IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received May 18, 2019, accepted July 2, 2019, date of publication July 18, 2019, date of current version November 25, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2929816

An Intelligent Knowledge Extraction Framework for Recognizing Identification Information From Real-World ID Card Images

LIN ZUO[1], WENYU CHEN[2], HONG QU[2], LI HUANG[2], ZHENG WANG[2], AND YONG CHEN[3]

[1]School of Information and Software Engineering, University of Electronic Science and Technology of China, Chengdu 610054, China
[2]School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 610054, China
[3]School of Mechanical and Electrical Engineering, University of Electronic Science and Technology of China, Chengdu 610054, China

Corresponding authors: Lin Zuo (linzuo@uestc.edu.cn) and Wenyu Chen (cwy@uestc.edu.cn)

This work was supported in part by the National Science Foundation of China under Grant 61877009 and Grant 61573081, and in part by the Sichuan Provincial Science and Technology Plan Project under Grant 2018GZ0396.

ABSTRACT In this work, we study the problem of recognizing identification (ID) information from unconstrained real-world images of ID card, which has extensively applied in practical scenarios. Nonetheless, manual ways of processing the task are impractical due to the unaffordable cost of labor and time consumption as well as the unreliable quality of manual labeling. In this paper, we propose an intelligent framework for automatically recognizing ID information from images of the ID cards. Specifically, we first conduct marginal detection using a multi-operator algorithm and then localize the region of ID card from all the proposed candidate regions with SVM classifier. Furthermore, we segment linguistic characters from the card region by an improved projection algorithm. Finally, we recognize the specific characters by an eight-layer convolutional neural network. We perform extensive experiments on a Chinese ID card dataset to validate the effectiveness and efficiency of our proposed method. The experimental results demonstrate the superiority of proposal over other existing schemes.

INDEX TERMS Identification information recognition, intelligent framework, convolutional neural network.

I. INTRODUCTION

In recent years, extracting textual information from images has received considerable concerns due to the rapid development of information technology which requires an enormous growth of accessible information. The explosive growth of smart device has been improving the quality of our daily life. For instance, we may take photos by the camera of cell phones to record and share many significant occasions and information (e.g., vital document, number and street). However, extracting text information like personal details, iconic numbers printed on smart cards (such as, ID card, and credit card) is a challenge. Those information are commonly used in banks, airports, security company and other places where high accuracy in identification and recording is required. Traditionally, this task can be completed by manual identifying and inputting digital information into system. However, the reliability and efficiency cannot be guaranteed. The rapid development of machine vision techniques enable automatic recognizing information from images with certain additional equipment. However, some challenging issues still remain to be addressed:

- Constrained place: Usually, the additional equipment requires the ID cards to be processed in some fixed position with the assistance of professional technicians to extract the information. In other words, customers have to personally go to a specified location to have theirs ID cards being processed. Undoubtedly, this way is quite time-consuming and the user experience is very poor, especially when the operation is urgent, such as stock exchange.
- Strict environments: The additional equipment is highly sensitive to illumination environment, too bright or too weak light may affect the recognition results, which inevitably requires expensive cost of professional technician to assist the operation.
- Recognition accuracy: Due to the scarcity of sufficient real-world training images and effective recognition approaches, existing identification systems cannot guarantee the accuracy of recognition. Moreover, the robustness of the performance should be further enhanced too.

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

In a nutshell, a highly accurate approach is demanded from bank, financial institution and corporation to realize the timeliness personal business for exacting customer information from images, enabling customers to perform the task as quickly as possible by themselves. It is similar to the vehicle license plate recognition (LPR) technology which processes natural images photographed by outdoor cameras without interventions from staffs. Oftentimes, these images are not readily to be processed due to complex background and a wide range of illumination conditions. The challenging issue of the LPR is to localize the license plate, which is similar to the ID card recognition problem in this study. The purpose of our study is to promptly segment characters from images with various backgrounds and illumination conditions [1].

Unlike the LPR problem in which a high quality camera is used, the problem to be solved is much more complex than that of LPR, because images could be taken by a diversity of phone cameras with different qualities [2], [3]. After we locate the ID card region, another important stage is the character recognition which typically uses a statistical pattern matching method [4]–[6]. This method, however, is sensitive to diverse images datasets, noise and the robustness of a particular learning algorithm [7], [8]. Another method which is extensively used for pattern matching is the deep artificial neural networks. Deep artificial neural networks are an important advance in solving recognition problems [9], [10]. It excels at discovering intricate structures and requires a very few assistance from human beings [11].

In [12], the BP neural network was used to recognize the Chinese ID card numbers and the ability of ID card number recognition was greatly improved. However, the study only focused on the ID card numbers while ignoring other text information that are equally important. In [13], an approximative Bayes optimality liner discriminant analysis (BLDA) was presented for Chinese handwriting character recognition, this model reduced the searching spaces of BLDA significantly and acquired good accuracy of recognition. The proposed model not only recognized the numbers but also successful recognized the Chinese characters on ID cards which have a great difference with handwriting characters. It is worth mentioning that Google has an open source system, namely Tesseract, which has been widely used for text recognition by researchers for western characters recognition [14]. Unfortunately, Tesseract does not has a good performance for Chinese characters recognition.

In this paper, an intelligent framework for automatically extracting ID information from images of ID cards is proposed, the flowchart of framework is illustrated in Fig. 1. By using a multi-operator algorithm, the marginal of ID cards are detected, and the region of ID card is located by the SVM classifier. An improved projection algorithm is employed to segment the linguistic characters from the card region. A tailored eight-layer convolutional neural network is used to recognize the specific characters. A set of experiments are conducted to examine the effectiveness and efficiency of the proposed method. The results demonstrate that the proposed method can improve the efficiency in Chinese ID card recognition.

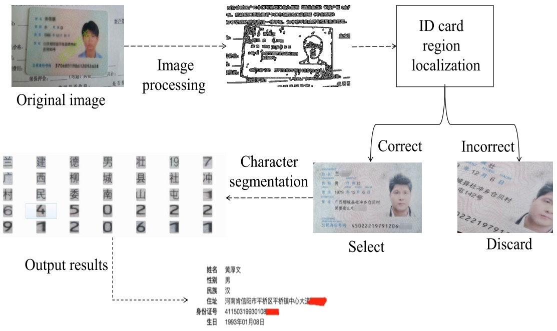


FIGURE 1. The flowchart of our proposed scheme.

