Deep Feature Fusion Multiple Instance Learning for WaDang Recognition

Chao WEN1, Zhan LI1,2, AiPing Li2, Jian QU1
1Research Center for Intelligent Interaction and Information Art, Northwest University, 710069, Xi’an, China
2School of Information Science and Technology, Northwest University, 710127, Xi’an, China
Corresponding author: Chao WEN (wenchao@nwu.edu.cn), Zhan LI (lizhan@nwu.edu.cn)

This work was supported by the China Postdoctoral Science Foundation (Grant No. 2014M552479, 2014M552478), Humanities and Social Science Youth Funded Project of MOE of China (Grant No. 17YJCZH186), Natural Science Basic Research Plan in Shaanxi Province of China (Grant No. 2013JQ8022), Scientific Research Project of Education Department of Shaanxi Province (Grant No. 2013JK1181, 2014JK1724).

ABSTRACT WaDang recognition can provide valuable information for digital protection of cultural relics. However, WaDang recognition is challenging due to a great amount of variations, such as different characters, different background pattern, different scale and rotation. In this paper, we try to solve these problems by using hand-crafted and deep features and integrate them into multiple instance learning, in which the features of instances are divided into two parts and their similarity is computed by multiple kernel learning. Experiments on our collected WaDang images, MUSK and COREL datasets show that the proposed algorithm is effective and its performance is compared with other state-of-art algorithms.

INDEX TERMS Multiple instance learning, Convolutional neural networks (CNN), WaDang recognition

I. INTRODUCTION
As a well-known cultural relic in China, WaDang enjoys the same reputation as oracle bone inscriptions, bronzes, painted pottery, bamboo slips, seals and mud seals. Being one of the components of ancient Chinese architecture, WaDang originated in Chinese Western Zhou era more than 3,000 years ago. Thus, to predict the WaDang images is very important in digital protection of cultural relics in China. However, in our collected WaDang image set, the task of WaDang recognition is very difficult because there are one, two, three, four, five or even more than six Chinese characters in WaDang dataset and same character appearing in different WaDang images has more than 20 or even 100 kinds of forms. We choose some WaDang images and show them in Figure 1.

Based on our observation, WaDang images usually consist of Chinese characters inside the circle and has the following characteristics: 1) The Chinese characters contained in WaDang images can be acquired in the intersects of specific circles and straight lines by utilizing the geometric symmetry. 2) The task of WaDang recognition is very difficult for the variations existed in it, which need rich features including hand-crafted and deep features to express WaDang images and capture the local, geometry and global features. 3) If we give the whole WaDang image label not the characters in it, the WaDang recognition is naturally casted as multiple instance learning (MIL) problem. In the MIL paradigm, we only know the WaDang image (bag) contains specific character (positive instance), but we do not know where it is, nor which character (instance) is positive.

Considering the above characteristics, this paper firstly adopts circle and line detection methods to eliminate the non-character signals in WaDang image and to obtain the Chinese characters, in which the hand-crafted Scale Invariant Feature Transform (SIFT) and mathematical morphologic features and deep convolutional neural networks (CNN) features are obtained. Finally, we divide the features into two parts: one is SIFT, the other is morphological and deep features. The similarity of these two parts is computed by multiple kernel and feed them to support vector machine (SVM) to recognize WaDang images.

FIGURE 1. Some examples of WaDang images
To emphasize the main contribution of this paper, we summarize the following distinct advantages of our method.

1) Since the WaDang recognition task is never conducted before, we firstly collect and build a WaDang image dataset whose detailed information will be described in Section IV, and study the WaDang image recognition problem, which is valuable for understanding other cultural relics.

2) To obtain accurate predictions for WaDang, we propose multi-scale sparse method and joint dictionary learning to obtain the SIFT bases and multiple concept points of MIL.

3) Rich and robust features are needed to describe the WaDang images and get better performance which lead us to combine the hand-crafted features and deep features together.

4) Experiments on WaDang, MUSK and COREL show that our proposal can outperform the state-of-art baselines including traditional MIL and deep MIL algorithms.

II. RELATED WORKS

A. DICTIONARY LEARNING

Dictionary learning algorithms can be divided into two categories: unsupervised learning and supervised learning.

