Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2024 SCPE. Volume 25, Issues 4, pp. 2664–2670, DOI 10.12694/scpe.v25i4.2978

MULTI-MEDIA IMAGE AND VIDEO OVERLAY TEXT EXTRACTION BASED ON BAYESIAN CLASSIFICATION ALGORITHM

LIANGLIANG YIN*AND ZIQIANG WANG†

Abstract. With the continuous development of Internet technology, using multimedia for virtual application has become a new way. In this paper, by introducing the bayesian classification algorithm, multimedia graphics and video for carding, testing and implementation of edge image of fine processing, matching the corresponding text for quick positioning, at the same time use the filter for image and video background picture, further classification, distinguish between background and text, the extraction of superposition character. Simulation results show that the bayesian classification algorithm is effective and improves the efficiency and accuracy of image and video processing.

Key words: Video text tracking, Video text segmentation, Multimedia information retrieval, Bayesian classification algorithm

1. Introduction. With the continuous development of social economy, the presentation modes of multimedia have become rich and diversified [1,2]. However, people have increasingly high requirements for the quality of multimedia, so how to carry out effective text overlay and improve the readability of multimedia images and videos has become a key difficulty [3,4,5]. Superimposing text on multimedia images and videos makes it easier to read or view multimedia, and easier to spread it. In multimedia text extraction, often can be divided into two categories according to the segmentation method, one is the use of static multimedia images or video screenshots for detection and extraction; The other is to use dynamic, multi-frame multimedia images and videos for text detection [6,7,8]. Bayesian classification method has solid mathematical theoretical foundation and is a practical method for remote sensing image classification. However, Bayesian serial classification algorithms based on CPU are very time-consuming when processing larger images.

Since the text in the multimedia video is artificially added in the later stage, the general existence period is about 3 seconds in order to be recognizable [9,10]. Therefore, it is more important to extract text in multimedia images and videos, that is, to extract text effectively within the 3-second time range when text appears. To determine the importance of different key features extracted, some scholars build the corresponding weighting function to calculate the different weights of each feature to distinguish the contributions of different feature categories to improve the performance of the naive Bayes model. The industry has also made a lot of attempts, such as the use of text matching algorithm for detection, first multimedia detection, then according to the existing samples for monitoring, and finally achieve two-step text extraction; The similarity of gray value of characters is used for quantitative calculation and differentiation, so as to search the text area, and multiple multimedia videos or images are used for unified correction, so as to realize the text tracking. In addition, some scholars use the linear analysis method to enlarge the relevant text in multimedia to reduce the difficulty of recognition and improve the accuracy of recognition [11,12,13,14]. In recent years, many studies at home and abroad have used GPU in the fields of remote sensing image fusion, hybrid image decomposition and real-time target detection, to accelerate the processing of the algorithm and obtain efficient parallel computing power. The graphics processor is computing intensive, highly parallel, small size and high cost performance, which provides favorable conditions for solving data-intensive computing.

However, in general, these studies still have some limitations. For example, the computational amount is relatively complex, especially for the global search, which is time-consuming and laborious, and the efficiency cannot be improved. For a large number of multimedia backgrounds, it is still difficult to distinguish them well,

[*] School of Big Data Science, Hebei Finance University, Baoding, 071051, China
[†] Jiangsu Ocean University, Jiangsu, 222005, China (Corresponding author, 20161219490jou.edu.cn)

especially for large similarity (small change in gray value), and quantitative filtering cannot be achieved [15]. Data analysis and mining have become the focus of research in the field of text classification. The classification of naive Bayes algorithm is simple and efficient, but the classification effect of the algorithm reduces the effect of the algorithm. Therefore, most scholars have done extensive research and improvement to improve the classification performance of naive Bayes.

Therefore, for the static text in multimedia image and video, this paper introduces bayesian classification algorithm, fast retrieval is built with new ways of tracking, secondly, using the corresponding matching edge features, and then, according to quickly find mode fixing the beginning and end, according to the bayesian classification to get the edge of the chart to distinguish, improve text region segmentation effect, Aims to explore the superimposed text extraction of multimedia images and videos.

2. Bayesian classification algorithm. The bayesian classification algorithm is stable, simple and easy to learn, and has good effect. It is widely used in various text classification. In particular, the feature words in commodity description texts or critical texts are generally short and concentrated, which are suitable for adopting naive Bayesian model. Bayesian classification method is a kind of classification method based on statistical model, and assumes the condition between each object, its principle is based on a local prior probability, using the Bayesian formula to calculate the posterior probability, namely the area belongs to a certain class of probability, choose the maximum posterior probability as the area belongs to the class.

