A new type of video text automatic recognition method and its application in film and television works

1st Yining Xu
Nanjing Keyuan Intelligent Technology Group Co., Ltd
Nanjing, China
inyhsu@outlook.com

2ndMinghao Liu
Nanjing Keyuan Intelligent Technology Group Co., Ltd
Nanjing, China
2550668988@qq.com

3rd Fang Wang *Nanjing Arts Institute* Nanjing, China
495526487@qq.com

2023 6th International Symposium on Autonomous Systems (ISAS) | 979-8-3503-1615-5/23/$31.00 ©2023 IEEE | DOI: 10.1109/ISAS59543.2023.10164415

***Abstract*— In the current post-production of film and television, editing and school teams need to identify each frame of image in the video and modify the sensitive words in it, which consumes a lot of manpower, material resources and financial resources. This article proposes an image recognition method based on attention mechanism to overcome this problem, which can quickly recognize specified texts in videos. Using the convolutional neural network and codec as the framework, the target text is used as the operation object of the attention mechanism, so as to realize the rapid recognition of pictures containing the target video text. To validate the proposed method, a TV series is taken as an example to test the characters involved in the video, and the results show the effectiveness of the proposed video text automatic recognition method.**

Keywords—Film, Convolutional neural network, codec, video text

I. INTRODUCTION

Since the concept of image text recognition was proposed by Tausheck in 1929 [1], a large number of scientific researchers have been interested in image text recognition and conducted in-depth research on its technology. In the initial stage of image text recognition, the traditional font matching method was mainly used for text recognition. However, due to the problem of complex text fonts and many types of text, the recognition method did not achieve a good recognition effect. The era of deep learning is gradually approaching, and neural networks such as CNN and RNN have been widely used in image text recognition technology by scientific researchers [2-4], making image text recognition a certain breakthrough in technology, and also obtained certain research results. However, there are still many problems to be solved in image text recognition. First, how to solve the fitness problem in image text recognition. Second, how to integrate the rich and precise text semantic information carried by text into the image text recognition task. Third, how to recognize image text in natural scenes. In the task, attention mechanism is used to extract various features of image text. Most of the key feature information that affects text recognition. The solution of all these problems will help to promote the image-text recognition. The further development of different theories and the innovation of technology.

Yao et al. [5] used image clustering to learn character stroke features, used the Hough voting algorithm to detect characters in the image, and performed character recognition through random forests on the basis of stroke features and Hough features. Belongie et al. [6] proposed an improved text recognition algorithm based on the spatial constraint relationship between characters and the confidence of characters. What's more, new research ideas have been brought to the research of image text recognition. Researchers have gradually adopted deep learning methods to improve the performance of image-text recognition systems by adding deep learning techniques to traditional algorithms. For example, Alsharif [7] and others took the lead in applying convolutional neural network (CNN) to character-based image text recognition tasks, using convolution to complete complex operations including segmentation, correction and character recognition, and using a fixed dictionary The hidden Markov model (HMM) generates the final recognition result.

However, in terms of film and television production, the method of image text recognition has not been widely used, and the current method does not take the attention mechanism into consideration. Therefore, the pertinence and effectiveness of target recognition still need to be improved. Thus, this paper proposes new type of video text automatic recognition method. Based on the visual features of the original image text, the proposed model introduces the text semantic features of the image, and the interactive attention mechanism is further used to realize the interaction and fusion of image semantic features and visual features, so as to enhance the image text recognition performance.

II. PRELIMINARIES

A. Image Preprocessing

Image preprocessing is mainly to trim the image before recognition, through some trimming methods to make the image have a better recognition effect when it is recognized. In the process of image text recognition, image preprocessing plays a vital role, and the processing involved mainly includes noise reduction, enhancement, tilt correction and binarization..

Corresponding Author: Fang Wang, **Affiliation:** *Nanjing Arts Institute* **City:** Nanjing, China **Email:** *495526487@qq.com*.

B. Attention mechanism

The visual attention mechanism is an effective method to help us focus on one thing. Firstly, scan the global image, secondly, identify the target area that needs attention, and then invest more energy in the area that needs attention, thereby indirectly suppressing useless interference information.

