Optimization Algorithm of Visual Multimodal Text Recognition for Public Opinion Analysis Scenarios

Xing Liu, Fupeng Wei, Qiusheng Zheng, Wei Jiang, Liyue Niu, Jizong Liu, and Shangshou Wang

Abstract Existing techniques for Monitoring public opinion on the internet generally rely on routinely mining text content from web pages, but they are unable to swiftly and effectively identify text content in images and videos. Therefore, a major challenge for multimodal information extraction in internet opinion scenarios is the quick and reliable identification and recognition of textual content in images and videos. Based on the state of the art in the field of optical character recognition (OCR), this paper proposed a improved method to visual multimodal text identification for scenarios such as internet opinion analysis. In order to enhance the learning effect of our overall model, this work upgrades the collaborative mutual learning (CML) distillation approach in the text detection module based on the combination of a classic distillation strategy and a deep mutual learning (DML) strategy. Next, the large kernel pixel aggregation network (LK-PAN) is then suggested as a solution to the earlier inadequacy in identifying multi-scale and text with extreme aspect ratios. In order to efficiently mine the contextual data in images or videos and to fulfill the goal of enhancing text recognition's mistake correcting capabilities, Transformer is finally implemented in the text recognition module. According to the experimental findings on the video dataset, our technique increases the F1 score by 17.97% and the recognition speed by 29.7%. The model provides important technical support for public opinion analysis in multimodal fields.

Keywords Public opinion analysis · Multimodal · Distillation · Text recognition

X. Liu · Q. Zheng · L. Niu · J. Liu · S. Wang
The Frontier Information Technology Research Institute, Zhongyuan University of Technology, No.1 Huaihe Road, Zhengzhou 450007, China

F. Wei (✉) · W. Jiang
North China University of Water Resources and Electric Power, No.136, Jinshui East Road, Zhengzhou 450046, China
e-mail: weifupeng@yeah.net

X. Liu · Q. Zheng · L. Niu · J. Liu · S. Wang
Henan Key Laboratory On Public Opinion Intelligent Analysis, No.1 Huaihe Road, Zhengzhou 450007, China

S. Li (ed.), *Computational and Experimental Simulations in Engineering*, Mechanisms and Machine Science 146, https://doi.org/10.1007/978-3-031-44947-5_97

1 Overview

As the all-media era progresses, people now interact with social hotspots and share their personal opinions through online platforms like Weibo, Zhihu, Jitterbug, and Racer. In addition to traditional text narratives, short videos and photos are becoming increasingly important carriers for conveying personal emotions because of their vivid and rich contents. However, there are also organizations that use the ability of images and videos to hide important details, such as water armies, network automata, etc., to spread false information. If they are not stopped in time, they could lead to a bad public image and cause significant societal losses.

Traditional approaches of monitoring public opinion online rely on a single data source, mostly text data crawled from web pages. The ability to recognize text in videos and images falls behind the ability to recognize text in multimodal data. In 2018, for the problem of distorted text detection, Liu et al. [1] proposed a deep character embedding network (CENet), which can extract the image data within the bounding box as a multiscale feature mapping and is thus characterized as embedding vectors. This makes text detection a clustering task in the character embedding space. In response to the underwhelming performance of the majority of existing approaches for curved text detection, Chen et al. [2] suggested an instance-aware segmentation-based text detection method for atypical scenarios in 2019. The main goal is to create a model for attention-guided semantic segmentation that correctly labels the weighted boundaries of text sections [3]. The approach has shown some promise on datasets with curved text. 2020 To solve the expensive issue of training a text recognition model, which necessitates a huge amount of data spanning as much variation as possible, Luo et al. [4] presented a text image enhancement method. The method starts with a collection of unique base points and then employs joint learning to close the gap between the two separate processes of data improvement and network optimization. Experiments show that the method produces training samples for the recognition network that are better suited for it. In 2021, Fang et al. [5] propose the autonomous, bidirectional, and iterative scene text recognition model Autonomous Bidirectional and Iterative Network (ABInet), which has an autonomous institution that proposes to block the gradient flow of visual and language models for display language modeling, followed by a bidirectional feature representation. This addresses the limitations of language models, including implicit language modeling, one-way feature representation, and noise. In the same year, Yong et al. [6] used a crawler technique to obtain text data of a "popular event" from Baidu's index, preprocessed these data to build a logistic differential equation model of online public opinion, and then used the Sine Cosine Algorithm (SCA) to solve it by combining the processed data. Zhang et al. [7] suggested a cross-modal depth metric learning-based oracle character recognition method in almost the same amount of time to address the problem that it is difficult to acquire topical oracle character samples in oracle character images. They were able to recognize topical oracle characters across modalities by modeling the common feature space and nearest neighbor classification of imitation and topical oracle characters.

The current web opinion monitoring techniques primarily depend on routinely crawling text content from web pages, which makes it difficult to rapidly obtain and identify text content from images and videos. Accordingly, this paper updates the Collaborative Mutual Learning (CML) distillation of the original PP-OCR V2 in the text detection module and then enhances the PP-OCR [8, 9] model of OpenVINO [10] based on the most recent developments in the field of text recognition and the characteristics of data in public opinion analysis. It is suggested that Large Kernel Pixel Aggregation Network (LK-PAN), a Pixel Aggregation Network (PAN) module with large perceptual fields, be used to address prior shortcomings in detecting text with multiple scales and extreme aspect ratios [11]. In order to effectively mine the contextual data of text line images and increase the text recognition module's capacity for mistake correction, a transformer network has been introduced [12]. Finally, the experimental findings demonstrate that the new model put forward in this study enhances both the accuracy and speed of text recognition in both images and videos.

2 Related Work

2.1 The Overall Framework of Traditional Opinion Text Recognition

Three modules of the PP-OCR model text detection, text detection frame correction, and text recognition are deployed and accelerated using OpenVINO. The implemented model can swiftly recognize and extract text data from pictures or video sources. The model's initial structure is depicted in Fig. 1.