The model is obtained from the learning of historical data, and then the model is used to judge the classification of text[16-20]. Bayesian formula is one of the core algorithms of machine learning. The occurrence of any event is not completely accidental, and is often based on the occurrence of other events. The general conditional probability is to determine the effect from the cause, and the posterior probability is to determine the cause from the effect. The Bayes formula in probability theory is shown in Equation 2.1:

$$P(B|A) = \frac{P(B|A)P(B)}{P(A)} \tag{2.1}$$

Type of $P(B|A)$ as the prior probability, said the incident occurred on the basis of the event B: the probability P (B) as the A posteriori probability, is based on past experience and analysis of the probability of event B.

In text classification technology,B represents category and text feature. When $B = \{B_1, B_2, ..., B_n\}$ is a complete group, its sum is a complete set; $A = \{A_1, A_2, ..., A_m\}$ represents a set of text features. Convert Equation 2.1 to the following form:

$$P(B_i|A_1, A_2, ..., A_m) = \frac{P(B_i) \prod\limits_{t=1}^{m} P(A_t|B_i)}{\sum\limits_{j=1}^{n} P(B_j) \prod\limits_{t=1}^{m} P(A_t|B_j)} \tag{2.2}$$

Naive Bayesian model is to choose the highest posterior probability. From Equation 2.2, the denominator is the same for all categories, so the numerator must be the largest in order to have the highest posteriori probability. The naive Bayesian model is described as:

$$B_{max} = argmax[P(B_x = B_i) \prod_{t=1}^{m} P(A_t|B_x = B_i)] \tag{2.3}$$

Equation 2.3 indicates to find a value such that the value of $P(B_i) \prod\limits_{t=1}^{m} P(A_t|B_i)$ is the largest at $B_x = B_i$.

The problem of zero probability may exist in actual data training. The so-called zero probability problem is when the posterior probability is calculated, the component of an eigenvector never appears in the training set, and the whole posterior probability is calculated to be zero. In Equation 2.2, as long as one $P(A_t|B_i)$ in $\prod\limits_{t=1}^{m} P(A_t|B_i)$ is 0, it will cause the whole formula to be 0. To solve this problem, you do Laplacian smoothing, you add one to the vector that never happens, so Laplacian smoothing is also called plus one smoothing. It can be expressed by Equation 2.4 :

$$P(A_t|B_i) = \frac{|D_{A_t,B_i}| + 1}{|D_{B_t}| + N} \tag{2.4}$$

where $P(A_t|B_i)$ represents the probability that a feature A_t belongs to B_i ; D_{A_t,B_i} represents the occurrence times of A_t of a certain feature under B_i classification. D_{B_t} represents the number of samples belonging to the B_i classification, and N represents the number of values of the characteristic J A_t.

The more times the word appears in the corpus, the less important it is. Colloquial understanding is that a word is not common (low IDF), but it appears in the article of high frequency (high TF), then it is likely to be the key word of the article. Tf-idf algorithm is defined as follows:

$$TFIDF_{t,j} = TF_{t,j} * IDF_t \tag{2.5}$$

Equation 2.5 refers to the TF-IDF weighted value of a word in a file, where $TF_{t,j}$ is the number of times that a word c_t appears in the file d_j, which is represented by Equation 2.5; IDF_i is the inverse frequency of the word, indicated by the Equation 2.6.

$$TF_{t,j} = \frac{n_{t,j}}{\sum N_j} \tag{2.6}$$

where $n_{t,j}$ is the number of times the word c_t appears in the file, and N_j is the total number of words in the file.

$$IDF_t = log\frac{|D|}{|\{j : c_t \in d_j\}| + 1} \tag{2.7}$$

As previously analyzed, if a word does not appear in the file, the denominator will be 0, so Laplace smoothing is used to add 1. In general, the longer the word is, the more explicit the information it expresses. Therefore, tF-IDF is improved by considering the weight of word length W_i, W_i which is the ratio of the word length of this word to the longest length of feature words in the document, and Equation 2.8 is obtained.

$$TFIDF_{t,j} = TF_{t,j} * IDF_t * W_i = \frac{n_{t,j}}{\sum N_j} * log\frac{|D|}{|\{j : c_t \in d_j\}| + 1} * \frac{L_t}{L_{max}} \tag{2.8}$$

In the traditional Bayesian classification algorithm, all text feature vectors have the same status and are equally important for classification decision. However, in practice, there are some redundant or modal words, which are irrelevant to classification and are polluted by noise, which reduce the accuracy of classification. In view of this situation, tF-IDF feature weighting is used to improve the traditional Bayesian classification algorithm, and the following formula is obtained after taking logarithm:

$$B_{max} = argmax[P(B_x = B_i) \times \sum_{i=1}^{m}(logP(A_t|B_x = B_i) + logTFIDF_{t,j})] \tag{2.9}$$

