatterjee, and D. Sheet, "Multiple instance learning of deep convolutional neural networks for breast histopathology whole slide classification," in *Proc. IEEE 15th Int. Symp. Biomed. Imag. (ISBI)*, Washington, DC, USA, Apr. 2018, pp. 578–581, doi: 10.1109/ISBI.2018.8363642.

[31] L. Liu and G. Kuang, "Overview of image textural feature extraction methods," *J. Image Graph.*, vol. 14, no. 4, pp. 622–635, 2009.

[32] A. Vedaldi and B. Fulkerson, "VLFeat: An open and portable library of computer vision algorithms," in *Proc. Int. Conf. Multimedia (MM)*, Firenze, Italy, 2010, pp. 1469–1472, doi: 10.1145/1873951.1874249.

[33] G. D. Lowe, "Distinctive image features from scale-in-variant interest points," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

[34] X. J. Shi, Z. R. Chen, H. Wang, D. Y. Yeung, W. K. Wong, and W. C. WOO, "Convolutional LSTM network: A machine learning approach for precipitation nowcasting," in *Proc. Comput. Vis. Pattern Recognit.*, Boston, MA, USA, 2015, pp. 802–810, doi: 10.5555/2969239.2969329.

[35] F. J. Ordóñez and D. Roggen, "Deep convolutional and LSTM recurrent neural networks for multimodal wearable activity recognition," *Sensors*, vol. 16, no. 1, pp. 115–140, 2016, doi: 10.3390/s16010115.

[36] S. Zhang, Y. Yang, J. Xiao, X. Liu, Y. Yang, D. Xie, and Y. Zhuang, "Fusing geometric features for skeleton-based action recognition using multilayer LSTM networks," *IEEE Trans. Multimedia*, vol. 20, no. 9, pp. 2330–2343, Sep. 2018, doi: 10.1109/TMM.2018.2802648.

[37] J. Yang, Y.-G. Jiang, A. G. Hauptmann, and C.-W. Ngo, "Evaluating bag-of-visual-words representations in scene classification," in *Proc. Int. Workshop Workshop Multimedia Inf. Retr. (MIR)*, 2007, pp. 197–206.

[38] F. Tajeripour, M. Saberi, and S. F. Ershad, "Developing a novel approach for content based image retrieval using modified local binary patterns and morphological transform," *Int. Arab J. Inf. Technol.*, vol. 12, no. 6, pp. 574–581, 2015.

[39] A. K. Bhunia, A. Bhattacharyya, P. Banerjee, P. P. Roy, and S. Murala, "A novel feature descriptor for image retrieval by combining modified color histogram and diagonally symmetric co-occurrence texture pattern," *Pattern Anal. Appl.*, vol. 23, no. 2, pp. 703–723, May 2020, doi: 10.1007/s10044-019-00827-x.

[40] R. F. Rachmadi, K. Uchimura, G. Koutaki, and K. Ogata, "Single image vehicle classification using pseudo long short-term memory classifier," *J. Vis. Commun. Image Represent.*, vol. 56, pp. 265–274, Oct. 2018, doi: 10.1016/j.jvcir.2018.09.021.

[41] C. Carson, S. Belongie, H. Greenspan, and J. Malik, "Blobworld: Image segmentation using expectation-maximization and its application to image querying," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 8, pp. 1026–1038, Aug. 2002, doi: 10.1109/TPAMI.2002.1023800.

[42] K.-M. Ahn, Y.-S. Kim, Y.-H. Kim, and Y.-H. Seo, "Sentiment classification of movie reviews using Levenshtein distance," *J. Digit. Contents Soc.*, vol. 14, no. 4, pp. 581–587, Dec. 2013, doi: 10.9728/dcs.2013.14.4.581.

[43] S. Fekri-Ershad and F. Tajeripour, "Multi-resolution and noise-resistant surface defect detection approach using new version of local binary patterns," *Appl. Artif. Intell.*, vol. 31, nos. 5–6, pp. 395–410, Jul. 2017, doi: 10.1080/08839514.2017.1378012.

[44] B. Xu, W. Li, D. Tzoumanikas, M. Bloesch, A. Davison, and S. Leutenegger, "MID-fusion: Octree-based object-level multi-instance dynamic SLAM," in *Proc. Int. Conf. Robot. Autom. (ICRA)*, Montreal, QC, Canada, May 2019, pp. 5231–5237.

[45] B. Babenko, M.-H. Yang, and S. Belongie, "Robust object tracking with online multiple instance learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 8, pp. 1619–1632, Aug. 2011, doi: 10.1109/TPAMI.2010.226.

DAXIANG LI received the M.S. degree from the Department of Electronics Science, Northwest University, China, in 2005, and the Ph.D. degree from the Northwest University, in 2015. He is currently an Associate Professor with the School of Telecommunication and Information Engineering, Xi'an University of Posts and Telecommunications, China. His research interests include image retrieval, image classification, object detection and tracking, multi-instance learning, and deep learning.

YUE ZHANG received the B.S. degree from the Xi'an University of Posts and Telecommunications, China, in 2018, where she is currently pursuing the master's degree with the School of Telecommunication and Information Engineering. Her research interests include image retrieval, machine learning, and pattern recognition.