Assign the attention coefficient to the input feature sequence is the keypoint of attention mechanism in deep learning, so that each feature sequence has a weight representing its own importance degree. The greater the influence of the feature sequence on the model, the greater the weight of the attention allocation. The Attention mechanism really entered the field of vision of researchers in the paper "Recurrent Models of Visual Attention" published by the Google Mind Team in 2014 [8]. In this paper, an implementation idea of the attention mechanism is proposed. At present, the mainstream attention mechanism is mainly composed of Query, Key and Value. The specific structural framework is shown in Figure 2-1 Attention framework. The constituent elements of Source are composed of multiple sets of <Key, Value> data pairs. The Attention mechanism is implemented by calculating the correlation of each Key value in Query and Source to obtain the correlation coefficient of the corresponding Value value, and then obtain the final Attention by weighting and summing the Value values. The above process is represented as shown in Equation 1.

$$Attention(Query, Source) = \sum_{i=1}^{n} Similarity\,(Query, Key_i) \star Value_i \quad (1)$$

Where, n stands for the Key number and the $Source$ number, $Similarity(Query, Key_i)$ stands for the similarity function between $Query$ and Key_i.

In 2017, [9] was published, which used the self-attention mechanism to learn text representation. The work obtained good results, which led to a research hotspot about self-attention. The traditional attention mechanism has a high dependence on external feature data and is not good at capturing the internal correlation of information. Therefore, a self- attention mechanism is proposed to improve the above problem, and its algorithm is shown below.

$$Att(Q, K, V) = \omega(QK^T)V \quad (2)$$

where, $Att(Q, K, V)$ stands for the self-attention mechanism, Q, K, V stands for the Query Vector, Key Vector, and Value Vector, respectively, ω stands for the activation function. $Att(Q, K, V)$ is also the sum of the vector weights.

Q, K, V can be obtained by multiplying the input sequence X by three matrices Wq, Wk, and Wv respectively.

$$Q = W_q X \quad (3)$$

$$K = W_k X \quad (4)$$

$$V = W_v X \quad (5)$$

The self-attention mechanism is often applied to codecs to solve problems such as image text recognition, because the self-attention model can learn the dependencies between different regional features.

C. Convolutional neural network

A convolutional neural network (CNN) is a feedforward neural network with connectivity between layers inspired by animal visual cortex [49]. CNN is mainly divided into three layers: convolution layer, pooling layer and fully connected layer. These modules are used to complete the task of feature learning and classification. The basic unit of convolution layer is neuron, which is the basic processing unit of neural network, and is usually expressed in the form of multi-input and single-output. The whole input and output process of a neuron can be expressed as

$$y_j = f(b_j + \sum_{i=1}^{n}(X_i * w_{ij})) \quad (6)$$

where, X_i stands for the input of the network, w_{ij} stands for the weight of the linked neurons, b_j stands for the bias of the internal parameters, and f stands for the activation function.

In the convolution layer, neurons and local areas of the upper layer are connected by the convolution nucleus [10]. The convolution kernel is usually a weight matrix, and different convolution layers can be convoluted through the convolution kernel. The convolution kernel is similar to a sliding window, which slides at a specific step size in the entire input signal. CNN is mainly used for high-dimensional input such as images. Through convolution operation of convolution kernel, local features of images can be extracted.

III. TEXT RECOGNITION BASED ON ATTENTION MECHANISM

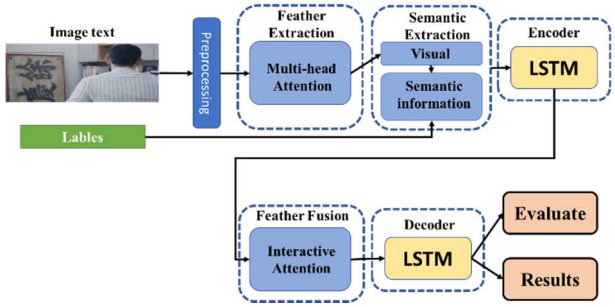
Most codecs for image text recognition use image visual features to encode and decode information, ignoring the semantic information expressed by image text. In order to enhance the ability of model recognition, this paper takes the Seq2Seq neural network model as the basic framework, introduces the semantic image text information, and this paper also proposes a new type of video text automatic recognition method, which based on attention mechanism. This chapter starts with the idea of building the model, and introduces in detail the extraction and feature fusion methods of image features and text semantic features of the image in the model. Finally, the realization of timing sampling mechanism of decoder is introduced.

A. Algorithm model

The current popular models ignore the semantic information between images and text. The model proposed in this article integrates the semantic information of images and text. The model proposed in this article further extracts image text semantic information based on image visual information features, and extracts association information between two information features based on interactive attention mechanism. By using semantic information to "focus" on image text structure information, and using image structure information to focus on semantic expression of image context.