3. Fast tracking and segmentation of video text with multi-frame edge information. In order to effectively extract the text of multimedia image and video overlay, this paper divides the Bayesian algorithm to better extract the multimedia overlay text, which is mainly divided into two aspects: tracking and detection. Specifically, it can be divided into three steps:

1. Multimedia image and video tracking, set a certain period of time, respectively calculate the Bayesian classification method;
2. Track the superimposed text. If the superimposed text is detected in the first step, the starting and ending position of the text will be tracked.
3. Multiple multimedia images and videos are fused so that more edge features can be extracted, which is conducive to classification and segmentation.

3.1. Text monitoring and verification. Calculate the overlapping area of the text region detected in the two frames of multimedia. Set RA_i as the ith text region of the reference frame, and RB_j as the region corresponding to the jth text region of the reference frame in the verification frame, then the overlap ratio S_o of the text region of the two frames is shown in Equation 3.1:

$$S_o = |RA_i \cap RB_j|/|RA_i| \tag{3.1}$$

Calculate the similarity S_e of the Edge Map of the corresponding text area of the two frames, which can be calculated by Equation 3.2 and Equation 3.3:

$$S_e = \frac{\sum\limits_{(x,y)\in RA_i \cap RB_j} (EM_t(x,y) \times EM_{t+10}(x,y))}{\sum\limits_{(x,y)\in RA_i \cap RB_j} EM_t(x,y)} \tag{3.2}$$

$$EM = \begin{cases} 1, \text{if pixel(x,y)is edge,} \\ 0, otherwise \end{cases} \tag{3.3}$$

The corresponding pixel brightness $L_{i-10}(x,y), ..., L_{i+10}(x,y)$ of 10 frames before and after the reference frame is fused as shown in Equation 3.4:

$$L_m(x,y) = min(L_{t-10}(x,y), ..., L_{t+10}(x,y)) \tag{3.4}$$

The number of edges generated after edge detection will be greatly reduced, so the false detection will be further reduced.

3.2. Fast tracking algorithm for static text. After monitoring and verifying the text object, a new text object is created. The text object is next tracked, and since there is no position change in the static text area, there is no need to search in adjacent frames. In this paper, a method based on binary search method is proposed to determine the start and end frames of text region. The similarity of edge bitmap is used as the matching feature. If the similarity (S_e) between the edge bitmap of the corresponding region in the current frame and the edge bitmap of the literal object is greater than a certain threshold (T_s), the match is considered successful. In the following pseudo-code form, the process of searching forward text area for the end frame is given. The algorithm of searching backward for the start frame only needs to reverse the search direction. The algorithm first searches for the lower bound of the end frame and then searches for the lower bound of the end frame.

```
① First set end ref + track step, last match = ref to ②;
② Calculate the edge graph of $Frame_{end}$ and $S_e$ corresponding to the text area in Framere If , end = end +
      track step, last match = end, turn to ②; otherwise, step = end last match, turn to ③;
③ step step/2  If step < tolerance  Output last match as the forward boundary of the text object. Otherwise,
      mid = last match + step, calculate the edge graph of frameid and $S_e$ corresponding to the text area
      in frameref. If $S_e > T_s$, last match = mid, otherwise end = mid, turn to ③.
```

In the experiment, this paper takes the threshold T_s as 05, the step track step of searching the lower bound is 150 frames, and the minimum step tolerance of searching the lower bound is 1.

4. Simulation experiment and analysis.

4.1. Experimental data set and evaluation criteria. The simulation experiment mainly includes two parts: one is to detect the effectiveness of Bayesian classification algorithm, that is, to detect multiple images or videos of multimedia; The other is to detect the detection results of superimposed text, especially the matching results by using the classified edge features. According to the needs of simulation experiments, different multimedia images and videos (3) are selected as data.

For the extraction of superimposed text, the following indicators are mainly used:

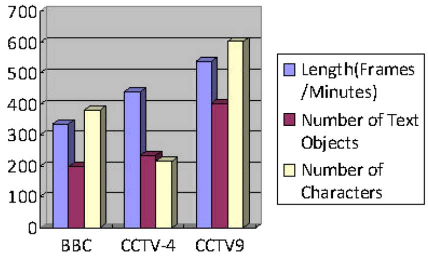

Fig. 4.1: Video type and text object (data sample)

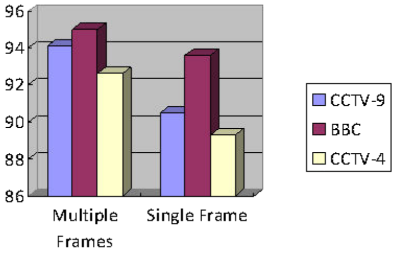

Fig. 4.2: Experimental results of text detection and tracking

(1) Recall: the ratio between the number of superimposed characters detected by the method and the actual number of characters.