Unsupervised learning method learns dictionary by minimizing signal residuals, such as greedy algorithm and orthogonal basis tracking algorithm. For example, [1][2] adopt greedy algorithm and orthogonal basis tracking algorithm to find the basis of the signal. In [3], K-SVD algorithm is proposed to learn overcomplete dictionary from image blocks by K-means clustering method. Lee [4] regards dictionary learning as least squares problem and uses Lagrange to solve it. Wright et al. [5] used the whole training sample as a dictionary in the task of face detection, and achieved good results.

Supervised learning method either generate a dictionary for each class, and then combine the dictionaries of each class to generate a new dictionary, or take full account of the constraints of classification signals when generating dictionaries, so as to generate the required dictionary. In [6][7] integrates training samples and dictionary learning process into an objective function to learn a new base matrix. [8] uses aggregated information bottleneck (AIB) to iteratively merge two dictionary atoms, thus reducing the mutual information between dictionary atoms and class labels. [9-11] optimizes the parameters of classifier and dictionary jointly, so as to improve the classification performance by using the learning dictionary.

B. MULTIPLE INSTANCE LEARNING

As a special learning schema, multiple instance learning is suitable for expressing incomplete labels and ambiguity problems [12]. According to the learning mechanism of multiple instance learning, it can be roughly classified into two classes: instance-based method and structure-based method. The instance-based multiple instance learning algorithm belongs to a bottom-up approach, which first determines the label of each instance, and then infers the bag label [12-16,41]. The structure-based multiple instance learning algorithm, on the contrary, belongs to the top-down approach, which learns the labels of the instances [17-25] after determining the label of the bag.

In these methods, one called concept points method has been mostly studied, including ARP [12], DD [13], DD-SVM [24] and MILES [25] etc. ARP [12] and DD [13] algorithms consider that all positive bags should contain positive instances, and thus these positive instances should densely compact in a rectangle or ellipse region of feature space, which can be used to infer the bag and instance label. These two methods only acquire a single concept point. Unlike above two methods, DD-SVM [24] and MILES [25] both tempt to find multiple concept points, and their common idea is to transform multiple instance bag into a fix-length vector, so as to transform multiple instance learning into a standard single instance learning (supervised learning) problem. The difference between these two methods is that DD-SVM uses DD algorithm to obtain multiple concept points, while MILES simply uses all the instances in the training bags as concept points, and then uses 1-norm SVM to select multiple concept points.

Given the success of multiple instance learning, it has been directly applied to many image understanding tasks. For example, Maron [13] adopts single Blob with Neighbors (SBN) method based on diversity density (DD algorithm score) to classify images, while Zhang [26] proposes an EM-DD algorithm to speed up the efficiency of DD algorithm for image classification. Rahmani [27] proposes a multiple Instance Semi-Supervised Learning (MISSL) algorithm, where he regards MIL problem as a semi-supervised learning problem and directly uses standard Semi-Supervised Learning algorithm for image retrieval. In addition, Wang [28] also proposes a new maximum-interval multiple instance learning algorithm for image classification and annotation. Besides these methods, Hoffman [29] proposed a method of recognizing objects by combining multiple instance learning and expression learning; Li [30] and Hossien [31] proposed positive instance selection and potential kernel function method separately to solve multiple instance learning; Wang [32] proposed a new multiple instance optimization model by relaxing the labeling conditions of multi-instances; Guan [33] constructed a multiple instance learning method using auto-regressive hidden Markov model to solve the problem of behavior recognition; Wu [34] proposed a novel multi-instance bag and instance space mapping method; [35] combines events, sentiments, as well as the quantitative data into a MIL framework to predict the stock market.

Since the deep learning algorithms, especial the convolutional neural networks, has significantly improved the performance of image understanding task [36-37]. Many deep MIL algorithms have been proposed to deal with the
traditional MIL [38, 41], the Panchromatic (PAN) and multispectral (MS) imagery classification [39], image classification and auto-annotation [40]. In these methods, [38] feed the instances or bags to CNN network and proposed two kind methods: mi-Net and MI-Net with or without deep supervision or residual connections; [41] adopted a two-layered neural network to achieve attention-based MIL pooling and test the performance on MUSK, MINIST and Breast Cancer etc.; [39] proposes an end-to-end learning framework based on deep multiple instance learning by the joint spectral and spatial information fusion to classify the PAN MS imagery, while [40] models the object proposals and possible text annotations as two instance sets of MIL and use them to label images.