The construction idea of this article is: on the one hand, the multi head attention mechanism is used to extract visual

feature information of images, and on the other hand, the semantic module proposes the image text semantic feature



information. The encoder encodes the extracted visual features and semantic features. The interactive attention mechanism and concat algorithm are used to decode the image semantic features of the "concerned" image visual features and the image visual feature decoder of the "concerned" image semantic features, and output the recognition results. The model construction mainly consists of five different parts: feature extraction, semantic extraction, decoder, and feature fusion, and decoder section. Figure 1 shows the model framework.

Figure 1. A Framework of Codec Text Recognition Model Based on Attention Mechanism

B. Preprocessing

The SCSD data set and ITD data set used in the experiment in this paper are standard 280 * 32 color images, so in the process of image preprocessing, only graying processing is used, that is, the R, G and B components of the image are weighted and averaged with different weights.

$$Gray = 0.114B + 0.587G + 0.299R \quad (7)$$

where $Gray$ is the pixel gray value after image processing. B is the value of pixel Blue channel; G is the value of pixel Green channel; R is the value of pixel Red channel.

C. Image feature extraction

This article adopts a multi head attention mechanism to extract deep image features. The number of bulls is set to 8; The dimensions of Q, K, and V are 64; The obtained features by the 8 single-head attention constructions are fused into 512-dimensional feature vectors. The fused features are linearly transformed by the fully connected neural network to obtain the input attention information and realize the deep feature extraction of the input information. The specific model is shown in Figure 2.

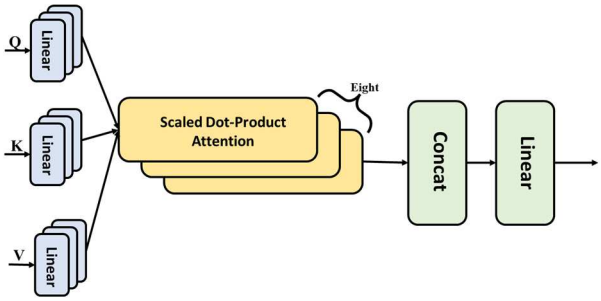


Figure 2. Multiple attention model

D. Visual feature representation

In order to better extract the visual features and context relations in the image text, the codec in this paper uses the bidirectional long short-term memory neural network (BiLSTM) as the basic component. The specific implementation of the proposed method in image visual feature representation is to use BiLSTM to learn the image deep visual feature representation X . The calculation equations can be shown as follows:

$$\overrightarrow{x_t} = \overrightarrow{LSTM}(I_t) \quad (8)$$

$$\overleftarrow{x_t} = \overleftarrow{LSTM}(I_t) \quad (9)$$

$$x_t = [\overrightarrow{x_t}, \overleftarrow{x_t}] \quad (10)$$

$$X = [x_1, x_2, \dots, x_t, \dots, x_n] \quad (11)$$

where I_t stands for the input of the image features extracted by multi-attention. $\overrightarrow{x_t}$ and $\overleftarrow{x_t}$ stands for visual feature representation obtained through forward and reverse LSTM networks.

E. Text feature extraction

At present, the codec of image character recognition model focuses on image visual features for feature extraction. In most cases, the model works well. However, when the image is blurred, the performance of model text recognition is degraded. The model mainly uses visual features for text recognition lead to bad performance. In the case of blurred images, it is difficult for the model to capture high-quality image visual feature information, resulting in reduced recognition performance. The text image is different from the ordinary image. The text image contains both the morphological features of the text and the semantic features of the image text. By extracting semantic information features of image text, its importance in image text recognition is enhanced, and the shortcomings of current models in text semantic information are compensated.

F. Semantic information extraction

The input image features capture the semantic information in the image text through two linear layers.

$$Z = W_2\sigma(W_1 I + b_1) + b_2 \quad (12)$$

where Z stands for the semantic feature vector of linear layer prediction, W_1 , W_2 , b_1 , and b_2 stands for the trainable weight, σ stands for the activation function.

The loss function can be represented by

$$L_{sem} = 1 - \cos(Z, em) \quad (13)$$

where em stands for the word vector of embedded text obtained by the pre-trained word2vec model.

G. Feature fusion

This article combines visual features with semantic information features to form an interactive attention mechanism. The visual feature extraction process is shown below.

$$D = \sum_{t=1}^{n} \partial_{1t} x_t \quad (14)$$

$$\partial_{1t} = \frac{\exp(u_t)}{\sum_{t=1}^{n} \exp(u_t)} \quad (15)$$

$$u_t = Tanh(W_1 y_t + b_1) \quad (16)$$

Where x_t stands for the t_{th} feature vector in X, y_t stands for t_{th} feature vector in Y, ∂_{1t} stands for the weight attention between Y and x_t, $Tanh$ is the nonlinear activation function, W_1 is the weight matrix, and b_1 is the bias. The fusion feature

vector representation calculation can be realized by the following formula.

$$R = concat(D, Q) \tag{17}$$

Where R stands for the fusion vectors, D stands for the visual feature representation, and Q stands for the semantic feature representation.

