Journal of Ambient Intelligence and Humanized Computing
https://doi.org/10.1007/s12652-019-01599-2

ORIGINAL RESEARCH

Automatic text location of multimedia video for subtitle frame

Qingmei Lu[1] · **Yulin Wang**[2]

Received: 25 June 2019 / Accepted: 23 November 2019

Abstract
The traditional multi-media video text auto-positioning is too dependent on man-made function, and its disadvantages are mainly embodied in the aspects of strong subjectivity, large quantity of work, slow processing speed and the like. Through the establishment of the basic framework of multimedia video text search, the feature vectors of multimedia video text are calculated. On the basis of the feature of the independent video captioning (ICA), the method of automatic positioning of the multimedia video text is studied. The method comprises the following steps of: extracting a plurality of ICA image features of an independent multimedia video text sub-block by adopting an independent video caption-based ICA feature extraction method; performing image candidate character line automatic detection based on a plurality of ICA image features, adopting an adaptive iterative positioning algorithm to quickly obtain a character candidate line, In the process of automatic text block detection, the flaw of text candidate line detection is checked, and the text candidate block is obtained by threshold calculation. The experimental results show that the proposed method has the highest accuracy in recognizing video subtitles with independent caption based ICA features, and the recall and precision of the text lines and blocks of the video image are located by the proposed method, and good automatic positioning effect.

Keywords Subtitles frame · Multimedia · Video text · Automatic positioning · Independent subtitle base · Adaptive iteration

1 Introduction

In recent years, with the rapid development of electronic technology and multimedia technology, multimedia video text has become an indispensable way of expressing information in people's lives. The continuous development of multimedia information technology is the foundation for the rapid advancement of the digital video industry (Armstrong 2017). In the civil field, the widespread use of digital cameras, mobile phones and video cameras has simplified the way people record their lives. In the industrial field, remote sensing satellite monitoring, video surveillance, film and television entertainment, etc., produce a large amount of video, images and other data at all times. Multimedia video often has low utilization rate, so in order to obtain effective multimedia video information, it must be based on video image analysis (Lu et al. 2017; Gao et al. 2018). Traditional multimedia video retrieval relies too much on human, and its main shortcomings are reflected in the aspects of strong subjectivity, large engineering volume and slow processing speed. Therefore, seeking high-quality multimedia video automatic positioning method has become a hotspot for relevant researchers. Effective analysis of text data in multimedia video is the basis of fast video retrieval. Automatic extraction and recognition of text information is an important channel for Multimedia video retrieval (Głomb et al. 2011).

Yin et al.(2019) proposed a new method based on Adaboost for video text localization. The connected domain in the video image is extracted. After analyzing the video text region, the five types of features of the video text are extracted. Using these five types of features, the Adaboost strong classifier is constructed by classification and regression decision tree. Send the candidate text area to the strong classifier to get the correct text area. This method has a good positioning effect on the text, image size and color of the video frame image. However, the method has a low

✉ Qingmei Lu
luqingmei113@163.com

[1] Data Science and Technology, North University of China, Taiyuan 03005, China

[2] Computer School, Wuhan University, Wu Han 430072, China

Published online: 05 December 2019

recall rate. Abdulrahim and Salam (2016) proposes a subtitle localization algorithm based on the rich feature of subtitle area edge information. The algorithm can accurately locate the four boundaries of the top, bottom, left and right of the subtitle, delete unnecessary background areas, and reduce the amount of subsequent operations. Text positioning is performed by comparing the degree of overlap to determine a number of frames taken out. The identification method has high positioning accuracy, but the recognition accuracy of the method is low. Yan et al. (2018) proposed a method combining Harris corner detection and corner density. The corner distribution obtained by Harris algorithm was used to filter out the background corner points, and the segmentation points were determined according to the distance between corner points. In this method, heuristic rules and connected domains are combined to obtain the final text. This method has high positioning accuracy, but poor positioning effect. Liu et al. (2017) proposes a two-stage subtitle detection and extraction algorithm. The subtitle frame and the subtitle area are separately detected, thereby improving detection efficiency and accuracy. The first stage performs subtitle frame detection, and the second stage performs subtitle area detection and extraction on the subtitle frame. This method can reduce the number of frames to be detected, but its operability is poor. Wang and Pan (2017a, b) proposed a multi-feature-based keyword extraction algorithm, TFL-WS algorithm. By analyzing the characteristics of the video containing rich related text information, considering the attributes of the candidate words such as part of speech and word span, the extended synonym word forest is used to extract the keywords. The extracted content is expressed by the extracted keywords. But the extraction speed of this method is slow.