III. PROPOSED METHOD

In this section, we will describe proposed algorithm in detail. The proposed algorithm regards the WaDang image and Chinese characters as bag and instances in multiple instance learning.

The proposed method contains four stages which is shown in Figure 2. First, in the stage one, the method obtains the Chinese characters regions (instances) in the intersection of circles and lines. Next, we extract hand-crafted (SIFT and morphological features) and deep features by CNN and divide them into two parts in the stage two. Then, we learn the multi-scale sparse bases and multiple concept points to convert the two parts of each character region to two different fix-length vectors in the stage three. At last, in the stage four, we adopt the multiple kernel learning to fuse these two fix-length vectors and use SVM to predict the label of WaDang. In the following sections, we will explain the stage one in section III.A, stage two in section III.B, stage three in III.C and stage four in III.D in detail.

A. OBTAINING THE CHARACTER REGIONS

To obtain the character regions in WaDang, we first binarize the image, and then find the circles and lines to acquire their intersection regions.

Since the color and texture contained in the WaDang image is relatively simple, we apply threshold method to make image binarization. Assuming the input image is \( f(x, y) \) and the output binary image function \( g(x, y) \), we can obtain the following binarization formula

\[
g(x, y) = \begin{cases} 
0 & f(x, y) < \text{Threshold} \\
255 & f(x, y) \geq \text{Threshold} 
\end{cases}
\]

The steps of RCD circle detection algorithm are summarized as follows:

1. Randomly select three non-collinear pixels to generate a candidate circle.
2. Search the pixels in the candidate circle and accumulate its number.
3. If the ratio of the number of pixels on the candidate circle to the circle circumference satisfies a given condition, then the circle is considered to be a real circle.

The original image (left) and binarized image (right) WaDang containing one word "宫" are shown in Figure 3.

The Chinese characters are located in the intersection areas of two or more circles and the straight lines. Thus, the first step we obtain the Chinese character regions by detecting circles and straight lines in WaDang images.

Fast and accurate detecting circle plays an extremely important role in the field of computer vision. Hough transform [42], which is widely used in circle detection, maps the edge points in the image space into the parameter space, and then determines the circle center and radius. Because of the high computation time and storage space complexity, we adopt the random circle detection method (RCD) [43], which is faster and has stronger robustness. The RCD firstly randomly selects three non-collinear pixels to generate a candidate circle. Then search the pixels in the candidate circle and accumulate its number. If the ratio of the number of pixels on the candidate circle to the circle circumference satisfies a given condition, then the circle is considered to be a real circle.

The steps of RCD circle detection algorithm are summarized as follows:

1. Randomly select three non-collinear pixels to generate a candidate circle.
2. Search the pixels in the candidate circle and accumulate its number.
3. If the ratio of the number of pixels on the candidate circle to the circle circumference satisfies a given condition, then the circle is considered to be a real circle.
(1) Assuming all pixel in an image is \( S \), \((x_i, y_i) \in S\) is a pixel on a circle \( C \) with circle center \((m, n)\) and radius \( r\), the circle equation can be denoted as
\[
(x_i - m)^2 + (y_i - n)^2 = r^2
\]
which can be rewritten to a more general equation
\[
x_i^2 + y_i^2 = 2mx_i - 2ny_i - d = 0
\]
where \( d = r^2 - m^2 - n^2 \).

(2) In order to determine the parameters of the circle \( C \), at least three pixels \((x_i, y_i), (x_j, y_j),\) and \((x_k, y_k)\) not on a straight line are needed.

By substituting these three points into formula (3), the following results can be obtained
\[
m = \frac{2(y_j - y_i) - 2(y_k - y_i)}{4[(x_j - x_i)(y_k - y_i) - (x_k - x_i)(y_j - y_i)]}
\]
\[
n = \frac{2(x_j - x_i) - 2(x_k - x_i)}{4[(x_j - x_i)(y_k - y_i) - (x_k - x_i)(y_j - y_i)]}
\]
\[
r = \sqrt{(x_i - m)^2 + (y_i - n)^2}
\]