The image or video dataset is first prepared in the first step (Stage 1) and used as the input for the following stage. Secondly, the photos or videos are pre-processed in the second step (Stage 2) using the OpenVINO Toolkit. Following that, the input data are progressively subjected to inference detection by three PP-OCR model submodules (text detection, text detection frame correction, and text recognition). The text detection module marks the area with text and gets the Bounding Box containing text; the text detection correction box module corrects the direction of the text inside the Bounding Box; and the text recognition module performs text recognition on the corrected area. Finally, In Stage 4, the extracted text information is visualized.

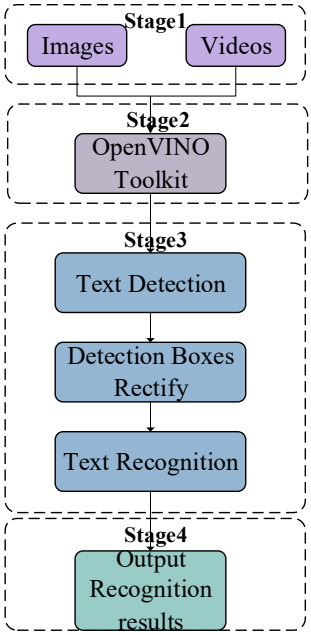

Fig. 1 Structure diagram of the model before improvement

2.2 PP-OCR Model

The method of Optical Character Recognition (OCR) involves turning a handwritten or printed image of text into computer-encoded text [13, 14]. The process is to determine the shape by detecting light and dark patterns in the image, and then translate the detected shape in the image into text by a recognition algorithm.

A text recognition module called PP-OCR is based on Baidu's deep learning flying paddle architecture, which recently combined deep learning, model compression, and the OCR fields. The technique offers end-to-end recognition in addition to the conventional method of text detection followed by text recognition. The PP-OCR model consists of three primary modules: text detection, detection bounding box correction, and text recognition, as shown in Stage 3 of Fig. 1.

The text detection module makes use of Differentiable Binarization (DB) [13], a segmentation-based scene text detection method that divides the heatmap produced by segmentation methods into bounding boxes and text regions. Then, using binarization operations, it marks the regions containing text using bounding boxes. Traditional binarization processes often use preset thresholds that cannot be modified to accommodate complex and changing detection scenarios. The binarization operation is inserted into the partition network for optimization, which in turn leads to the adaptation of the threshold value in each region of the heat map. The input height of the original code is 32. First statistical training sample image aspect ratio distribution

In terms of image dataset, we chose ICDAR 2015 because it has the highest aspect ratio with the training dataset (containing 1000 training sets and 500 test sets).

To address the scenario where the bounding box appears to be skewed in orientation and to correct the text detected in the image, the detection bounding box correction module employs a text orientation classifier. The specific operation flow of the direction classifier is as follows: The text angle classification is mainly applied to the case where the image is not at 0°, in which case the detected text lines in the image must go through a conversion process. However, when testing a large number of images, some overly long text shows obvious errors, so this paper processes the images in the improved model. If the text was too long, it was truncated; if the text was too short, it was copied and expanded to the size of the input image. The modified overly long text test results showed a significant improvement. After the detection area is directionally turned, it helps to improve the accuracy of text recognition.

The convolutional recurrent neural network (CRNN) is used in the text recognition module [15, 16]. The technical challenge of end-to-end OCR is how to handle the indeterminate long sequence alignment problem. The CRNN network first uses Convolutional Neural Network (CNN) to extract image features, taking the idea of solving indeterminate long speech sequences in speech recognition and modeling it as a time-dependent lexical or phrase recognition problem and then applies Bidirectional Long Short Term Memory Unit (Bi-LSTM) to resolve the indeterminate long sequence text problem [17, 18]. Finally, the Connectionist Temporal Classification (CTC) model [19] is introduced to resolve the segmentation alignment problem of samples in order to increase the model's applicability and lessen the significant amount of segmentation annotation work.

2.2.1 DB Algorithm

The main text detection algorithms are shown in Table 1. Text detection methods are classified into two types based on regression and segmentation.

Because regression-based algorithms have difficulty obtaining smooth text wrap-around curves for curved text, researchers have proposed image segmentation-based text detection algorithms. The segmentation procedure is as follows: First, we classify at the pixel level and determine whether each pixel point is a text target to obtain a probability map of the text region. Next, we process the probability map to obtain

Table 1 Main algorithms for text detection [20, 21]

	Horizontal text detection	Skewed text detection	Bending text detection
Regression-based text detection methods	CTPN, Textbox, etc	EAST, MOST, etc	CTD, LOMO, etc
Segmentation-based text detection methods	PAN, Seglink++, DB, PSENet, etc		

the wrap-around curve of the text segmentation region, which has better advantages for irregular text detection.

To address the time-consuming problem of subsequent processing caused by threshold binarization processing, PP-OCR uses a DB algorithm with a learnable threshold method. A binarization function that closely resembles a step function is created [22], and Eq. (1) defines the binarization function:

$$\hat{B}_{i,j} = \frac{1}{1 + e^{-k(P_{i,j} - T_{i,j})}} \tag{1}$$

where, $\hat{B}$ is the threshold value of text segmentation that the segmentation network can learn end-to-end during training, T is the adaptive threshold map learned from the network, and k stands for the magnification factor, and (i, j) represents the pixel coordinate locations on the image. The design of the binarization function allows the segmentation network to learn the threshold of text segmentation end-to-end during training, and automatically adjusts the threshold since the improvement of accuracy and also improves the text detection performance. DB architecture is shown in Fig. 2. The picture or video dataset is first created and utilized as the model's input. Second, The ResNet-18 network is used in the Backbone module to extract the feature information of the model input. Following 3*3 convolution upsampling and right after the feature map, the Fusion module aggregates the feature information of 1/8, 1/16, and 1/32 (indicating the scale of the input image or video) with the feature layers of 1/4, and describes the text by forecasting the threshold map and probability map of the feature layers. Finally, the fixed threshold is binarized in the post-processing stage to obtain the approximate binarization map, and then the text box is obtained from the approximate binarization map.