IV. Case Study

Environment

- Computer Name: LAPTOP-I6CEIQ3I; Operating System: Windows 11 Pro 64-bit (10.0, Build 22621); System Manufacturer: LENOVO; System Model: 20KH002HUS; Processor: Intel(R) Core(TM) i7-8650U CPU @ 1.90GHz (8 CPUs), ~2.1GHz; Memory: 16384MB RAM.
- Python 3.7.0; Pytorch 1.6.0; TorchVision 0.6.1;
- IDE: Pycharm 2020.1.4

Data

- The data comes from a TV play that is actually in trial and has not been released yet. Some fragments were taken for testing. In order to increase the diversity of the image, Gaussian noise, stains, horizontal and vertical lines, left and right blurring, etc. are randomly added when generating the image.

A. Performance evaluation of image and character recognition

The performance evaluation of image and character recognition system often adopts the methods of Levenshtein Distance Similarity (LDS) [11], accuracy rate, recall rate and F1 value to analyze the performance of character recognition model.

(1) Levenshtein Distance Similarity

The image character recognition model will have the problem of missing or multiple detection in the character recognition process, so the LDS is an essential evaluation standard. The LDS is a method to measure the difference between two sequences (strings). The LDS of two sequences is the minimum number of times required to edit a single character (such as insert, delete, and modify) divided by the word length when using one word to modify another word. The smaller the LDS is, the better.

In order to view the text recognition effect more intuitively, LDS is defined as 1-average editing distance. For example, sequence 1 is "*stri1*", sequence 2 is "*stri2*", sequence 1 needs to change sequence 2 once, the character length is 5, the average editing distance is 1/5=0.2, and the average editing distance similarity LDS=0.8.

(2) Precision

The precision rate is defined as the proportion of the number of characters recognized to the total number of characters recognized, which can reflect the situation of missing recognition and multiple recognition, but cannot reflect the situation of wrong recognition. The accuracy calculation formula is shown in Equation 18.

$$Pre = \frac{N_{acc}}{N_{pre}} \times 100\% \tag{18}$$

Where N_{acc} stands for the number of correctly recognized characters in the image sample, and N_{pre} stands for the total number of characters recognized.

(3) Recall

Recall rate is defined as the proportion of the number of correct characters to the actual number of characters, which can reflect the situation of recognition errors, but can not reflect the situation of multiple recognition and missing recognition. As shown in formula 19:

$$Rec = \frac{N_{acc}}{N_{rea}} \times 100\% \tag{19}$$

Where N_{rea} stands for the actual number of characters.

(4) F1 value

Because precision and recall cannot comprehensively evaluate the performance of image and character recognition model. Considering the accuracy rate and recall rate, the calculation of F1 value is defined as shown in formula 20.

$$F1 = \frac{2*(Pre*Rec)}{(Pre+Rec)} \tag{20}$$

B. Image and character recognition model based on interactive attention

The image and text recognition model based on interactive attention adopts the self-attention mechanism, which is based on the semantic information feature "focus" on the visual feature, and based on the visual feature "focus" on the semantic information feature. In order to analyze the mutual "concern" between image semantic information and image visual information and its impact on the performance of image character recognition more deeply, this section carries out the ablation experiment of corresponding image recognition model.

In order to visually show the convergence of image and character recognition model based on interactive attention under different rounds, this section lists the experimental results of loss, Pre, Rec, and average editing distance in the training using dataset model, as shown in Figure 3 and Figure 4, where the x-axis is the running round.

