Text Detection Model Based on Parallel SVM Algorithm

Yan Zhang & Yufang Song
Department of Foreign Language, Qinhuangdao Institute of Technology, Qinhuangdao, Hebei, China

ABSTRACT: With the development of the computer science and technology, the advanced algorithm rapidly develops in the field of the text detection and recognition and reduction. However, how to establish a model to be faster and more accurate access to more complex variants or the texts with missing information is to be solved urgently. In this paper, MATLAB parallel computing toolbox is applied to develop the parallel SVM. And this paper improves the operational efficiency through the distributed algorithm and selects the appropriate algorithm parameters to improve the accuracy of the model detection through cross validation. Based on the parallel MATLAB platform of multiple computers, this paper carries out a simulation test and concludes that the accuracy of the complex text detection is 98%. This paper establishes a text detection model that has a high accuracy and a high robustness even in a complex background.

Keywords: big data; parallel computing; SVM text detection

1 INTRODUCTION

In human life and production activities, most of the information acquired is the visual information, such as the license plate, auto logo or vehicle model are required to be recognized in daily traffic process. In the literature work, there is a need to recognize a great number of text information, and the information is often incomplete due to the action of nature. For example, the bill font may fade for years, resulting in difficulty in the commercial forensics or other commercial activities and other problems to be solved urgently. The text plays a mainstay role in scientific and technological exchanges. The development of the text feature extraction and detection technology plays an important role in the progress in daily academic exchanges. As the technology of the big data becomes mature, how to quickly and accurately acquire and recognize the text features and convert into the text at the technology terminal is often found in reports. In Handwritten Script Recognition System for Languages with Diacritic Signs, Cosin, H. researches the application of the neural network in the handwritten script recognition and carries out classification detection through letters, words and fragments. In Identity Recognition using an Artificial Intelligence based on Artificial Immune System, Jadiel Caparrós da Silva researches the application of the artificial intelligence in the recognition of the human facial expression, introduces the immune system in the algorithm and obtains a good recognition result. However, for the detection of more complex variants or the texts with missing information, currently, there is no effective technology for detection recognition or further repair, and the recognition rate and operating efficiency of the previous models is often not satisfactory with a longer period of recognition. Given that the image text is mostly the mass data, this paper applies the distributed algorithm for parallel computing and SVM for the text feature recognition, establishment of a high-precision text detection model and improvement of its robustness based on the idea of big data.

2 MODELING

2.1 Distributed algorithm

The image text is mostly the mass data. The traditional process mode is not only time-consuming, but also presents a downward trend with the increase of accuracy of the data size. The distributed system rapidly develops with the development of the computer hardware and the emergence of the big data concept, and its application gradually becomes mature. The distributed parallel processing can not only improve the accuracy of the text detection, but also increase the
detection efficiency. However, its programming difficulty and algorithm complexity are criticized by people. The parallel toolbox developed by MATLAB is not required to be developed from the bottom, but only needs to simply rewrite the programming languages and add parafor-loops in a loop for loop optimization.

The parallel computing toolbox developed based on MATLAB platform is used to solve the above problems. PCT of MATLAB provides two types of computing methods: dadafirst and jobfirst. The latter method can be used for computing after establishment of a complex model because the data has hereditability. MATLAB server has robustness for the scheduler, which can be compatible with MSCC, APBSpro and other scheduling packages.

2.2 SVM principle

SVM is another new learning method after the emergence of the neural network. SVM makes \( f(x_1, x_2, \cdots, x_n) \) achieve a certain mapping link in \( R^n \rightarrow R \) criterion. Such algorithm is supported by the theory with reliable robustness and accuracy. The theoretical basis of this model and the preferential criterion of the algorithm are shown in Figure 2.

Two types of problems can be realized based on SVM, that is, linear transformation criterion and non-linear transformation criterion. The training matrix is defined as follows:

\[
\{(x_i, y_i), i = 1, 2, \cdots, l\}
\]

Where: \( x_i \) is part of the neighborhood rule \( (R_i) \), then \( y_i = 1 \). If \( x_i \) is a part of the neighborhood rule \( (R \neq R_i) \), then \( y_i = -1 \). The first class of the plane or hyperplane is defined as follows:

\[
w^T x + b = 0
\]

This set can converge the sample matrix in \( R_i \) and \( R \neq R_i \), indicating that this matrix has a regular and clear neighborhood rule, that is:

\[
\begin{align*}
w^T x + b & \geq 1 \quad \text{Where: } y_i = 1 \\
w^T x + b & \leq -1 \quad y_i = -1
\end{align*}
\]

The link weight between a certain element in the matrix and a certain neighborhood in (2) is:

\[
\varepsilon_i = y_i (w^T x_i + b) = |w^T x_i + b|
\]

Map-Min-Max process is given to \( w \) and \( b \) in (3). Substitute \( \frac{w}{\|w\|} \) and \( \frac{b}{\|w\|} \) into \( w \) and \( b \), the interval of the Euclidean space is defined as follows:

\[
\delta_i = \frac{w^T x_i + b}{\|w\|}
\]

The allowable interval of the Euclidean space for the convergence of this matrix in \( R_i \) and \( R \neq R_i \) is defined as follows:
\[ \delta = \min \delta_i \quad i = 1, 2, \cdots, l \]  
\[ N \leq \left( \frac{2R}{\delta} \right)^2 \]  
\[ 0 \leq a_i \leq C \]  
\[ \sum a_i y_i = 0 \quad i = 1, 2, \cdots, n \]  

The above problem is the linear classification criterion. For the non-linear transformation criterion, it is unable to transform in its domain, which must be mapped to the domain \( R \times [8] \).