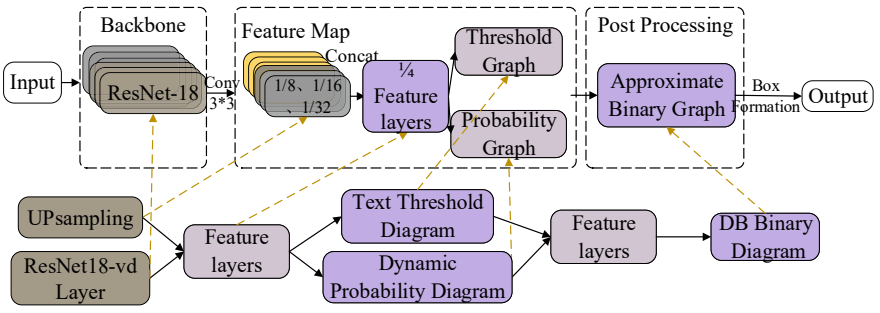

Fig. 2 DB overall architecture diagram

2.2.2 CRNN Network

Text recognition is one of the key subtasks of PP-OCR, which mainly completes the recognition of text content within the bounding box area and outputs the text content in the image and the corresponding confidence level. A single test image is used as the input to the CRNN network in the first step of the text recognition process, which is depicted in Fig. 3. Secondly, it goes through the image correction pre-processing module, which corrects skewed images and distorted text in images to reduce the difficulty of the visual feature extraction module. The visual feature extraction module then employs a convolutional neural network to extract the input image's feature data in order to produce the visual feature V. Then, the sequence feature extraction module uses the extracted visual feature information V as input, and the sequence feature extraction module goes on to extract the image's contextual feature information to produce the sequence feature information L. Finally, the prediction module processes L and outputs the result, the recognized text information.

A CRNN network is used by PP-OCR to recognize text. Assuming input x and output y, we want the posterior probability $p(y|x)$ to be as large as possible [23], and the CTC Loss function (assuming that the outputs of the RNN are independent of each other at each moment) is defined as shown in Eqs. (2) and (3).

$$p(\pi|x) = \prod_{t=1}^{T} y_{\pi_t}^t, \quad \forall \pi \in L^T \tag{2}$$

$$p(y|x) = \sum_{\pi \in B^{-1}(y)} p(\pi|x) \tag{3}$$

where B^{-1} is the mapping function of the set of all paths of y. $y_{\pi_t}^t$ represents the probability of π_t selecting a character at time step t, π_t represents the output character corresponding to path π at time step t. The CTC Loss function is next obtained as shown in Eq. (4).

$$\mathcal{L}(s) = -\sum_{(x,y)\in S} \ln \sum_{\pi \in B^{-1}(y)} \prod_{t=1}^{T} y_{\pi_t}^t, \quad \forall \pi \in L^T \tag{4}$$

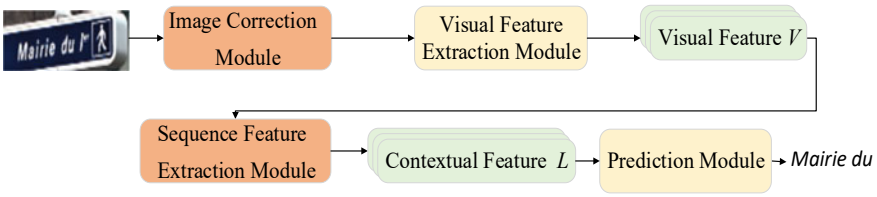

Fig. 3 Text recognition flow chart

The sample set is denoted by S in the equation above, and the maximum prediction probability can be reached by utilizing a dynamic programming approach to determine the ideal value of the loss function using the HMM's (Hidden Markov Model) Forward–Backward algorithm concept [24] as shown in Eqs. (5) and (6).

$$p(l|x) = \sum_{s=1}^{|l|} \frac{\alpha_t(s)\beta_t(s)}{y_{l_s}^t} \tag{5}$$

$$-\ln(p(l|x)) = -\ln\left[\sum_{s=1}^{|l|} \frac{\alpha_t(s)\beta_t(s)}{y_{l_s}^t}\right] \tag{6}$$

In the above Eq. (5) $\alpha_t(s)$, denotes the probability sum of all paths passing through character s at time step t at 1 minus time step t, and $\beta_t(s)$ denotes the probability sum of all paths passing through character s at time step t from t minus moment T. CRNN networks use mainstream convolutional structures, such as Resnet, MobileNet, etc. [25, 26], For the problem of a large amount of contextual information in the input data, CRNN introduces Bi-LSTM to enhance the contextual relationship modeling, and the final sequence is input to CTC for decoding, which avoids the problem of unaligned predictions and labels.