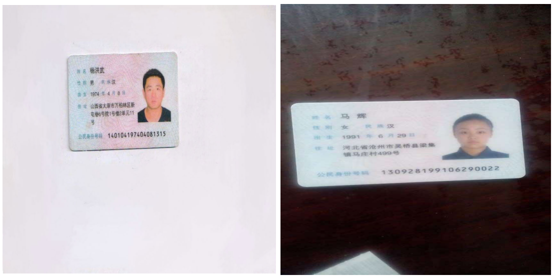


FIGURE 2. The raw images usually have many problems. For example, the useful information of the left image is less than 40%. The right one is a distorted image as the camera is not vertically aligned to ID card.

II. LOCALIZING ID CARD REGION

A universal ID card recognition project should have a good ability to obtain highly accurate ID card regions even in complicated situations. As the quality of images taken by different people may be the same, we learned the idea from the question of dehaze images, including image enhancement-based approaches and image restoration-based approaches [15], [16]. As images may be taken by different people, various image issues are to be expected including a copy of ID card with a large proportion of empty space as shown in Fig. 2 (left): where the useful information in the picture is less than 40%, or may be a photo taken by a mobile phone or other handhelds as depicted in Fig. 2 (right): depending on how the handheld device was held, the image is not vertically aligned to ID card. The distortion may directly affect ID card identification. These reasons among others render the reasons why original image must be preprocessed to acquire exact ID card region from the original images.

We use three steps to remove noise and obtain an accurate position of ID card. In the first step, the foregrounds and backgrounds should be separated and the color images are binarized to remove any form of noises. The multi-operators is used to detect the edge of ID cards. The multi-operators algorithm will sort out all eligible regions which have rectangle block, also called candidates. In the second step, an SVM model will be trained to classify all the candidates and filter out the obvious incorrect regions. In the third step, a confidence algorithm is used to calculate the most likely ID card region. The comprehensive details are shown below.

A. DENOSING

In order to remove the noise and make the image smoother, we first use the Gaussian low-pass filtering and image binarization which can help eliminate noises and interference. Filtering is the most basic operation of image processing. In the broadest sense of the word ''filtering'', the value of a filtered image at a given location is a function of the values of the input image in a small neighborhood of the same location. The Gaussian low-pass filtering computes a weighted average of pixel values in the neighborhood [17]. By the Gaussian low-pass filtering, the noise in the images can be significantly removed. It is worth mentioning that even though the binaryzation is powerful for denoising, it has its own drawbacks. The most important one among many drawbacks is the lose of useful color properties if images are binarized.

The kernel of size $(2k+1)*(2k+1)$ Gaussian filter function is given by:

$$G_{xy} = \frac{1}{2\pi\sigma^2} e^{-\frac{(x-(k+1))^2+(y-(k+1))^2}{2\sigma^2}}$$
$$(1 \leq x, y \leq (2k+1))$$

We input images, and the Gaussian kernel is used for the Gaussian filter to convolve images, so as to make images smooth and eliminate the noise and interference.

B. MULTI-OPERATOR IN EDGE DETECTING

Using the multi-operator to detect edges is the core of our proposed method. Without this stage we cannot obtain the marginal information of ID card for the ensuing segmentation and recognition.

Sobel operator: The Sobel operator is a famous detection algorithm in the field of image processing and machine vision [18], and it can create image with emphasized edges. It uses two odd number matrices, one for horizontal changes and the other for vertical changes, as kernels for convoluting the original image to calculate approximations of the derivatives. Compared to other edge operator, the Sobel has two main merits: it adds some smoothing effect to the random noise of the image. Because the Sobel is the differential of two rows or two columns, the elements of the edge on both sides has been enhanced [19].

$$G_x = \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix} *A$$

$$G_y = \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{bmatrix} *A$$

where A is the source image, Gx and Gy are two images of which each point contains the horizontal and vertical derivative approximations respectively, $*$ denotes the 2-dimensional signal processing convolution operation. We then calculate the gradients direction as follows:

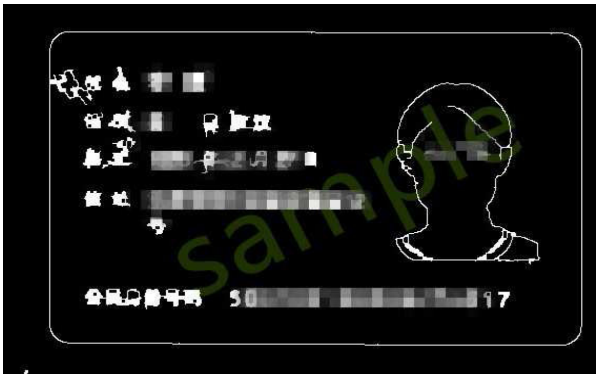


FIGURE 3. Intermediate result of the multi-operator algorithm which can acquire exact marginal information. The purpose of adding mosaics is to protect the sensitive information of the ID card.

$$\Theta = atan(\frac{G_y}{G_x})$$

where Θ is 0 for a vertical edge which is lighter on the right side, Θ is π for the anther case.

Canny operator: The Canny edge detector uses a multi-stage algorithm to detect a wide range of edges in images and it's optimal at any scale [20]. It has a simple approximate implementation in which edges are marked at maxima in gradient magnitude of a Gaussian-smoothed image. Empirical experience in the area of machine vision recommends that the canny edge detection provides good and reliable detection.

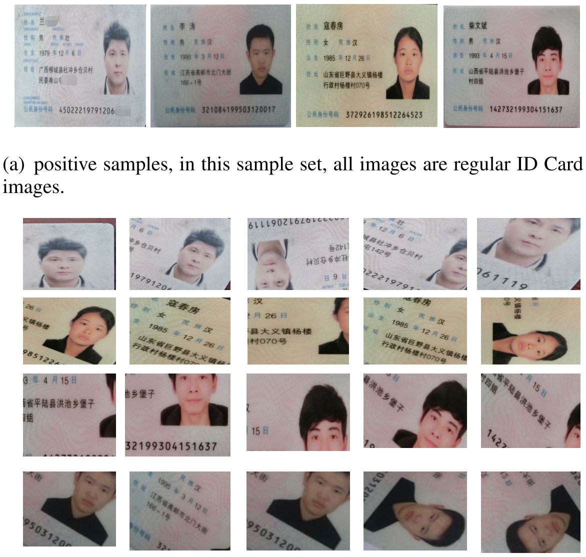
We used multigroup parameters of the Sobel operator and Canny edge detector together. Multigroup parameters and combined results ensure that the algorithm can acquire the exact marginal information. In fact, experiments show that the results of combined of the two operators outperform any single operators as shown in section V.