### 3 MODEL SOLUTION

The text is the image with features. And the features are divided into the features with statistical significance and architectural features. The architectural features can be modeled and classified by defining the basis matrix. The architectural features in the image can be extracted by \textit{Sobel} operator. Its basic definition is as follows: The pixel matrix at the boundary can be weighted for solving the extremum. The mathematical description is as follows:

\[ S = \begin{bmatrix} f(x+1, y+1) + f(x+1, y) + f(x+1, y-1) \\ f(x-1, y-1) + f(x-1, y+1) \end{bmatrix} \]

\[ S = \begin{bmatrix} f(x-1, y+1) + f(x-1, y) + f(x+1, y-1) \\ f(x+1, y-1) + f(x+1, y+1) \end{bmatrix} \]

Training for the text randomly selected to be tested can not only ensure the randomness of training requirement, but also achieve the diversity of the training samples, so as to prevent missing detection. And the text randomly selected is as follows: Four score and seven years ago, our fathers brought forth on this continent, a new nation, conceived in Liberty, and dedicated to the proposition that all men are created equal.

After the extraction of a certain feature by \textit{Sobel} operator as a training matrix and the connection with SVM program, it conducts classification learning by SVM. In the learning process, SVM parameters are
adjusted to ensure the detection accuracy and numerical stability.

Three desktop computers in the group are interconnected with LAN. Detect the connectivity through Ping, operate GUI and start mdc. In the dispatching center, there is a need to add the host column and create Job detector and connecting points. In preparing the SVM function, there is a need to adopt parfor and other commands to extend the code to executable parallel state. This paper assigns SVM learning tasks to three computers. Its pseudo-code and schematic diagram is shown in Figure 4:

```
matlabpool open local 3
parfor i
    A(i) = Sobel(D)
end
plot(A)
matlabpool close
```

Figure 4. Pseudo-code and schematic diagram.

There is no SVM parallel code in the toolbox, so train, Searchsource, Creatework and other independent function sets are self-prepared for computing. The features are extracted by Sobel operator, and the feature matrix is used as the training matrix to connect with SVM program and enter the test matrix for testing. This paper adopts jobs and tasks methods to create the parallel computing. The steps are as follows:

1. First, seek for a set distribution system by Searchsource and create the object scheduler js (i) by assigning the task management jobs.
2. Create jobs on the object scheduler and name it as work (i); at the same time, define and train the text feature matrix by train ( ); make definition and addressing of the path and name of the module in the creatework ( ).
3. Assign the work (i) as a set of jobs with the number of n; set up SVM parametric variations in train module and creatework module.
4. Integrate the partitioning jobs into the background through submit ( ), and complete the assigned tasks.
5. To take back the results and push to the host by the function of getA_OA ( ) after the iterative computation reaching the specified limit of error.

```
js=searchsource('type', 'jobmanager', 'name', 'node')
work(i)=creatework(js, 'FileD', '{train.m}', 'pathD='D:\')
createTa(work(i), @train, n, 'all')
submit(createTa)
Result = getA_OA(work(i))
```

Figure 5. Schematic diagram of distributed computing process

The schematic diagram of the above process is shown in Figure 5. And the main code is shown as follows:

```
js=searchsource('type', 'jobmanager', 'name', 'node')
work(i)=creatework(js, 'FileD', '{train.m}', 'pathD='D:\')
createTa(work(i), @train, n, 'all')
submit(createTa)
Result = getA_OA(work(i))
```

SVM parameter selection in the third step has an important impact on the accuracy rate of learning. This paper adopts the random noise cross-validation method for optimization of parameters. After training,
c= 0.25, g= 2, and the classification accuracy reaches 98%. The contour line of the parameter selection is shown in Figures 5 and 6.

The combination with the distributed algorithm and support vector machine requires using a large number of sub functions, so the user graph interaction interface is written to improve the universality of this model. Many functions can be free of being transferred through writing GUI user interface, and there is only a need to click on the button for file leading-in and text detection. The prepared GUI interface and the final detection results are shown in Figure 7.

4 CONCLUSION

As can be seen from the simulation results, the parallel computing based on the distributed algorithm can significantly improve the operational efficiency, and the SVM algorithm developed by the MATLAB parallel computing toolbox is applied to achieve 98% of the model recognition rate. In terms of the parameter selection, the cross validation method introduced also has a good effect on the parameter selection. Moreover, compared with the genetic algorithm, PSO algorithm and other complex algorithms, the model is simple and effective. The establishment of the graphical user interface system improves the universality of the model and reduces the difficulty in use of the system, and it also has a good secondary development performance, thus bringing convenience to subsequent code optimization and module development for the users.

REFERENCES