2.3 Inference Acceleration Engine

Due to the limited computing power and storage at the edge, models have to be deployed to the edge after training in the cloud, at which time they need to be pruned, quantized, etc. [27]. In order to assure tolerated accuracy loss following model compression and to avoid potential incompatibility issues when models are delivered to various edges, OpenVINO came into being. It is a collection of tools created by Intel based on its own hardware platform that helps speed up the creation of computer vision and deep learning applications. It includes a number of inference libraries, model optimization, and other deep learning-related materials. It can deploy algorithmic models online and is interoperable with models learned using a variety of open source frameworks. Figure 4 shows the workflow diagram. First, the model structure is selected in the training model phase, such as TensorFlow, Caffe, etc. The trained model is then converted into an intermediate representation that the inference engine can understand by using the model optimizer to improve the neural network model created in the previous phase. In the subsequent inference acceleration phase, the model is sped up for inference computation. Finally, the model optimizer and the inference engine are the core elements of the OpenVINO engine for producing user-ready cloud applications. The inference engine controls the loading and compilation of optimized neural network models while also supporting asynchronous operations. Model Optimizer, a cross-platform command-line tool [28],

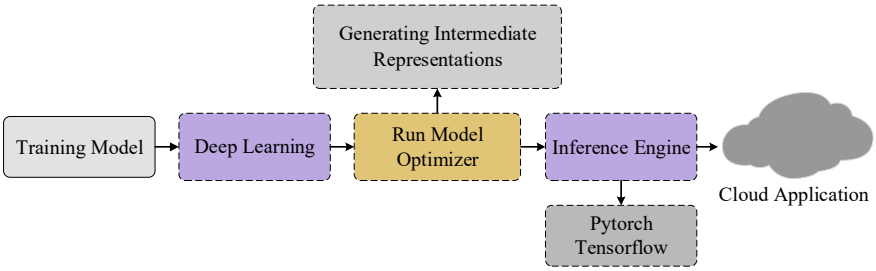

Fig. 4 Inference engine workflow diagram

converts the trained neural network from its source framework to an open-source intermediate representation for inference operations that is compatible with nGraph.

3 Improved Model

We address the technical issue that existing web opinion monitoring methods cannot rapidly acquire and identify textual information in visual multimodal scenes in this paper. Combining the characteristics of data in opinion analysis, the PP-OCR model deployed in the OpenVINO environment is improved and adapted based on the latest achievements in the current text recognition field. Before focusing on improvement, evaluate the pre-training model first and use the DB + CRNN combination model to determine whether the issue is with detection or recognition in the dense text images. try increasing image resolution and stretching the image within a specific range in order to sparse the text and enhance the recognition impact if the image contains small amounts of dense text. Improvements can be seen in two areas: (1) The LK-PAN network with a large feeling field is proposed in the detection module to upgrade the CML distillation strategy. (2) The transformer is introduced in the recognition module to mine the contextual data between text lines, while the original CTC decoder is changed to the Guided Training of CTC (GTC) method. Figure 5 shows the structure diagram of the improved model.

As shown in Fig. 5 above, the datasets of images and videos are first prepared in the first stage (Stage 1) as the model's input. The reading and inference of the input images and videos by the text detection and recognition model can be sped up by the second stage (Stage 2) of the inference acceleration engine, which primarily consists of two parts: the model optimizer and the inference engine. The third stage (Stage 3) then further improves the input data (image and video data) from the first stage to enhance the quality of the images and videos. In the fourth stage, strategies such as LK-PAN and DML will be added to the text detection module, as well as the Transformer network and GTC to the text recognition module (Stage 4). Finally, the experimental findings demonstrate that the enhanced model performs better on the dataset than the model without the enhancement.

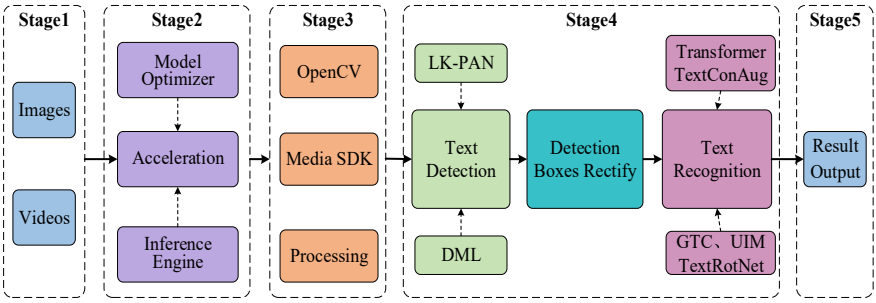

Fig. 5 Improved overall framework diagram

3.1 Detection of LK-PAN Networks

In this paper, we improve on the original PP-OCR V2's CML distillation strategy. It combines the traditional standard distillation of Teacher (Teacher) guiding students (Students) with Deep Mutual Learning (DML) mutual learning between the network of Students [29], as well as the features that the teacher network guides the students network while they are learning from each other. The LK-PAN of the PAN module with large sensory fields is used in this paper to optimize the teacher model. Figure 6 shows the LK-PAN framework diagram. First, the input feature information is extracted by ResNet50 network features, which are stacked by several similarly structured blocks that are the basic units of the residual network, i.e., residual blocks. This network can increase the model's training effectiveness and speed up model training. When the gradient and feature degradation problems are well solved as the model layers are deepened in order to obtain rich features, Secondly, after the LK-PAN module, the core is to increase the convolutional kernel in Path Augmentation (PAN), and the size of the convolutional kernel is expanded from 3*3 to 9*9 to enhance the perceptual field covered by each position in the feature map by increasing the size of the convolutional kernel. In image and video datasets, it can show good results for detecting text with large fonts and in text with extreme aspect ratios, while combining the LK_PAN network with the DML distillation strategy [30]. Finally, the information that passes through the LK-PAN network is concatenated.

In this paper, an Residual Squeeze-and-Excitation Feature Pyramid Network (RSE-FPN) with a residual attention mechanism is used for the students model [31]. The RSE-FPN framework is depicted in Fig. 7. The MobileNetV3 network processes the feature information first, automatically learning the relative relevance of each feature channel. The results are then used to boost helpful features and suppress features that are less useful for the present job based on the acquisition. The RSEConv network's subsequent layer receives its input from the feature data from the preceding stage. With the addition of a residual structure, this network transforms the convolutional layer in the FPN into a channel attention structure RSEConv layer with residual structure, which is better able to characterize the feature map. The feature data that has passed through the RSEConv network is finally concatenated.