C. DISCRIMINATION THE CANDIDATES

There are several candidate images after executing the multi-operator algorithm, most of which are not very useful. We therefore classify them. In this study, we use the Support Vector Machine (SVM) model and the confidence algorithm for selecting the useful images before proceeding the next step. The Support Vector Machine (SVM) [21] is a class of machine learning model used for classification and regression. Given a set of labeled training samples, each marked sample belongs to one of two categories: positive or negative, the SVM learns from those labeled samples to classify a new image into the two classes. The results of a trained SVM is a classification model that can be used to classify new examples.

The SVM exhibits excellent generalization performance (accuracy on test sets) in practice and have strong theoretical motivation in statistical learning theory [22].

In our case, we have 1200 positive and negative samples, respectively. The positive samples are all regular ID card images with complete information, whereas the negative



(a) positive samples, in this sample set, all images are regular ID Card images.

(b) negative samples with incomplete images.

FIGURE 4. The training samples for the SVM model.

samples are various kinds of irregular images with incomplete information as shown in Fig. 4. To ensure the correctness of the final result and reduce the computing complexity, we use the confidence algorithm to verify the results produced by the SVM and preserve final images which satisfy the criteria. Since we used multigroup parameters and combined the results of two operators, we may obtain more than one final image. Usually, we believe that they are all correct ID card images and we select directly the first one for the next step.

D. IMAGE CORRECTION

Since we use the real-world ID card data, many images are tilted and distorted, and they require corrective actions. We adopt corner detection which is widely used in the field of machine vision to acquire images features such as image location, video tracking, and target recognition, to correct these images. In our system, the corner detection focuses on the edge of ID card's profile and analyze the distribution of corner points.

After the image's corner points are acquired, we find the inclination of the image by using the Hough Transform [23], which is high accuracy transform algorithm to detect object's shape, such as detecting straight lines. Because the ID cards have high linearity, we can easily repair tilted images by figuring out their angle of the inclination as shown in Fig. 5.

III. SEGMENTING LINGUISTIC CHARACTERS

To recognize the ID card, the characters and numbers must be separated out from the image of ID cards, the performance of separation directly affect the accuracy of recognition. In this study, the segment method is based on the row scanning of the original ID card image along with the projection information from the vertical scanning.

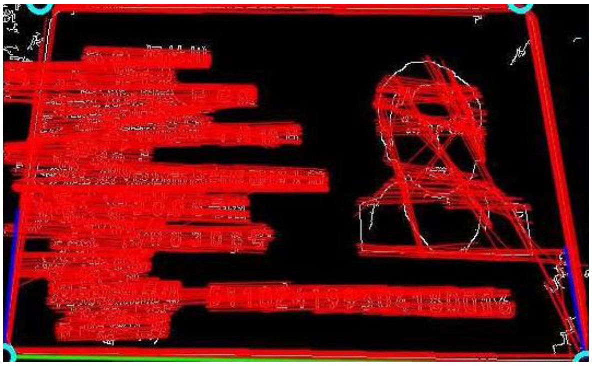
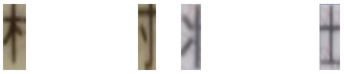


FIGURE 5. Correcting skew image by checking those lines.



(a) The left character is:村 (b) The right character is:壮

FIGURE 6. The problem of naive project algorithm used on Chinese character, the algorithm always separates the left-right structure character.

The segmentation falls into two types: the line segments and the character segments. The line segmentation uses the relative position of the ID card image because ID cards have uniform sizes and styles. In China, every ID card has the same layout with the facial photo being on the right side and the personal information on the left side. The position, size, and line space of the images are all fixed. According to this characteristic, the text message can be roughly determined by using the position of detected pixels where ID card's exact region in image can be obtained.

The characters segmentation uses vertical projection information in this study and it is modified to accommodate the complex structure in Chinese characters. The vertical projection requires that images are binary before segmentation, and the vertical projection can be calculated on every column pixels based on binary image. The naive projection works well for Latin characters. As letters are next to each other, and the projection algorithms are efficiently used. However, Chinese characters differ from western words. Chinese characters have their particular structure, it may have a gap even if only a word, such as Chinese character in Fig. 6. When separating Chinese characters, the naive projection algorithm always produces a recognition error. To solve this problem, we improve the projection algorithm to adapt to Chinese character segmentation. Experimental results show that the improved projection algorithm is much more accurate for Chinese character segmentation.

1) The width and space of numbers and Chinese characters should be recorded and treated as the threshold of segmenting criterion.
2) When the algorithm start scanning, it adds minutes of the start position sCH and end position eCH of character $n(n = 1, 2, \ldots)$, then judge whether the (sCH, eCH) is greater than the threshold:

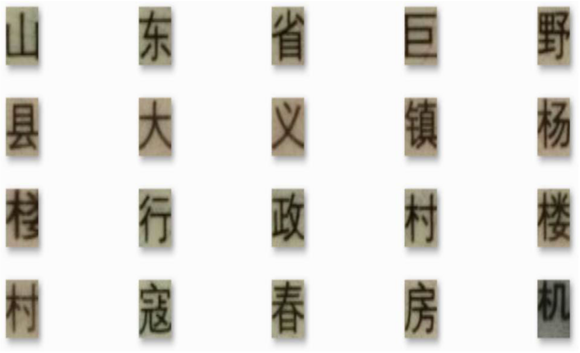


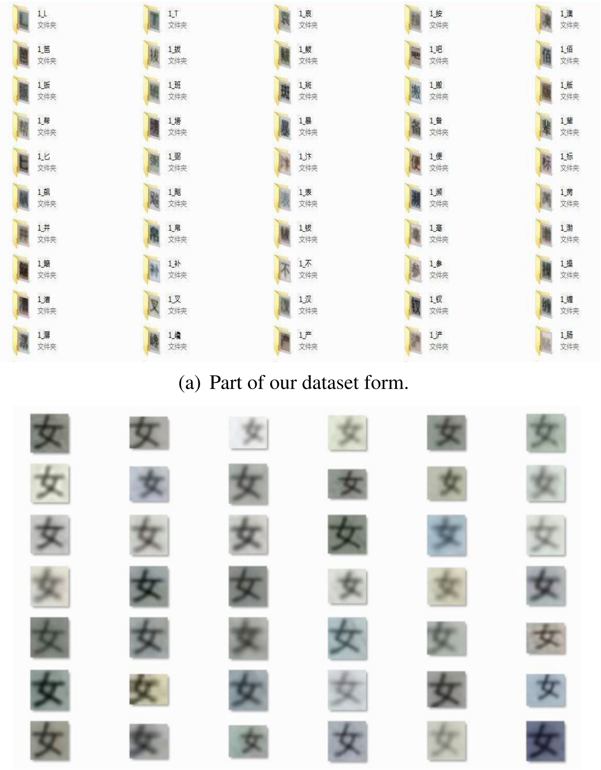
FIGURE 7. The results from characters segmentation.