(3) To detect other pixels on the circle \( C \), we use formula (7) to compute the distance \( d \) between image pixel \((x_j, y_j)\) and the candidate circle \( C \). That is to say, the pixels, whose distance is smaller than a given positive \( \alpha \), are points on the circle.
\[
d = \left| \sqrt{(x_i - m)^2 + (y_i - n)^2} - r \right| \leq \alpha
\]

(4) We compute the distances between all image pixels and the circle \( C \) boundary, and then select the candidate circle \( C \) with enough parameters \( \beta \). The candidate circle is considered to be a real circle. Here the calculation formula is defined as follows.
\[
\frac{N}{2\pi r} \geq \beta
\]

The same to circle detection, as one of the most basic elements in image processing, line detection is also a very important topic in the field of computer vision. Hough transform is an effective method to deal with such problems. However, the traditional Hough transform method also needs high computation and large storage space, so it can not guarantee real-time detection. Thus, in this paper, we adopt Random Hough transform [44] (RHT), which uses many-to-one mapping instead of one-to-many mapping of traditional Hough transform, which greatly reduces the amount of computation. In addition, RHT uses dynamic linked list structure to accumulate parameters and reduces the memory requirement.

The RHT line detection algorithm steps are as follows:

(1) Generate the boundary points \( D \) of the image and initialize parameter space \( P \) (used to store two parameters , i.e. the straight line and corresponding cumulative values).

(2) Randomly select two points \((x_1, y_1), (x_2, y_2)\) from set \( D \) which satisfies the distance condition \( \frac{1}{4} \) of the shortest line length).

(3) Calculate linear parameters \( \rho \) determined by two points \((x_1, y_1)\) and \((x_2, y_2)\).

(4) Find \( \rho \) in \( P \) satisfied \( |\rho - \rho_0| \leq \delta \), where \( \delta \) is the allowable error. If there exists \( \rho \), the value of \( \rho \) accumulates 1, \( N(\rho) = N(\rho) + 1 \); otherwise, \( \rho \) will be inserted into the set \( P \).

(5) Parametric space voting: if \( N(\rho) \geq T \) and \( T \) is a predefined threshold, the corresponding \( \rho \) is a line parameter and delete all pixels in this line from set \( D \) and reset \( P \); otherwise returns to step (2).

We show the detected circles and lines of WaDang image "竹泉當" in Figure 4 and Figure 5 separately.
The Chinese characters captured from the "竹泉宫当" and "宮" WaDang images are shown in Figure 6.

![Image of Chinese characters](image)

**FIGURE 6. The Chinese characters captured from WaDang**

**B. FEATURE EXTRACTION**

In this stage, we will extract two hand-crafted features (SIFT and morphological features) and CNN feature from the obtained Chinese character region (instance). For WaDang recognition, robust features are needed to describe the Chinese character which can be achieved by combining the local, geometry and global information. It should be noted that SIFT and morphological features can acquire local and geometric features, while CNN feature can acquire global features, which just meets the requirements here.

SIFT [45] proposed by David Lowe is not sensitive to position, scale and rotation. The main idea of SIFT is to find the extremum points in the scale space and then filter them to obtain the stable ones. Finally, it extracts the local feature descriptions around each stable extremum point.

The essence of SIFT algorithm is based on the idea of image feature scale selection [44]. The SIFT algorithm consists of the following four steps: 1) detect extreme points in multi-scale space; 2) determine key points; 3) determine the main direction of feature points; 4) generate the feature descriptors.

In addition to extracting hand-crafted feature SIFT, we also extract morphological features.

Assuming the outer rectangle of the WaDang character region is $P \times Q$, $P$ the length and $Q$ the width of the rectangle, $f(x,y)$ the Chinese character region, we define below five morphological features.

1) The ratio of length and width of external rectangle $r_1 = \frac{P}{Q}$ expresses the flatness of Chinese character.

2) Pixel ratio of character to external rectangles $r_2 = \frac{\sum_{x=0}^{P-1} \sum_{y=0}^{Q-1} f(x,y)}{P \times Q}$, where $\sum_{x=0}^{P-1} \sum_{y=0}^{Q-1} f(x,y)$ represents sum of the number of pixels in Chinese character region and $P \times Q$ sum of the pixel number in the entire outer rectangle.

3) Relative center of gravity in horizontal and vertical directions $r_3 = \frac{x}{P}, \quad r_4 = \frac{y}{Q}$, where $x$ and $y$ are the horizontal center of gravity and the vertical center of gravity of the Chinese character region, respectively.