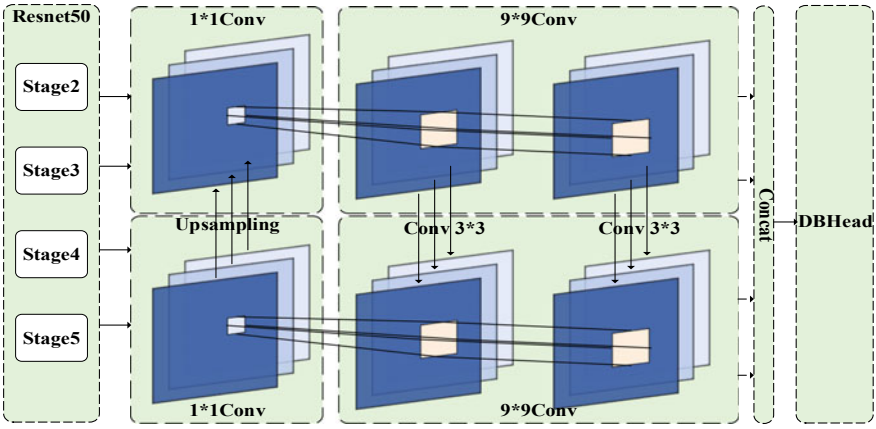

Fig. 6 LK-PAN framework diagram

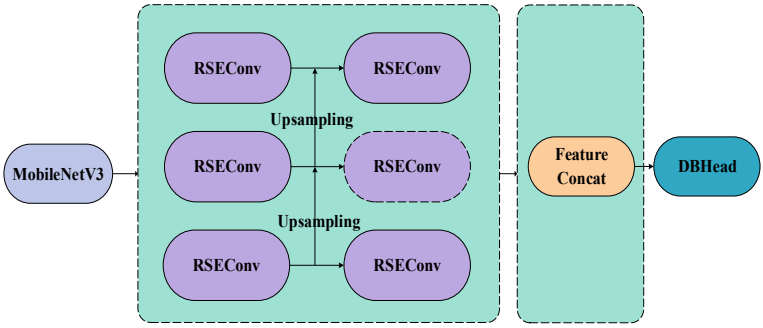

Fig. 7 RSE-FPN framework diagram

3.2 Improvement Strategy in Identification Module

To efficiently mine the contextual information of text line images, the recognition module ditches the RNN structure in favor of the Transformer network. The original CTC decoder can make quick inferences, but it has a low accuracy rate. Hence, the GTC technique was developed. As attention is more sensitive to spatial information and can precisely optimize the Spatial Transformer Network (STN) network [32], it is utilized to guide CTC training in order to facilitate the fusion of more features. The attention module is removed during inference without increasing inference time. First, a Gated Recurrent Cell (GRU) + attention module is added in the training phase, and the computational graph is partitioned to guide the learning of CTC. The computational flow is as follows:

The GRU is adopted to learn the attention dependency. At time-step t, The formula for x_t is shown in Eq. (7):

$$x_t = Softmax(W^T m_t) \tag{7}$$

where m_t is a hidden state of the GRU cell and W is a weight matrix. The hidden state m_t is updated via the recurrent process of GRU, As shown in Eq. (8):

$$m_t = GRU(y_{prev}, g_t, m_{t-1}) \tag{8}$$

where y_{prev} is the embedding vector of the previous output y_{t-1}. During training, y_{t-1} is replaced by the ground truth sequence. w_t represents the glimpse vector calculated As shown in Eq. (9):

$$w_t = \sum_{i=1}^{T} (\alpha_{ti} z_i) \tag{9}$$

where z_i is the feature sequence vector of $z^{1:T}$ at the time step i. L_t is the attention weight vector as follows Eq. (10):

$$L_t = Attention(m_{t-1}, z_i) \tag{10}$$

Our improved model uses the TextConAug module [33], which can be used to mine textual contextual information and thereby enrich the contextual information of the training data. Conventional data augmentation techniques include random flipping, cropping, adding noise, color scrambling, and other techniques, but their effectiveness is limited by the amount of original data. The quality of earlier data annotation efforts, which frequently included specialists in related subjects, was enhanced, but they were ineffective and expensive. The network makes use of the TextRotNet network [34], which employs a significant quantity of unlabeled data and is trained via a self-supervised technique [35], lowering the workload of labeled samples and significantly shortening the model training period without sacrificing recognition accuracy.

Finally, the text recognition module introduces the Unlabeled Images Mining (UIM) unlabeled data mining scheme, which has the ability to predict unlabeled data using recognition models with high accuracy, obtain pseudo labels, and use those with high confidence as training data for training models.

4 Systematic Experimental Process and Results Analysis

4.1 Experimental Environment

The experimental hardware configuration for this experiment in the Windows 10 environment is: central processing (CPU) using Intel (R) Core(TM) i5-7200U CPU @2.50 GHZ, 12G memory. The graphics processor (GPU) is NVIDIA Tesla P100 with 24G RAM, Python is the development language, and Pytorch is the framework.

4.2 Experimental Dataset

The model for text recognition on photos and videos is tested in this experiment using the free-to-use ICDAR 2015 dataset [13] and recorded video data, respectively. 500 test photos are randomly chosen, while 1000 images from the ICDAR 2015 dataset serve as the training samples. The video was captured for 1 min, 27 s, just for testing purposes. The original image and data annotation format is shown in Fig. 8 and Table 2 using an image from ICDAR 2015 as an example.