- If $eCH - sCH$ is very close to a pre-specified threshold, we believe that it is an acceptable segmentation.
- If $eCH-sCH$ is less than a pre-specified threshold, it means there may be a Chinese word which have a division like left-right structure and was segmented incorrectly. The algorithm should merge the current region and the next right region, and then discriminate with the threshold again. Usually, the process will be repeated twice through the algorithm to get a correct segmentation.
- The last situation is that $eCH - sCH$ is greater than threshold, and it means there that may be an adhesive character to be segmented again. By using the physical position to segment text line and the modified vertical projection method to segment characters, the accuracy of segmentation increases significantly. It can process some cases of Chinese characters such as the structural problems and the adhesion characters problems. The results of segmentation is shown in Fig. 7.

IV. CHARACTER RECOGNITION

Machine learning technologies are powerful in many aspects of modern society [9], such as detecting object in images [24], audio recognition [25], path-planning for mobile robots [26] and optimize classic question [27]. Currently, the common form of machine learning is supervised learning. To model a supervised learning of object identification in image, there is a need to collect a large dataset of the object we called ''label'', then use the ''label'' to correct the model's output. Lots of studies have demonstrated that the supervised learning is very useful in solving recognition problems. Recently, supervised learning models are increasingly being adopted in practical applications, especially in image recognition and natural language process-the accuracy of neural network model base on machine learning technology of handwritten digit corpus MNIST is greater than 98% [28].

Although machine learning is very successful in western characters and digit recognition, limited researches effort have been made on Chinese characters recognition, because it needs a relatively large data to train models for Chinese characters. In fact, the common characters in Chinese are more than 5000 which are considerably larger than the western characters. The core problem that to acquire enough character labels to train the model remains.



(a) Part of our dataset form.

(b) A folder of a word includes many segmentation of this word that cutout out from many ID card images to enlarge the diversity of our dataset.

FIGURE 8. Our dataset layout.

We utilize image post-processing to edit real ID card images, replace those words on card with ''raw'' words which not exist in training dataset and ignored syntax and grammar. It is worth mentioning that, we chose to edit real ID card images instead of creating image of a word directly because the words on ID card have their own font, size and style. We do so to make sure the artificial word data is consistent with the real word so that it can be useful to train our model. Figs. 9 and 10 show the ''man-made'' image and the segmentation results, respectively. Our training dataset was enlarged to 3000 words by our approach. Our experimental results show that the recognition accuracy of our model has improved tremendously since our training dataset increased by artificial data.

The deep learning technology has widespread applications, from web searches to content filtering on social networks to recommendations on e-commerce websites [9]. In speech recognition, using the deep learning structure can achieve record-breaking results on a standard benchmark in a small vocabulary [29] and quickly develop on a large vocabulary [30]. There are many kinds of neural network models, the Convolution Neural Network (CNN) is one among many. The CNN is designed to process data that come

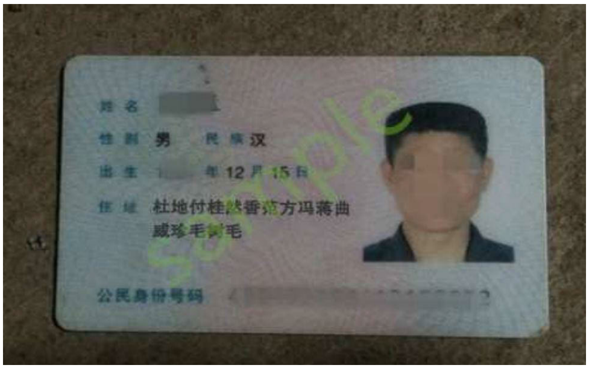


FIGURE 9. In order to enlarge our training dataset, we used "raw" word that not exist in our earlier training dataset to "create" some "artificial" ID cards and ignored syntax and grammar.

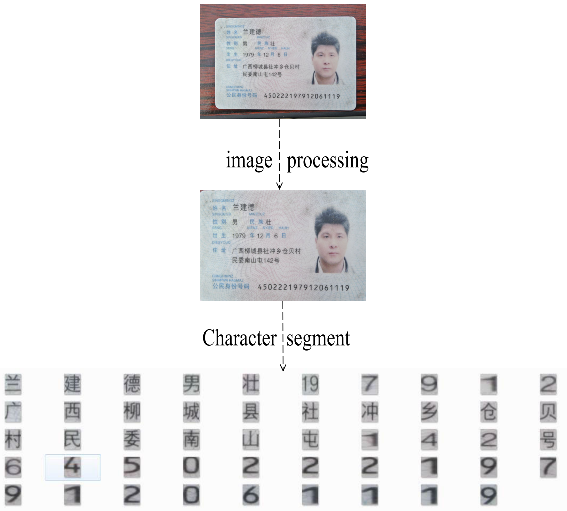


FIGURE 10. The segmentation result of an "artificial" ID card, the segmentary effect and the pieces (samples) are the real data when we train our model. The recognition accuracy has been significantly increased since these artificial samples were added into the train set.

in the form of multiple arrays like an image composed of arrays containing pixel [31]. It is a particular type of deep, feedforward network that is much easier to train and generalized than networks with full connectivity between adjacent layers [32], [33]. The CNN has many practical achievements in image processing and widely used in the field of machine vision. In the early 1990s, the CNN model has used for face recognition. Now, CNNs are the dominant model for recognition and detection task [9], [34], [35].