Therefore, in order to improve the accuracy of automatic text positioning of multimedia video, this paper proposes a multimedia video automatic text positioning method for subtitle frames. Based on the independent video subtitle ICA feature of multimedia video text, the method adopts adaptive iterative localization algorithm, which can quickly obtain candidate lines and candidate blocks of video text.

2 Multimedia video text search

2.1 Basic framework of multimedia video text search

The metadata that video has is a summary of multimedia video text. It is closely related to video content, and it plays an important role in the analysis and interpretation of video. Metadata is an important way to obtain multimedia video text, and it is also one of the important information sources for multimedia video text search (Liu et al. 2019a, b).

Before performing multimedia video text positioning, the text need to be searched first. The search process uses Lucene's text indexing and search functions. It is not a complete program, but a software library that allows users to add search functionality to their applications via Lucene (Ghosh et al. 2018). Lucene can index and search data parsed in text without having to care about the source, format, or even language of the data. As long as it can be converted to text format, it can index and search document data in various formats. The search program has many applications: it can be used as a backend component to search for specific sets of content (such as mail messages, local files, and so on). Run on a Web site server to search for product listings, determine document sets, and so on. For most Internet content indexing, processing large-scale search requests, the search architecture of the search program is shown in Fig. 1.

The first function that needs to be implemented is the indexing operation, which needs to be completed in several separate steps (shown in the left half of Fig. 1): Get the original text metadata; Parsing the original text metadata, extracting text information, such as parsing binary files and Word documents, and converting them into text files that are easy to index; Index the created document. After the document is indexed, the search operation can be implemented on the index. The search process (shown in the right half of Fig. 1) includes: User interface; Method of constructing a programmable query statement; Execute the query statement (retrieve the matching document); Show query results. The module in Fig. 1 is implemented by Lucene. The following is a brief introduction to each module in the figure:

1. Acquire Content: It can be collected by web crawlers or spider programs. This step is not necessary, such as

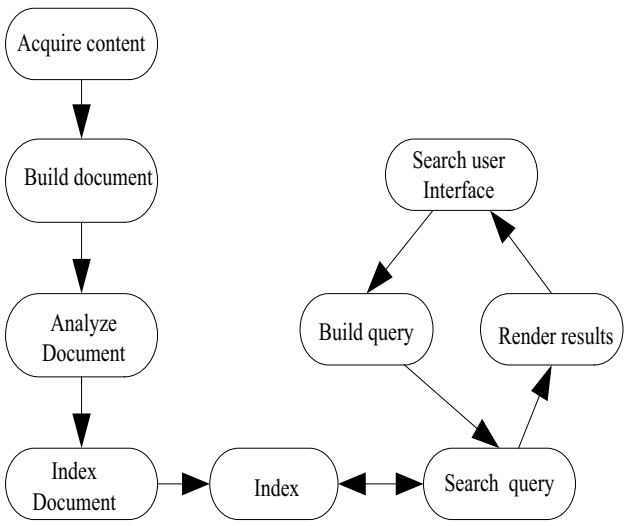

Fig. 1 Basic framework of text search engine

indexing a set of XML documents that already exist on the file system.

2. Build Document: Before searching for the acquired content, a document (also called a component) needed to be created for the search engine to use. The document mainly includes several fields with values, such as title, author, main body, abstract, and so on. This requires designing how the captured content is split into appropriate documents and domains, usually one document, one web page, and one domain. After the scheme is designed, the text in the original content needs to be extracted and written into each document, including parsing binary files (PDF documents, Word documents, etc.) and removing fixed tags for some text encoding format files (XML, HTML, etc.).
3. Analyze Document: The main task is to divide the text into independent lexical units (equivalent to "words"), including the processing of linked words, grammatical corrections, whether to insert synonyms into the lexical units, word stem extraction, and so on. Lucene provides a number of profilers that can easily control these steps.
4. Build Query: Process the text entered by the user into the query object according to the set query syntax. Query statements include Boolean operations, phrase queries, wildcard queries, and weighting for multiple subqueries, etc. Lucene provides a powerful development package called a query parser to do this.
5. Search Query: Indexes are retrieved with queries that return results that are collected, filtered, and sorted by query request. The specific operation is shown in Fig. 2.