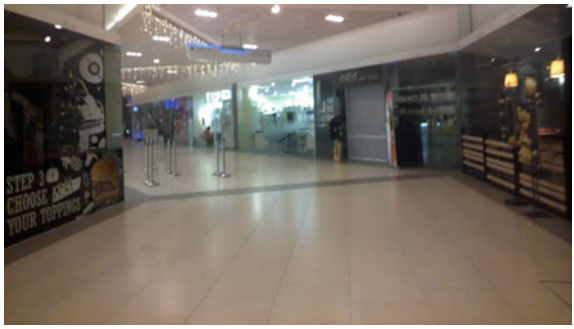

Fig. 8 A sample image from ICDAR 2015

Table 2 A sample image in ICDAR 2015 annotated with formatted text, which is corresponding to Fig. 8

Test image	Marked out text	Coordinate point
Img_110.jpg	STEP	[[1, 380], [71, 371], [74, 416], [4, 431]]
	CHOOSE	[[4, 432], [101, 405], [105, 450], [6, 476]]
	YOUR	[[3, 477], [76, 456], [80, 505], [4, 527]]
	TOPPINGS	[[76, 453], [174, 424], [178, 471], [78, 502]]
	ORO	[[750, 155], [795, 151], [795, 176], [750, 181]]

4.3 Evaluation Metrics

This experiment uses accuracy P, recall R, and F-value as its assessment metrics. The ratio of samples that were properly identified to all samples in the test dataset is known as the accuracy rate. Recall is defined as the proportion of accurate predictions to all accurate predictions in the test dataset [36]. F-value is a representation of the average of P and R. The formulae are shown in Eqs. (11), (12) and (13).

$$P = N\prime/N \tag{11}$$

$$R = N\prime/M \tag{12}$$

$$F = \frac{(\alpha * \alpha + 1)}{(P + R)\alpha * \alpha} P * R \tag{13}$$

where: N and M stand for the number of words projected to be correct overall and the number of words that were actually correct overall. N' stands for the number of words identified properly by the model. respectively. α represents learning rate, When $\alpha > 1$, The F-value is largely influenced by the recall rate. and when $0 < \alpha < 1$, The accuracy rate has a stronger impact.

4.4 Experimental Results

4.4.1 Comparison of Experimental Effects

Table 3 compares the network model's impact on the video dataset's detection speed before and after the enhancement. Table 3 shows that the model's average recognition speed for text in video has increased from 19.2 frames per second to 24.9 frames per second, a 29.7% improvement over the model's previous performance.

On the image and video datasets, Fig. 9a, b compares the accuracy, recall, and F-value effects of the network models both before and after the modification. Figure 9a, b shows that when Transformer network and the UIM unlabeled data mining scheme are added, the detection effect is slightly better than the model before improvement. The F-value of the text detection effect on the image dataset ICDAR2015 is improved

Table 3 Comparison of the effect of model detection speed before and after improvement

Test dataset	Network Model	Average detection speed (frames/s)
Recorded video data	Model before improvement	19.2
	Improved models	**24.9**

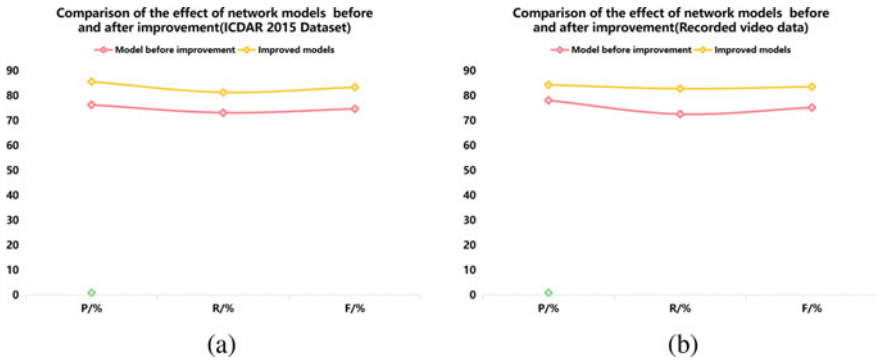

Fig. 9 **a, b** Comparison of the effect of network models before and after improvement

by 10.11%, and the F-value of the text detection effect in video is improved by 17.97%.

Table 4 compares the text detection effect in the image before and after the enhancement using a single image from the dataset ICDAR 2015, namely Fig. 8. Table 4 findings demonstrate that the upgraded model not only has a greater identification rate for "PPINGS" text but also a higher level of recognition confidence. As a result, the enhanced model's recognition capacity has significantly increased.

Table 4 Comparison of image text detection effect before and after improvement

Network model	Image text detection effect	Detection confidence
Model before improvement	STEP30	0.881
	OPPINGS	0.945
Improved models	STEP30	0.900
	CHOOSE	0.991
	PPINGS	0.939

The text detection effect of gambling video is compared in Fig. 10a, b before and after the enhancement and Fig. 11a, b show the comparison of the text detection effect of violent video before and after the improvement. The findings in Figs. 10 and 11 demonstrate that the enhanced approach used in this paper has a stronger identification capacity since it can accurately recognize more text in the video.