A CNN model has a series of layers. Most of CNN models have one convolution layer and one pooling layer followed to obtain features and a fully connected layer as hidden layers, then a classify layer to output, all hidden layers are free to stack. To obtain better features, our CNN model, with twice stacked two conventional layers and one max-pooling layer, is different with conventional CNN model structure. The CNN has fully connected layers for nonlinear classification. The inputs of the model are the images of Chinese characters and the outputs are the corresponding recognition results. The architecture of our model is shown in Fig. 11.

Convolutional layer: Generally, a classical CNN model with one convolution layer followed by a max-pooling layer should perform well when obtaining features. Empirical experiences suggest that deep architecture with multiple layers could achieve a good performance. As the convolution layers play critical roles in this study, we add one more convolution layer before max-pooling layer to extract deeper and significant information. The input of a convolutional layer is a feature map $\{z_i = (u, v) \in \Re^d\}$ where $(u, v) \in \Omega_i$ are image coordinates and d denotes d scalar features. The output is a new feature map z_{i+1}, such as:

$$z_{i+1}^k = g_i(W_{ik} z_i + b_{ik})$$

where W_{ik} and b_{ik} denote the k-th filter kennel and bias respectively. The $g(\cdot)$ is a nonlinear activation function. In our study, the Rectified Linear Unit (ReLU) [36] is used, and we will introduce it later.

Max-pooling Layer: We conduct a max-pooling layer: $h_i = \max z_i^2(u, v)$ after two convolution layers to capture the most important feature, where 2 denote the max-pooling that uses the feature map of the second convolutional layer. Max-pooling can maintain the translation invariant of image, it means that the same feature will be active even the image undergoes translations. It is very useful because our images come from the real world photos, it always have a litter left or right translation or tilt, but we need the model still accurately classify it regardless of its position.

Fully-connected Layer: We conduct a fully-connected layer $h = [h_1, \ldots, h_k]^T$ for classification purposes, where k is the number of filters.

In addition to the above layers, the neural network is composed of input layer and output layer. The input layer is the Chinese character images we introduced before, while the output layer is a softmax layer which we used to classify characters and then output the recognition results [37]. Inspired by Nature Language Processing (NLP) tasks, such as language model [38], [39], machine translation [40], [41], we used the vocabulary probability as our output. In this approach, the dimension of the softmax layer is the same as the vocabulary size, representing the probabilities of the corresponding words. Although the number of parameters in the layer of softmax is huge, this method is a promising way to predict words in NLP tasks. We present and analyze the experiment results in Section V.

V. EXPERIMENTS

In this section, we present and discuss our experiments results.

To handle the data problem, firstly, we search available dataset on Internet. However, it does not work because the ID card has its peculiar fonts and those available datasets are unsuitable. Besides, there are very limited references on Chinese ID card identification. Although some companies have developed their ID card identification SDK for limited use and commercial applications. However, as the principles and algorithms of ID card are not either publicized or reported

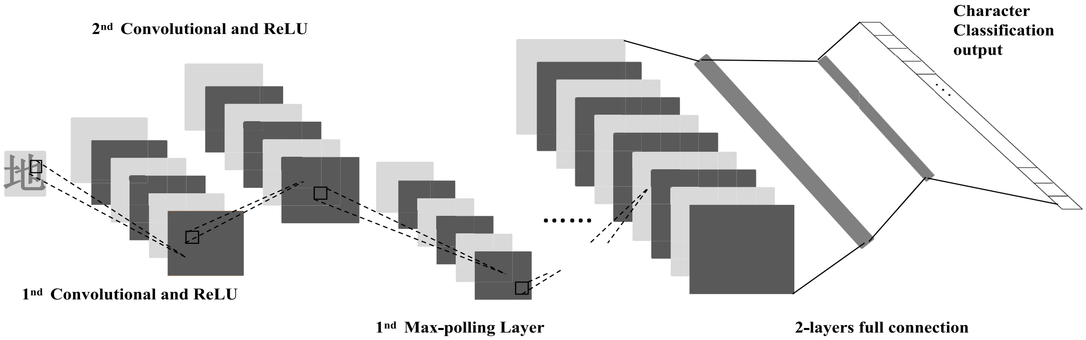

2nd Convolutional and ReLU
Character Classification output
地
......
1nd Convolutional and ReLU
1nd Max-polling Layer
2-layers full connection

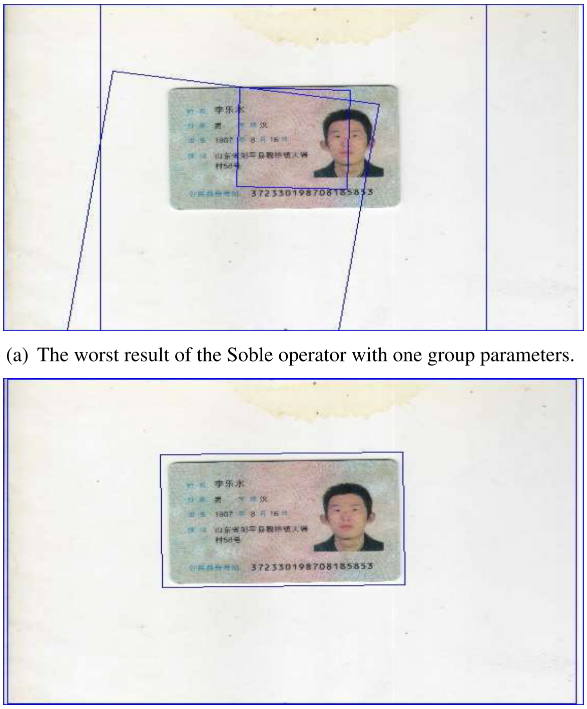
FIGURE 11. The model architecture.

in the literature, we cannot make a full comparison with these methods. To solve this issue, we create a train dataset by ourselves. We use method mentioned in Section III to obtain many character fragments, then we classify them by hand. This set has 1000 labels and more than 10000 pieces, each label has a folder to store pieces belonging to the label, tens of images samples per one label, including 10 numbers and a capital character ''X'' which is a special symbol in China ID card. Each character sample is cut out from real-world ID card image so that our model can get very accurate results. The dataset is shown in Fig. 7.

However, after experiments, our model could not work well with our trained dataset because there were still many characters that could not be identified. We checked the intermediate results and found out that those characters did not exist in the training dataset. In other words, the dataset was still too small to work well for identification. Unsurprisingly, as I mentioned earlier, the common Chinese characters are more than 5000 words, our training dataset was only one-fifth. To overcome this barrier, we try a ''man-made'' method to extend our training dataset.