(2) False alarm rate: that is, the ratio between the number of superimposed characters (false detection) detected by the method and the actual number of characters.

For the evaluation of characters, character recognition rate and accuracy are mainly used. The recognition rate is the ratio of the correct characters to the actual total number, and the accuracy is the ratio of the correct characters to the total number.

4.2. Text object detection and tracking algorithm experiment. The Bayesian classification algorithm is calculated and segmented for Figure 4.1. The specific results are shown in Figure 4.2. From the results, the text detection effect of multimedia multi frame image or video is much greater than that of single multimedia image. In addition, it can improve the accuracy of corresponding text positioning.

This paper compares the text calculated by Bayesian classification algorithm with the results of manual annotation and uses the corresponding results to calculate the error between them. Due to the effectiveness of Bayesian classification algorithm, the edge features provided are more accurate. Therefore, the error between them is very small, that is, the judgment accuracy of starting position and ending position is very high.

4.3. Experiment of multi frame text region enhancement and segmentation algorithm. On the basis of the above simulation experiment, continue to fuse multiple multimedia images or videos for superimposed text extraction of Bayesian classification algorithm. The results of the simulation experiment are shown in Figure 4.3. From the results, the recognition rate and accuracy extraction of superimposed text are significantly improved after the fusion of multiple multimedia images.

The Bayesian classification method is used to distinguish the background and superimposed text of multimedia images and videos, track and detect the superimposed text in the form of blocks, and segment the superimposed text and image according to the threshold method. Specifically, as shown in Figure 4.4, a certain threshold value is selected according to local or overall.

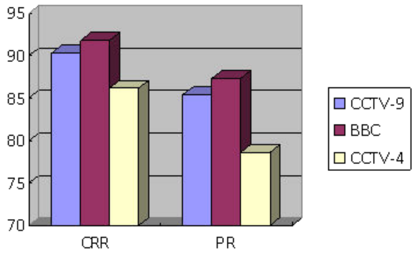

Fig. 4.3: Experimental results of text segmentation

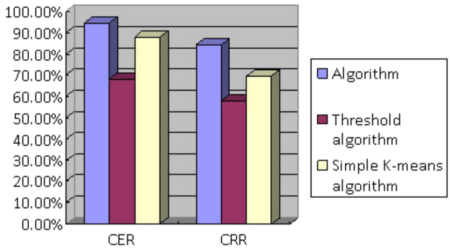

Fig. 4.4: Performance comparison of algorithms

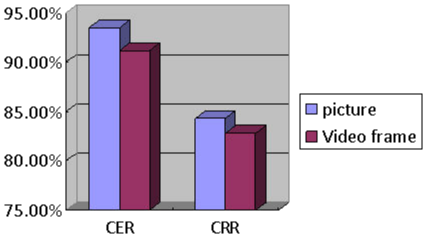

Fig. 4.5: Different image segmentation capabilities

Through the establishment of Bayesian classification method, the text is segmented accordingly. Among them, the text in multimedia image and video is divided into one kind and the other background is divided into one kind by Bayesian classification method. On this basis, clustering processing is carried out. However, it should be noted that if the background color of text is close to that of multimedia image and video, Further classification is needed. The recognition results are shown in Figure 4.5. Bayesian classification algorithm (CER) has high accuracy.

5. Conclusions. The Bayesian net is an important tool for processing uncertainty information, providing a way to represent causal information that has been successfully used in medical diagnosis, statistical decision making, expert systems, etc. With the continuous development of mobile Internet, people have higher and higher requirements for multimedia video and images. In order to better provide high-quality services, this paper attempts to introduce Bayesian classification algorithm. By combing the patterns of text extraction superimposed on multimedia images and videos, the two-stage model is determined by using the correlation model, that is, the text is tracked first and then detected; Secondly, the search method is used to quickly locate the beginning and end of the text; Finally, the background is filtered according to the edge feature results obtained by Bayesian classification method to realize the effective extraction of superimposed text. Simulation results show that Bayesian classification algorithm is effective, can improve the effect and accuracy of segmentation, and lay a foundation for multimedia image and video processing. In further work, we can explore the possibility of applying other Bayesian net types to target extraction, and then better represent the relational constraints between target and environment.