4) The ratio of longest line length and shortest line length $r_5 = \frac{\max(f(x,y))}{\min(f(x,y))}$.

5) Statistics of independent parts (number) are included in the Chinese character $f(x,y)$, which to some extent reflects the composition parts of the internal strokes of WaDang character.

Finally, in view of the powerful performance of CNN, we use CNN to extract the global feature of WaDang character.

1) Architecture: We use the pretrained VGG-16, which has total 16 weight layers, 13 layers of which are convolutional layers and the rest 3 of which are fully connected layers. The second fully connected layers are 4096 units whose output are used as features. That is to say, we can extract 4096-D feature vector for each WaDang Chinese character. To make the VGG-16 fit our task, we replace the last 1000-way fully connected layer with a 2-way fully connected layer (whether it is a given Chinese Character or not). Here, we use softmax function as the final prediction function.

2) Network Training: The network is first pretrained on ILSVRC2012 dataset and then fine-tuned on WaDang Chinese character regions obtained in section III-A. There are around 3100 WaDang Chinese characters in total which is not sufficient to train a network. Thus, the WaDang Chinese character region is resized to 256 × 256, and a fix-size 128 × 128 sub image is randomly cropped and horizontally flipped to enhance the training data. We use stochastic gradient descent to fine-tune our network with a batch size of 128, learning rate of 0.0001, momentum of 0.9, and weight decay of 0.0005. We stop our training after 30 epochs since the accuracy stops increasing.

**C. THE MULTI-SCALE SPARSE BASES AND MULTIPLE CONCEPT POINTS**

After obtaining the two hand-crafted features (SIFT and morphological features) and CNN feature of Chinese character region, we divide these three features into two parts: SIFT feature as part 1, morphological and CNN features as part 2. Then, we construct multi-scale sparse base and multiple concept points by using dictionary learning and joint dictionary learning, which will be employed to project
the part 1 and part 2 features and generate two fix-length vectors.

We first introduce the dictionary learning and then explain our multi-scale sparse bases and multiple concept points construction method.

Assuming there are a signal \( x \in \mathbb{R}^d \) and a base \( U = \{u_i\}_{i=1}^k \), \( U \in \mathbb{R}^{d \times k} \), the signal \( x \) can be expressed as the combination of base \( U \), \( x = \sum_{i=1}^{k} y_i u_i = UY \), \n
\[
Y = [y_1 \ y_2 \ ... \ y_k]^T.
\]

Sparsity means that the non-zero elements in \( Y \) is as few as possible. Thus, the sparse mathematical model can be expressed as follows

\[
\min_{Y} \|Y\|_0 \quad s.t. \quad x = UY \tag{9}
\]

where \( \|Y\|_0 \) is the L0-norm of vector \( Y \).

Since the L0-norm is a NP hard problem and not applicable in real application, Donoho et al. [46] [47] used the L1-norm to replace L0-norm in formula (9), and it has been proved that minimizing the sparse solution of L1-norm approximation to L0-norm in most under-determined equality constraint. Considering that the signal \( x \) may have noise, the equation constraints can be expressed as \( x = UY + \varepsilon \) and the optimization model formula (9) can be transformed to formula (10).

\[
\min_{Y} \|Y\|_0 \quad s.t. \quad x - UY \leq \varepsilon \tag{10}
\]

By using Lagrange function, the optimization problem of formula (10) can be expressed as

\[
\arg \min_{Y} \|x - UY\|_2 + \lambda \|Y\|_1 \tag{11}
\]

where \( \lambda \in \mathbb{R}^+ \) is regular parameter and balance the sparsity of coefficients and the approximation of signals. The optimization (11) is a convex optimization problem and can be solved by the least square method of L1 regularization.

In this paper, we adopt L1-norm optimization model to obtain our SIFT base. We assume there contains \( N \) WaDang Chinese character regions \( \{I_i\}_{i=1}^N \) of \( K \) classes and the SIFT features \( S_i = \{s_{ij}\}_{j=1}^6 \) in Chinese character region \( I_i \), where \( t_i \) is the SIFT feature number of Chinese character region \( I_i \). In order to express our model conveniently, we line the SIFT features of \( N \) images together and obtain a set \( S = \{s_i\}_{i=1}^M \), \( M = \sum_{i=1}^N t_i \).