A. MULTI-OPERATOR EXPERIMENTS

Generally, the photos of ID cards are instantly photographed, we should first locate the ID card in picture. As mentioned Section II, we use multi-operator to obtain the ID card profile and then to acquire the region of ID card in an image. Initially, we used only the Sobel operator to detect the ID card profile in an image, but the results may be unstable. As a variety of images will be collected, a single operators may not be able to locate the ID card in each individual image. The worst result is shown in Fig. 12(a), the location of ID card was not identified correctively. The problem happens because the ID card's shape in images are diverse as we said before, so we combined the Soble operator and the Canny operator and give five group parameters for dealing with complex situation. As expected, the profiles of most ID cards can be extracted in this way, the result is shown in Fig. 12(b).



(a) The worst result of the Soble operator with one group parameters.

(b) Five group parameters used the Sobel and Canny operator can obtain most of ID card's profile in images.

FIGURE 12. Multi-operator experiments.

TABLE 1. The performances of three activation functions, the higher the score, the better the model.

	Activation Functions	Recognition Rate
1	Sigmoid	76.0
2	Tanh	97.2
3	ReLu	98.1

B. CHARACTER RECOGNITION EXPERIMENTS

We present our recognition experiment from the convolution neural network model in this section. Because our model will be used in some practical applications, we invite humans to evaluate our model with human perception of the recognition quality.

Dataset and Evaluation: As mentioned before, we use two datasets to test our model, the small one includes 1000 Chinese character image samples while the largest set includes more than 3000 image samples. We carried out a human evaluation and the reference label evaluation, the human evaluation is the human perception of the model's recognition

TABLE 2. **The performances of different structure of hidden layers. The 1,2 models are basic models, *d* denotes the basic model stack times, 3,4 models are twice stacked model based 1,2 model respectively, 5 model is three times stacked model 2.**

	Hidden Layer	Number of Hidden Layers	Dimension of Features	Recognition Rate
1	1-convolutional layer	3	7200	90.30
2	2-convolutional layers	4	6272	94.30
3	twice 1-convolutional layer($d = 2$)	5	3136	94.70
4	twice 2-convolutional layers($d = 2$)(Our model)	7	2304	98.90
5	three times 2-convolutional layers($d = 3$)	10	256	99.10

quality. We asked human rater to rate the recognition results in a side-by-side comparison way.

Training details: Our character recognition moder uses a seven-layer CNN model to output a probability of $p(c|x)$ over our dataset alphabet C including 122/300 words, 10 digits and a special capital letter "X", giving a total of 133/311 classes. The input $\{z_1 = x\}$ of the CNN are gray-scale clipped character images of 30 × 30 pixels, the images is convolved with 64 filters of size 3 × 3 and pooled with activation function in size of 2 × 2. The loss function is a cross entropy function and all the parameters of the model are jointly optimized to minimize the loss function over a training set using Stochastic Gradient Descent (SGD) and back-propagation.

Because recognition accuracy of neural networks is affected by using different activation functions, we compare our model with three activation functions:

- Sigmoid function: a bounded in [0,1] and differentiable real function which is computed as: $sigmoid(x) = \frac{1}{1+e^{-x}}$.
- Tanh (Hyperbolic function): Tanh function is a solution to the nonlinear boundary value problem, bounded in [−1,1], and it is computed as: $\tanh(x) = \frac{e^{-x}-e^{x}}{e^{x}+e^{-x}}$.
- ReLU (Rectified Linear Unit) [42]: A simple function expressed as: $f(x) = \max\{0, x\}$. Actually, the ReLU function becomes a standard nonlinear activation function of CNN, because it provides a simple calculation and significantly reduce the number of iterations.

We use 100 ID card images to test the accuracy of the three activation functions with the same parameters as our model. Table 1 shows the performances of the three activation functions. Appearently, the ReLU function improves the recognition quality in all the cases. Therefore, we chose the ReLU function as the model's activation function.

The performances of neural network also depends on the structure of hidden layers, it can be greatly influenced by the number of hidden layers. Using a conventional convolution neural network always has a pooling layer after one convolution layer. We test some combinations of model for high accuracy. All cases of combinations used 100 ID card images to test as the activation functions test, the result shown in Table 2. Although the three times stack CNN structure acquires higher accuracy, taking account of the amount of time consumption, we chose the twice stacks CNN structure.

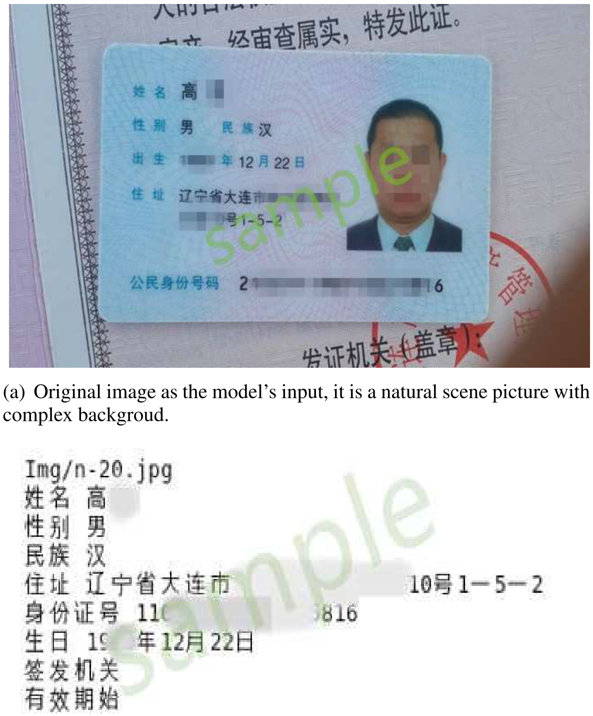
According to the above experimental results, our model uses the ReLU activation function, two convolutional layers and one max-pooling layer structure at last, the results show our model is reliable and satisfactory in terms of accuracy requirement of many scenarios. Before our work, most security companies manually fill in the identification information, resulting in intensive human labor and a high percentage of error. In order to solve the problem of designing laborious hand-typing work and increase the processing speed, we developed the proposed approach and introduced human evaluation to illustrate the advantages of our algorithm in terms of reducing manual labor and improving the accuracy rate. We used 120 real-world ID cards excluding in our training samples to evaluate the recognition rate. The results are shown in Table 3 by human side-by-side evaluation. Since there are some anti-counterfeiting marks and patterns on the back side, it is not an easy task to locate character areas and recognize information. Moreover, the boundary between the periods is highly interfering, leading to recognition errors. Therefore, the back recognition rate is often lower than the front side. It should be noted that all the training datasets are all from real world ID cards, and due to the complexness of Chinese characters, it is normal that the character recognition rate is a little bit lower than the number recognition.