Data Availability. The data used to support the findings of this study are available from the corresponding author upon request.

Funding Statement. This study did not receive any funding in any form.

REFERENCES

[1] Sravani M, Maheswararao A, Murthy M K. Robust detection of video text using an efficient hybrid method via key frame extraction and text localization[J]. Multimedia Tools and Applications, 2021, 80(3):1-16.

[2] Lei, Tao, Liu, et al. A Method of Effective Text Extraction for Complex Video Scene[J]. Mathematical Problems in Engineering: Theory, Methods and Applications, 2016, 6(7): 1-9.

[3] Goto H. Versatile Text Extraction System for Text-to-Speech Reading Assistant Camera[J]. Studies in Health Technology and Informatics, 2015, 217(5):392-397.

[4] Lee J, Park J S, Hong C P, et al. Illumination-Robust Foreground Extraction for Text Area Detection in Outdoor Environment[J]. Ksii Transactions on Internet & Information Systems, 2017, 11(1):345-359.

[5] D Brodić. Text Line Segmentation With Water Flow Algorithm Based on Power Function[J]. Journal of Electrical Engineering, 2015, 66(3):132-141.

[6] Neshov N, Popova A, Garcia J, et al. Finding URLs in images by text extraction in DCT domain, recognition and matching in dictionary[J]. International journal of reasoning-based intelligent systems, 2015,3(6):46-57.

[7] Bontempi L, Vassanelli F, Cerini M, et al. Video-Assisted Thoracoscopic Monitoring of Laser Lead Extraction by Femoral Route[J]. Innovations Technology & Techniques in Cardiothoracic & Vascular Surgery, 2018, 13(3):233-235.

[8] Alotaibi S S. Optimization insisted watermarking model: hybrid firefly and Jaya algorithm for video copyright protection[J]. Soft Computing, 2020,5(2):98-105.

[9] Shetty S, Devadiga A S, Chakkaravarthy S S, et al. Optical Character Recognition for Alphanumerical Character Verification in Video Frames[J]. Advances in Intelligent Systems & Computing, 2015, 324(5):81-87.

[10] Mentzer N, Paya-Vaya G, Blume H. Analyzing the Performance-Hardware Trade-off of an ASIP-based SIFT Feature Extraction[J]. Journal of Signal Processing Systems, 2016, 85(1):83-99.

[11] Ayed A B, Halima M B, Alimi A M. MapReduce Based Text Detection in Big Data Natural Scene Videos[J]. Procedia Computer Science, 2015, 53(1):216-223.

[12] Gambhir M, Gupta V. Recent automatic text summarization techniques: a survey[J]. Artificial Intelligence Review, 2016, 47(1):1-6.

[13] Dang L M, Hassan S I, Im S, et al. Utilizing text recognition for the defects extraction in sewers CCTV inspection videos[J]. Computers in Industry, 2018, 99(4):96-109.

[14] Papastratis I, Dimitropoulos K, Konstantinidis D, et al. Continuous Sign Language Recognition Through Cross-Modal Alignment of Video and Text Embeddings in a Joint-Latent Space[J]. IEEE Access, 2020, 5(99):1-10.

[15] Shi X, Feng Z, Lei L, et al. Textural feature extraction based on time–frequency spectrograms of humans and vehicles[J]. Radar Sonar & Navigation Iet, 2015, 9(9):1251-1259.

[16] Tuna T, Subhlok J, Barker L, et al. Indexed Captioned Searchable Videos: A Learning Companion for STEM Coursework[J]. Journal of Science Education & Technology, 2017, 26(1):1-18.

[17] Iaas A, Mm A, Wan A, et al. Web Data Extraction Approach for Deep Web using WEIDJ[J]. Procedia Computer Science, 2019, 163(5):417-426.

[18] Tomiyasu F, Wang X, Mase K. Video cut extraction method for wide-angle multi-view videos using spatial relationship between ball and cameras[J]. Journal of the Institute of Image Electronics Engineers of Japan, 2016, 45(3):305-317.

[19] Koehler K, Eckstein M P. Temporal and peripheral extraction of contextual cues from scenes during visual search[J]. Journal of Vision, 2017, 17(2):90-100.

[20] Na I S, Le H, Kim S H, et al. Extraction of salient objects based on image clustering and saliency[J]. Pattern Analysis and Applications, 2015, 18(3):667-675.

Edited by: Zhigao Zheng
Special issue on: Graph Powered Big Aerospace Data Processing
Received: Dec 5, 2023
Accepted: Dec 25, 2023