Supposing \( s_i \) is a SIFT feature, \( U \) and \( y_i \) are sparse basis and sparse coefficient separately, our goal is to restore the \( S = \{s_i\}_{i=1}^M \) with the sparsity constraint. In this way, the L1-norm sparse model of SIFT can be expressed as

\[
\min_{U, y_i} \sum_{i=1}^M \|x_i - Uy_i\|_2^2 + \lambda \sum_{i=1}^M \|y_i\|_1 \tag{12}
\]

Then the sparse basis can be obtained by solving the optimization problem (12).

Suppose the learned sparse basis is \( U \in \mathbb{R}^{d \times N} \), where \( N \) is the dimension of base \( U \) and \( d \) the dimension of SIFT, we let \( V_0 = U \) as the first layer of multi-scale sparse model.

Then, we divide \( V_0 \) into \( t \) groups, each of whose size is \( \frac{N}{t} \) denoted as \( \{V_{0i}\}_{i=1}^t \). We again use \( \{V_{0i}\}_{i=1}^t \) as input signal, learn the sparse base \( V_i = \{V_{0i}\}_{i=1}^t \) and again consider \( V_i \) as the second layer of multi-scale sparse model.

Repeating above steps \( L \) iterations, We line all the learned \( V_i \) and obtain our multi-scale sparse base \( V = [V_0 \ V_1 \ ... \ V_L] = \{v_j\}_{j=1}^U \), where \( U \) is the dimension of \( V \).

After the multi-scale sparse base of SIFT is acquired, we then adopt joint dictionary learning to obtain the multiple concept points.

Assuming there are \( n_i \) Chinese character regions \( x_{ij} \in \mathbb{R}^d \) in WaDang image \( B_i \), we concatenate the morphological feature \( m_{ij} \) and CNN feature \( c_{ij} \) to form \( x_{ij} = [m_{ij}, c_{ij}] \). For the \( k \) class WaDang prediction task, We put all the instances in the \( i \) class together to form an matrix \( X_i \), \( X_i \in \mathbb{R}^{d \times N_i} \), \( i = 1, 2, ..., k \), and meanwhile denote the learned dictionary of class \( i \) as \( D_i \in \mathbb{R}^{d \times k_i} \), the sparse coefficient matrix \( A_i \in \mathbb{R}^{k_i \times N_i} \), where \( k_i \) is the cardinality of the dictionary and \( N_i \) is the total instance number in class \( i \). Since the core problem of multiple instance learning is ambiguity, the distance between these multiple concept points should be greater than a certain threshold \( \delta \), so we can attach optimization constraints to the joint dictionary learning model. Thus, the multiple concept point acquisition model is

\[
\min_{\{A_i, D_i\}_{i=1}^k} \sum_{i=1}^k \|X_i - D_i A_i\|^2_F + \lambda \sum_{i=1}^k \sum_{j=1}^{n_i} \|a_{ij}\|_1 \tag{13}
\]

\text{s.t.} \quad \|a_{ij} - a_{il}\|_2 \geq \delta, i \neq l
We put the learned dictionary $D_i \in \mathbb{R}^{d \times k_i}$ of all classes together to generate multiple concept points matrix
\[
D = \{d_i\}_{i=1}^{M}, M = \sum_{i=1}^{k} m_i.
\]

D. MULTIPLE KERNEL LEARNING
In the above dictionary learning and joint dictionary learning process, we have obtained the sparse base matrix $V$ and multiple concept matrix $D$. Next, we need to project the SIFT features (feature part 1), morphological and CNN features (feature part 2) to make them into two fix-length vectors, which will be further fused by multiple kernel learning.

For each WaDang image $B_i$ that contains $n_i$ Chinese character regions, we line all the SIFT features in these Chinese character regions and obtain $S_i = \{s_{ij}\}_{j=1}^{l_i}$ where $t_i$ is the SIFT number of WaDang image $B_i$. The project function is defined as
\[
\phi(B_i) = \left[ f(S_i, v_1), f(S_i, v_2), \ldots, f(S_i, v_{l_i}) \right]
\]
where \( f(B_i, v_k) = \max_{x_j \in S_i} \|x_j - v_k\|\).