TABLE 3. **Evaluation of recognition quality by human side-by-side evaluation. The number denotes recognition rate(%). The front of ID card includes information of name, gender, birthday, nation, address and ID card number of holder (1-6 in the table), the back have two lines of words are the information of expiry date and issuance authority.**

	Column	Recognition Rate
1	Name	90.45
2	Gender	95.83
3	Nation	93.23
4	Address	89.03
5	ID number	**99.07**
6	Birthday	96.78
7	The front	**94.07**
8	The back	89.50

C. MODEL RESULT EXHIBITION

Our model includes four parts: the image processing, ID card location, character segmentation and recognition. The overall performance of the model depends on the performance of these four sections. Therefore each section of the model



(a) Original image as the model's input, it is a natural scene picture with complex backgroud.

```
Img/n-20.jpg
姓名 高[redacted]
性别 男
民族 汉
住址 辽宁省大连市[redacted]10号1—5—2
身份证号 11[redacted]816
生日 19[redacted]年12月22日
签发机关
有效期始
```

(b) Output of our system, it is not 100% accurate, the first number of ID number is wrong, but it is acceptable.

FIGURE 13. **The recognition result.**

is equally important. The results of our model are shown in Fig. 13, the original picture is a natural picture such as a person readily photographed, our model extracts the ID card region in the complex background and recognize it. Although the result are not 100% correct, the first number of ID number is wrong as shown in Fig. 13, the method is still acceptable and can significantly decrease human one work load.

Moreover, it is worth mentioning that our Chinese character data set is only about 3000 character image labels in total, including artificial data and real data but not cover the common Chinese words. Our future work aims at solving these flaws in order to improve our model.

VI. CONCLUSION

In this paper, we proposed an effective machine learning approach for real-world Chinese ID card recognition, including all the techniques that are critical to ensure its accuracy. To the best of our knowledge, we are the first to use the deep learning method to identify Chinese ID card, and also we are the first to publicly report the use of deep learning methods to identity Chinese ID card, which will drive the academic development of the OCR identification technique. On our real-world ID card dataset, the recognition quality of our method approaches accuracy of 94% which can significantly decrease human work and highly recommendable for use in the industry. Moreover, for the limitation of the real-world ID cards dataset, we leverage an ID card characters dataset by training dataset ourselves to train a convolution neural network. Experimental results show that our model performs well, the recognition result is reliable with satisfactory accuracy for practical scenarios.

REFERENCES

[1] B. Shan, "Vehicle license plate recognition based on text-line construction and multilevel RBF neural network," *JCP*, vol. 6, no. 2, pp. 246–253, 2011.

[2] B. Yang, X. Zhang, L. Chen, H. Yang, and Z. Gao, "Edge guided salient object detection," *Neurocomputing*, vol. 221, pp. 60–71, Jan. 2017.

[3] Z. Wang, G. Xu, Z. Wang, and C. Zhu, "Saliency detection integrating both background and foreground information," *Neurocomputing*, vol. 216, pp. 468–477, Dec. 2016.

[4] G. Cheng, J. Han, L. Guo, X. Qian, P. Zhou, X. Yao, and X. Hu, "Object detection in remote sensing imagery using a discriminatively trained mixture model," *ISPRS J. Photogramm. Remote Sens.*, vol. 85, no. 9, pp. 32–43, 2013.

[5] J. Han, D. Zhang, G. Cheng, L. Guo, and J. Ren, "Object detection in optical remote sensing images based on weakly supervised learning and high-level feature learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 6, pp. 3325–3337, Jun. 2015.

[6] H. Sun, X. Sun, H. Wang, Y. Li, and X. Li, "Automatic target detection in high-resolution remote sensing images using spatial sparse coding bag-of-words model," *IEEE Geosci. Remote Sens. Lett.*, vol. 9, no. 1, pp. 109–113, Jan. 2012.

[7] L. Zhang, L. Zhang, D. Tao, and X. Huang, "Sparse transfer manifold embedding for hyperspectral target detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 2, pp. 1030–1043, Feb. 2014.

[8] G. Cheng, P. Zhou, and J. Han, "Learning rotation-invariant convolutional neural networks for object detection in VHR optical remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 12, pp. 7405–7415, Dec. 2016.

[9] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, May 2015.

[10] Y. Guo, Y. Liu, A. Oerlemans, S. Lao, S. Wu, and M. S. Lew, "Deep learning for visual understanding: A review," *Neurocomputing*, vol. 187, pp. 27–48, Apr. 2016.

[11] W. Liu, Z. Wang, X. Liu, N. Zeng, Y. Liu, and F. E. Alsaadi, "A survey of deep neural network architectures and their applications," *Neurocomputing*, vol. 234, pp. 11–26, Apr. 2017.

[12] X. S. Xin, Q. Song, L. Yang, and W. H. Jia, "Recognition of the smart card iconic numbers," in *Proc. Int. Conf. Electron., Inf. Comput. Eng.*, vol. 44, 2016, Art. no. 02087.

[13] C. Yao and G. Cheng, "Approximative Bayes optimality linear discriminant analysis for Chinese handwriting character recognition," *Neurocomputing*, vol. 207, pp. 346–353, Sep. 2016.

[14] W. Wu, J. Liu, and L. Li, "Text recognition in mobile images using perspective correction and text segmentation," *Int. J. Signal Process., Image Process. Pattern Recognit.*, vol. 9, no. 10, pp. 171–178, 2016.

[15] J.-B. Wang, N. He, L.-L. Zhang, and K. Lu, "Single image dehazing with a physical model and dark channel prior," *Neurocomputing*, vol. 149, pp. 718–728, Feb. 2015.

[16] Z.-J. Zhu, Y. Wang, and G.-Y. Jiang, "Unsupervised segmentation of natural images based on statistical modeling," *Neurocomputing*, vol. 252, pp. 95–101, Apr. 2017.