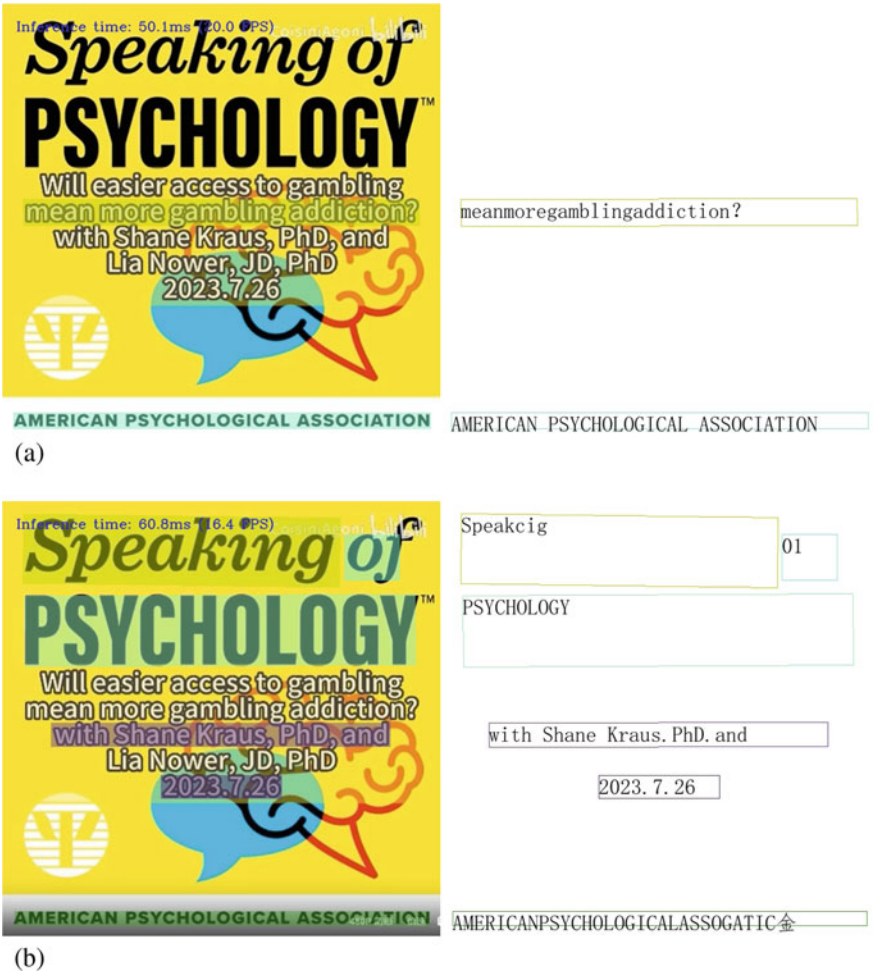

Fig. 10 **a** Experiment before improvement. **b** Experiment after improvement

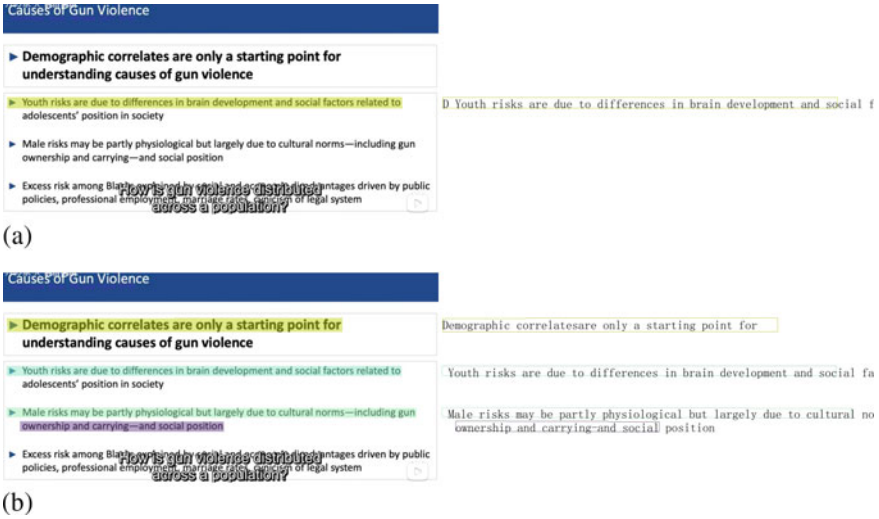

Fig. 11 **a** Experiment before improvement. **b** Experiment after improvement

5 Conclusion

In this paper, we proposed an improved method to visual multimodal text identification for scenarios such as internet opinion analysis. Furthermore, the method's validity has been experimentally validated. This paper makes three major contributions: Firstly, to improve the learning effect of the overall model, the text detection module upgrades the traditional distillation approach by integrating it with the DML mutual learning technique. Secondly, the large feeling wild PAN module is suggested in response to the prior shortcomings in identifying multi-scale and extreme aspect ratio text. Finally, the Transformer network is shown in the text recognition module to efficiently mine the contextual data of text line images to increase the text recognition's mistake correcting capabilities. The improved model enhances text identification in images and videos, and in the future, the identified text data will be combined with natural language processing methods. Currently, multimodal sentiment analysis has become a research hotspot, and we will further integrate the features of different modalities, effectively explore the relationship between features, improve the existing sentiment analysis models, and enhance the accuracy of sentiment analysis. In order to provide important technical support for the field of multimodal opinion analysis.

Acknowledgements The calculations were performed by using the high performance computing server in the Henan Key Laboratory on Public Opinion Intelligent Analysis.

Funding Statement Supported by the Key Research Projects of Henan Higher Education Institutions (No. 23A520031), Open Foundation of Henan Provincial Key Laboratory of Network Public Opinion Monitoring and Intelligent Analysis (No. HNPOL202101002) and the 2021 Henan Province Higher Education Teaching Reform Research and Practice Key Project (No. 2021SJGLX167).