Taking the similar way, the morphological features and CNN features in WaDang image $B_i$ can be transformed into a fix-length vector by using
\[
\psi(B_i) = \left[ f(F_i, d_1), f(F_i, d_2), \ldots, f(F_i, d_{M_i}) \right]
\]
where $F_i = \{x_{ij}\}_{j=1}^{w_i}$, $x_{ij}$ is the morphological features and CNN feature concatenation in Chinese character region and $W_i$ is the region number of WaDang image $B_i$.

In order to fuse the SIFT features, the morphological and the CNN features, we use multi-kernel learning to balance the impact of them on image similarity. The basic idea of multiple kernel learning is to construct new kernels by combining multiple kernels with a positive coefficient [48]
\[
k = \sum_{i=1}^{l} d_i k_i
\]
where \( \sum_{i=1}^{l} d_i = 1 \), \( 1 > d_i > 0 \), \( k_i \) is a kernel function satisfying Mercer’s theorem.

For our WaDang recognition task, suppose $l = 2$ in the formula (16), \( K_{RBF} (\psi(B_i), \psi(B_j)) \) is the kernel function of SIFT features of WaDang Image $B_i$ and bag $B_j$, \( K_{RBF} (\phi(B_i), \phi(B_j)) \) is the kernel function of morphological and CNN features, the kernel function of WaDang image $B_i$ and $B_j$ can be defined as
\[
K(I_i, I_j) = \alpha K_{RBF} (\phi(B_i), \phi(B_j)) + (1 - \alpha) K_{RBF} (\psi(B_i), \psi(B_j)),
\]
which integrate the hand-crafted and deep features similarity together, and finally feed to support vector machine to predict images.

IV. EXPERIMENTAL RESULTS AND ANALYSIS
We evaluate the performance of our method based on three datasets, including our collected WaDang dataset, and two traditional MIL datasets (MUSK and COREL). In the first experiment, we introduce our WaDang dataset and then compare our method with other state-of-art algorithms including SMILE [16], MILES [25], Attention and Gated-Attention [41], mi-Net and MI-Net [38]. We further adopt Musk and Corel to validate our algorithm on drug activity prediction and image classification, whose results proved the effectiveness of our algorithm in standard multiple instance learning.

A. WA-DANG DATASET
We cooperated with Qin Brick and Han Tile Museum in Xi’an to acquire the WaDang images by using high definition camera. The collected WaDang dataset is a small dataset which contains 843 WaDang images in total. There are 76 one Chinese character, 128 two Chinese characters, 15 three Chinese characters, 560 four Chinese characters, 22 five Chinese characters, 5 six Chinese characters, 5 seven Chinese characters, 14 eight Chinese characters, 6 nine Chinese characters, 5 ten Chinese characters and 11 twelve Chinese characters. Among these, the four words’ proportion is 50% and there are 2998 different Chinese characters in total. The image size is 400 \times 400 pixels. The WaDang images are labeled by cultural relics experts. The Chinese character region is about 82–252 pixels wide and 98–232 pixels high. Some selected samples of our WaDang images are shown in Figure 7.

![FIGURE 7: Some selected samples of our WaDang images](image)

To verify our algorithm, we split the WaDang dataset into two sets: WaDang1 and WaDang2, in which WaDang1 consists of two and four characters’ images and the WaDang2 are the other images not in WaDang1. Here, we use ten cross validation methods to test the performance. That is to say, we divide each class of data set into 10 parts, using 9 of them as training data and the rest as test data for experiment. The training set is further split into training and
validation sets. In the experiment, we choose "one-to-one" method to deal with multiple classes classification.

Table 1 shows the results of comparison between our method and SMILE [16], MILES [25], Attention and Gated-Attention [41], MI-Net with DS and MI-Net with RC [38]. For Attention and Gated-Attention, mi-Net and MI-Net with DS and MI-Net with RC, we only use CNN to extract the deep features and train deep neural networks. The other algorithms [16] [25] adopts the same methods which we have described in this upper section to deal with the SIFT features. That is to say, the base of SIFT is obtained by dictionary learning and the SIFT sets of Chinese character regions are projected to the base to obtain fix-length vectors. Then we concatenate the obtained fix-length vectors, morphological and deep CNN features and feed it to corresponding algorithms.