[17] C. Tomasi and R. Manduchi, "Bilateral filtering for gray and color images," in *Proc. ICCV*, Jan. 1998, vol. 98, no. 1, pp. 839–846.

[18] S. Gupta and S. G. Mazumdar, "Sobel edge detection algorithm," *Int. J. Comput. Sci. Manage. Res.*, vol. 2, no. 2, pp. 1578–1583, Feb. 2013.

[19] W. Gao, X. Zhang, L. Yang, and H. Liu, "An improved sobel edge detection," in *Proc. 3rd IEEE Int. Conf. Comput. Sci. Inf. Technol. (ICCSIT)*, vol. 5, Jul. 2010, pp. 67–71.

[20] J. Canny, "A computational approach to edge detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-8, no. 6, pp. 679–698, Nov. 1986.

[21] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995.

[22] V. N. Vapnik, *Statistical Learning Theory*, vol. 3. New York, NY, USA: Wiley, 1998.

[23] D. H. Ballard, "Generalizing the Hough transform to detect arbitrary shapes," *Pattern Recognit.*, vol. 13, no. 2, pp. 111–122, 1981.

[24] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 779–788.

[25] H. Lee, R. Grosse, R. Ranganath, and A. Y. Ng, "Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations," in *Proc. 26th Annu. Int. Conf. Mach. Learn. (ACM)*, 2009, pp. 609–616.

[26] H. Qu, S. X. Yang, A. R. Willms, and Z. Yi, "Real-time robot path planning based on a modified pulse-coupled neural network model," *IEEE Trans. Neural Netw.*, vol. 20, no. 11, pp. 1724–1739, Nov. 2009.

[27] H. Qu, Z. Yi, and H. Tang, "Improving local minima of columnar competitive model for TSPs," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 53, no. 6, pp. 1353–1362, Jun. 2006.

[28] D. Ciresan, U. Meier, and J. Schmidhuber, "Multi-column deep neural networks for image classification," 2012, *arXiv:1202.2745*. [Online]. Available: https://arxiv.org/abs/1202.2745

[29] A. Mohamed, G. E. Dahl, and G. Hinton, "Acoustic modeling using deep belief networks," *IEEE Trans. Audio, Speech, Language Process.*, vol. 20, no. 1, pp. 14–22, Jan. 2012.

[30] G. E. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 1, pp. 30–42, Jan. 2012.

[31] H. L. H. Li, J. Chen, and Z. Chi, "CNN for saliency detection with low-level feature integration," *Neurocomputing*, vol. 226, pp. 212–220, Feb. 2017.

[32] Y. Le Cun, B. Boser, J. S. Denker, R. E. Howard, W. Habbard, L. D. Jackel, and D. Henderson, "Handwritten digit recognition with a back-propagation network," in *Proc. Adv. Neural Inf. Process. Syst.*, 1990, pp. 396–404.

[33] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

[34] J. Tompson, R. Goroshin, A. Jain, Y. LeCun, and C. Bregler, "Efficient object localization using convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 648–656.

[35] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun, "OverFeat: Integrated recognition, localization and detection using convolutional networks," 2014, *arXiv:1312.6229*. [Online]. Available: https://arxiv.org/abs/1312.6229

[36] V. Nair and G. E. Hinton, "Rectified linear units improve restricted Boltzmann machines," in *Proc. 27th Int. Conf. Mach. Learn. (ICML)*, 2010, pp. 807–814.

[37] R. Collobert and J. Weston, "A unified architecture for natural language processing: Deep neural networks with multitask learning," in *Proc. 25th Int. Conf. Mach. Learn.*, 2008, pp. 160–167.

[38] Y. N. Dauphin, A. Fan, M. Auli, and D. Grangier, "Language modeling with gated convolutional networks," in *Proc. 34th Int. Conf. Mach. Learn.*, vol. 70, 2017, pp. 933–941.

[39] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*. [Online]. Available: https://arxiv.org/abs/1810.04805

[40] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[41] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," 2014, *arXiv:1409.0473*. [Online]. Available: https://arxiv.org/abs/1409.0473

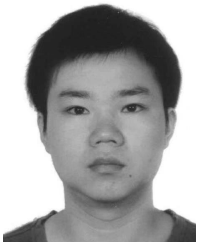
[42] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *Proc. 14th Int. Conf. Artif. Intell. Statist.*, 2011, pp. 315–323.

LIN ZUO received the Ph.D. degree in computer science from the University of Electronic Science and Technology of China, Chengdu, in 2011. From 2009 to 2010, she was a Visiting Predoctoral Fellow with Northwestern University, Evanston, IL, USA. She is currently an Associate Professor with the School of Information and Software Engineering, University of Electronic Science and Technology of China. Her research interests include artificial intelligence and neural networks.

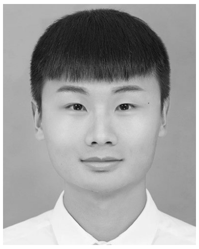
WENYU CHEN received the Ph.D. degree in computer science from the University of Electronic Science and Technology of China, Chengdu, in 2009, where he is currently a Professor with the School of Computer Science and Engineering. His research interests include computing theory, and machine intelligence and pattern recognition.

HONG QU received the Ph.D. degree in computer science from the University of Electronic Science and Technology of China, Chengdu, in 2006. From 2014 to 2015, he was a Senior Visiting Scholar with the Humboldt University of Berlin. He is currently a Professor with the School of Computer Science and Engineering, University of Electronic Science and Technology of China. His research interests include artificial intelligence and neural networks.

LI HUANG is currently pursuing the Ph.D. degree with the School of Computer Science and Engineering, University of Electronic Science and Technology of China. Her research interests include deep learning and natural language processing.



ZHENG WANG received the bachelor's and Ph.D. degrees from Zhejiang University, China, in 2017 and 2011, respectively. From 2014 to 2015, he was a Visiting Scholar with DLR, German Aerospace Center funded by CSC. He is currently a Postdoctoral Research Fellow with the School of Computer Science and Engineering, University of Electronic Science and Technology of China (UESTC). His current research interests include cross-media analysis, computer vision, and machine learning.



YONG CHEN is currently pursuing the degree with the School of Mechanical and Electrical Engineering, University of Electronic Science and Technology of China. His research interests include computer vision and visual slam.