References

1. Liu, J., Zhang, C., Sun, Y., et al.: Detecting text in the wild with deep character embedding network. In: Asian Conference on Computer Vision, pp. 501–517. Springer, New York (2018)
2. Chen, J., Lian, Z., Wang, Y., et al.: Irregular scene text detection via attention guided border labeling. Sci. China Inform. Sci. **62**(12), 1–11 (2019)
3. Baldi, P., Vershynin, R.: The capacity of feedforward neural networks. Neural Netw. **116**, 288–311 (2019)
4. Luo, C., Zhu, Y., Jin, L., et al.: Learn to augment: joint data augmentation and network optimization for text recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13746–13755 (2020)
5. Fang, S., Xie, H., Wang, Y., et al.: Read like humans: Autonomous, bidirectional and iterative language modeling for scene text recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7098–7107 (2021)
6. Yong, L.Q., Jia, W., Zhang, J.K.: Web opinion monitoring based on crawler technology and intelligent algorithm. Intell. Comput. Appl. **11**(04), 35–38 (2021)
7. Zhang, Y.-K., Zhang, H., Liu, Y.-G., et al.: Oracle character recognition based on cross-modal deep metric learning. Acta Autom. Sin. **47**(4), 791–800 (2021)
8. Du, Y., Li, C., Guo, R., et al.: PP-OCR: a practical ultra lightweight OCR system. arXiv preprint arXiv:2009.09941 (2020)
9. Du, Y., Li, C., Guo, R., et al.: PP-OCRv2: bag of tricks for ultra lightweight OCR system. arXiv preprint arXiv:2109.03144 (2021)
10. Castro-Zunti, R.D., Yépez, J., Ko, S.-B.: License plate segmentation and recognition system using deep learning and OpenVINO. IET Intell. Transp. Syst. **14**(2), 119–126 (2020)
11. Cao, Z., Hidalgo, G., Simon, T., et al.: OpenPose: realtime multi-person 2D pose estimation using part affinity fields. IEEE Trans. Pattern Anal. Mach. Intell. **43**(1), 172–186 (2021)
12. Wang, W., Xie, E., Li, X., et al.: Pvt v2: improved baselines with pyramid vision transformer. Comput. Vis. Media **8**(3), 415–424 (2022)
13. Liao, M., Zou, Z., Wan, Z., et al.: Real-time scene text detection with differentiable binarization and adaptive scale fusion. IEEE Trans. Pattern Anal. Mach. Intell. **34**, 11474–11481 (2022)
14. Liao, M., Wan, Z., Yao, C., et al.: Real-time scene text detection with differentiable binarization. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 11474–11481 (2020)
15. Xu, Y., Wang, Y., Zhou, W., et al.: Textfield: learning a deep direction field for irregular scene text detection. IEEE Trans. Image Process. **28**(11), 5566–5579 (2019)
16. Shi, B., Bai, X., Yao, C.: An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. IEEE Trans. Pattern Anal. Mach. Intell. **39**(11), 2298–2304 (2016)
17. Zheng, X., Chen, W.: An attention-based bi-LSTM method for visual object classification via EEG. Biomed. Signal Process. Control **63**, 102174 (2021)
18. Li, Z.-Y., Ge, H.-X., Cheng, R.-J.: Traffic flow prediction based on BILSTM model and data denoising scheme. Chin. Phys. B **31**(4), 040502 (2022)
19. Zhang, L., Zhao, Z., Ma, C., et al.: End-to-end automatic pronunciation error detection based on improved hybrid ctc/attention architecture. Sensors **20**(7), 1809 (2020)
20. Jiang, Y., Pan, J.Z., Chen, H.H., et al.: Traditional Chinese newspaper text detection based on segmentation methods. J. Jilin Univ. Eng. **2022**, 1–9 (2022)

21. Libing, G.: Research on natural scene text detection method based on feature fusion. Master's thesis, Shandong University (2021)
22. Vo, G.D., Park, C.: Robust regression for image binarization under heavy noise and non-uniform background. Pattern Recogn. **81**, 224–239 (2018)
23. Chen, T.B., Zhang, C.F.: Posterior probability map and complementary white model for secondary fusion of keyword recognition. J. Zhejiang Univ. Eng. **54**(06), 1170–1176 (2020)
24. Karahanoglu, N.B., Erdogan, H.: Compressed sensing signal recovery via forward–backward pursuit. Dig. Sig. Process. **23**(5), 1539–1548 (2013)
25. He, R., Liu, Y., Wang, K., et al.: Automatic cardiac arrhythmia classification using combination of deep residual network and bidirectional LSTM. IEEE Access **7**, 102119–102135 (2019)
26. Zhou, Y., Liu, Y., Han, G., et al.: Face recognition based on the improved MobileNet. In: Proceedings of the 2019 IEEE Symposium Series on Computational Intelligence (SSCI), pp. 2776–2781. IEEE (2019)
27. Ruan, X., Hu, W., Liu, Y., et al.: Model pruning based on dynamic sparse and feature learning enhancement. Chin. Sci. Tech. Sci. **52**(05), 667–681 (2022)
28. Qunhui, W., Yehua, W., Hao, W.: Key technology research and system construction of cross-level network, cross-architecture and cross-platform data sharing and exchange. Dual-Use Technol. Prod. **05**, 30–33 (2022)
29. Zhao, H., Yang, G., Wang, D., et al.: Deep mutual learning for visual object tracking. Pattern Recog. **112**, 107796 (2021)
30. Aguilar, G., Ling, Y., Zhang, Y., et al.: Knowledge distillation from internal representations. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 7350–7357 (2020)
31. Zhang, W.S., Zhu, Z.C., Zhang, Y.H., et al.: A cellular image segmentation method based on residual blocks and attention mechanism. J. Opt. **40**(17), 76–83 (2020)
32. Ghazi, M.M., Yanikoglu, B., Aptoula, E.: Plant identification using deep neural networks via optimization of transfer learning parameters. Neurocomputing **235**, 228–235 (2017)
33. Du, Y., Chen, Z., Jia, C., et al.: SVTR: scene text recognition with a single visual model. arXiv preprint arXiv:2205.00159 (2022)
34. Komodakis, N., Gidaris, S.: Unsupervised representation learning by predicting image rotations. In: International Conference on Learning Representations (ICLR) (2018)
35. Zhang, G., Nie, R., Cao, J.: SSL-WAEIE: self-supervised learning with weighted auto-encoding and information exchange for infrared and visible image fusion. IEEE/CAA J. Autom. Sin. **9**(9), 1694–1697 (2022)
36. Andreas, J., Rohrbach, M., Darrell, T., et al.: Learning to compose neural networks for question answering. arXiv preprint arXiv:1601.01705 (2016)