The best performance of WaDang1 and WaDang2 datasets are bolded in Table I. As can be seen from Table 1, our algorithm achieves the best performance in WaDang image prediction, which is nearly 4% higher than that of SMILE, MI-Net with DS, Attention and Gated-Attention. These results demonstrate the effectiveness of our deep feature fusion Multiple Instance Learning. The reason why our approach is more effective than SMILE [16] and MILES [25] is that our method can capture the SIFT base and MIL multiple concepts, which can express the images more efficiently. And, for mi-Net and MI-Net, our method fuses the hand-crafted features and deep feature, which can make better use both of them.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SMILE [16]</td>
<td>78.3%</td>
</tr>
<tr>
<td>MILES [25]</td>
<td>69.6%</td>
</tr>
<tr>
<td>Attention [41]</td>
<td>78.3%</td>
</tr>
<tr>
<td>Gated-Attention [41]</td>
<td>78.6%</td>
</tr>
<tr>
<td>mi-Net [38]</td>
<td>75.3%</td>
</tr>
<tr>
<td>MI-Net with DS [38]</td>
<td>78.1%</td>
</tr>
<tr>
<td>MI-Net with RC [38]</td>
<td>77.6%</td>
</tr>
<tr>
<td>Ours</td>
<td>82.1%</td>
</tr>
<tr>
<td>WaDang1</td>
<td>75.8%</td>
</tr>
<tr>
<td>WaDang2</td>
<td>76%</td>
</tr>
</tbody>
</table>

We compare our method with SMILE [16], MILES [25], Attention and Gated-Attention [41], DD-SVM [24], and MI-Net with DS [38] algorithms. The comparison results are shown in Figure 8. As can be seen from the Figure 8, our method far exceeds the traditional methods and deep learning method MI-Net with DS, and are basically consistent with the performance of SMILE.

C. COREL DATASET

In addition, we further test the performance of our algorithm on multiple instance learning image data sets COREL. The data sets include 20 objects, including African people and villages, beaches, historic buildings, buses, dinosaurs, elephants, flowers, horses, mountains and glaciers, food, dogs, lizards, fashion models, sunset scenes, cars, waterfalls, antique furniture, warships, skiing and deserts. The 2000 images in the COREL are divided into 20 categories, each of which contains 100 images.

Table 3 is a comparison of our algorithm with other algorithms on the datasets. The results show that our algorithm also achieves good performance on COREL datasets. Compared with other multiple concepts point methods, such as MILES [25], DD-SVM [24], Attention and Gated-Attention [41], and MI-Net with DS [38], our algorithm achieves the best performance. The reason is that we adopt multi-class joint dictionary learning method to find the optimal sparse basis as the conceptual point by fully considering the characteristics of multiple instance learning.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MILES [25]</td>
<td>68.7 [64.4 72.8]</td>
</tr>
<tr>
<td>Attention [41]</td>
<td>74.6 [72.9 76.3]</td>
</tr>
<tr>
<td>Gated-Attention [41]</td>
<td>76.7 [75.2 78.2]</td>
</tr>
<tr>
<td>DD-SVM [24]</td>
<td>67.5 [63.9 71.1]</td>
</tr>
<tr>
<td>MI-Net with DS [38]</td>
<td>76.6 [73.4 79.8]</td>
</tr>
<tr>
<td>Ours</td>
<td>78.1 [75.0 81.2]</td>
</tr>
</tbody>
</table>

V. Conclusions

In this paper, we propose a deep feature fusion multiple instance learning for WaDang image prediction. In the case that fully understand the characteristics of the WaDang image, this paper uses the symmetry of the WaDang to obtain the Chinese character regions and extract its SIFT, morphological and deep features in the intersection of circles and lines. Then, the SIFT base is obtained by multi-scale sparse method, and the multiple concept points are obtained by joint dictionary learning method. Finally, the similarity between hand-crafted features (SIFT and morphological feature) and deep feature is fused by multiple kernel learning. In the experiments of WaDang image, MUSK and COREL datasets, the algorithm has achieved good performance.

The algorithm in this paper depends on the learning performance of sparse dictionary. One feasible follow-up research method is to find more effective sparse dictionary and joint dictionary learning methods.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2929837, IEEE Access
REFERENCES


FIGURE 8. The comparison between our method and other methods on MUSK