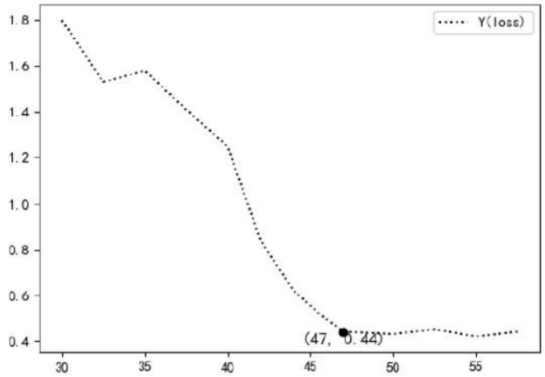


Figure 3. Loss value change curve

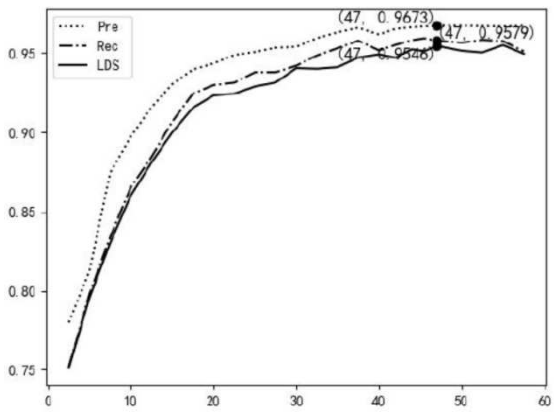


Figure 4. Evaluation parameter change curve

In order to further verify the effectiveness of the proposed codec text recognition model based on attention mechanism, we compare the proposed method with the traditional CNN method and the direct splicing semantic model. The results are shown in Table 1.

TABLE I. COMPARISON RESULTS

Methods	Index			
	Pre	***Rec***	***F1***	***LDS***
CNN	0.9036	0.8794	0.8907	0.8874
Direct splicing semantic model	0.9254	0.9016	0.9221	0.9032
The proposed codec text recognition model	0.9389	0.9103	0.9276	0.9101

we can see from Table 1 that the proposed codec text recognition model performs well in both Pre, Rec, F1, and LDS. The interactive attention mechanism is used to realize the mutual "attention" between the visual features and semantic features of the image text, and the effect of feature fusion based on image text recognition is better than that of concat direct splicing.

V. CONCLUSION

How to quickly and accurately batch extract text from images is an extremely important issue in the field of image and video processing. Identifying text in images can greatly improve the speed and accuracy of post review of film and television works, reducing unnecessary labor costs. However, traditional image recognition methods cannot meet the accuracy requirements and the extraction effect is poor. Therefore, this article proposes the A new type of video text automatic recognition method, which utilizes interactive attention to become the link between visual and semantic information features of linked image texts. The performance of the image text recognition model is optimized using multi head attention and timed sampling. The experimental results indicate that this method is effective. However, at present, the training speed of the model is relatively slow and the labeled data is less. The follow-up work will further improve the model from these two aspects.

REFERENCES

[1] Su Y M, Peng H W, Huang K W, et al. Image processing technology for text recognition[C]//2019 International Conference on Technologies and Applications of Artificial Intelligence (TAAI). IEEE, 2019: 1-5.

[2] Sharma N, Sharma R, Jindal N. Machine learning and deep learning applications-a vision[J]. Global Transitions Proceedings, 2021, 2(1): 24-28.

[3] Shinde P P, Shah S. A review of machine learning and deep learning applications[C]//2018 Fourth international conference on computing communication control and automation (ICCUBEA). IEEE, 2018: 1-6.

[4] Dhruv P, Naskar S. Image classification using convolutional neural network (CNN) and recurrent neural network (RNN): a review[J]. Machine Learning and Information Processing: Proceedings of ICMLIP 2019, 2020: 367-381.

[5] Yao C, Bai X, Shi B, Liu W. Strokelets: A learned multi-scale representation for scene text recognition[C]/Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on. iEEE, 2014: 4042-4049.

[6] Wang K, Babenko B, Belongie S.End-to-end scene text recognition[C]/Computer Vision(ICCV), 2011 IEEE. International Conference on IEEE, 2011: 1457-1464.

[7] Alsharif O, Pineau J. End-to-end text recognition with hybrid HMM maxout models[J]. E-print arXiv:1310.1811, 2013: 1-10.

[8] Mnih V , Heess N , Graves A. Recurrent Models of Visual Attention[J]. Advances in Neural Information ProcessingSystems, 2014:2204-2212.

[9] Vaswani A, Shazeer N, Parmar N. Attention Is All You Need[C]. Proceedings of the 31st International Conference on Neural Information Processing SystemsDecember. 2017: 6000-6010.

[10] Yin W,Schutze,Xiang B, Zhou B. ABCNN:Attention-based convolutional neural network for modeling sentence pairs[J]. Transactions of the Association for Computational Linguistics, 2016, 4:259-272.

[11] Zhang S, Hu Y, Bian G. Research on string similarity algorithm based on Levenshtein Distance[C]//2017 IEEE 2nd Advanced Information Technology, Electronic and Automation Control Conference (IAEAC). IEEE, 2017: 2247-2251.