Multimedia video text search method is similar to text retrieval method (Xu et al. 2018): In the pre-processing stage of text query analysis, the keyword text information in the query is obtained. Use these keyword text information as the smallest text query units, and then search against those query units. Search for text in multimedia video texts by TRECVID Known-item Search (Bakas et al. 2017). It includes a lot of irrelevant multimedia video text information, which needs to be analyzed to determine the search strategy, and then the text statement is preprocessed. Extract key text statements from multimedia video and search through these key text. Extracting key text can also be thought of as removing words that are not search-related from key text. Multimedia video text search is carried out by extracting nouns, verbs and adjectives that can be used for searching in text statements as keywords.

In multimedia video text search, users can express their search intention directly through text keywords or free text. The expression of the user's search content query intention is expressed in different query input methods. The choice of query input method is directly related to the subsequent description of multimedia video text data content and semantic correlation calculation method.

2.2 Multimedia video text feature vector calculation

Text search usually uses an exact matching method, while multimedia video text search is done by similarity matching between visual features (Sangaiah et al. 2019; Ren et al.2018; Lu and Liu 2019). Therefore, choosing the appropriate visual feature similarity measure has a great impact on the search results. The multimedia video text feature vector can be calculated using the similarity measure method, the MEL frequency cepstrum coefficient, and the voiceprint-based broadcaster Gaussian mixture model method.

2.2.1 Similarity measurement methods

(1) Vector space model

The elements of the vector feature are independent and of the same importance, and the D_1 distance between the two vectors A and B is expressed as:

$$D_1 = \sum_{i=1}^{N} |A_i - B_i| \tag{1}$$

D_2 distance is expressed as:

$$D_2 = \sum_{i=1}^{N} \left|(A_i - B_i)^2\right|^{1/2} \tag{2}$$

where N is the dimension of the feature vector.

Under certain conditions, the cross-entropy of the probability distribution of text keywords (Kullback–Leibler Divergence) (Shum and Liao 2001), and the local characteristics of the multimedia video text can be expressed as:

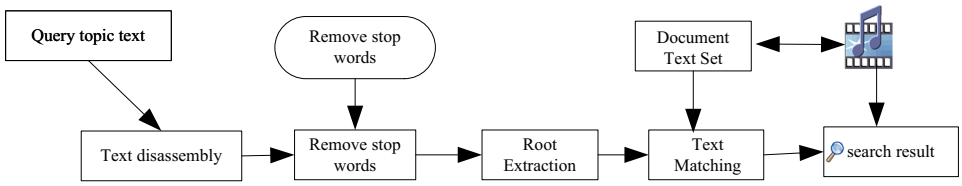

Fig. 2 Multimedia video text search process

$$v_i = [t_i, t_{i2}, \ldots, t_{iD}]^T$$
$$t_{id} = \frac{n_{id}}{n_i} \log \frac{N}{n_d} \tag{3}$$

where, n_{id} The number of times that d video text appears in x_i, and n_i is the sum of the occurrences of all text information in x_i; n_d is the number of d text information in the text information collection; N is the number of fixed text messages in the video text collection. The local feature similarity between text x_i and text x_j is calculated by the Cosine similarity function (Korzun et al. 2015; Kurzhals et al. 2016):

$$S_i(x_i, x_j) = \frac{v_i^T v_j}{\|v_i\|} \tag{4}$$

where $\|v_i\|$ is vector modulus function.

Assumed text search probability $p(r = 1|tx)$, the k–neighbor method can be used to calculate the point x relative to the set of χ/x, the local visual content consistency of the K' nearest neighbors is estimated:

$$p(r = 1|tx) = \frac{1}{k'} \sum_{i=1}^{k'} s(x, x_i) \tag{5}$$

where $x_i \in \chi/x$, $i = 1, \ldots, k'$ is x in $\chi \backslash x$ K' nearest neighbors. To simplify parameter settings, make $K = K'$, $s(x, x_i)$

2.2.2 The MEL frequency cepstrum coefficient (MFCC)

The frequency cepstrum of the speech signal can characterize the speaker's channel and excitation source to reflect the physiological differences of the speaker. In order to better reflect the characteristics of human hearing, the frequency cepstrum on the text frequency is used (Wang and Pan (2017a, b); Sheng et al. 2009). Figure 3 shows the calculation process of text frequency cepstrum. The correspondence between the frequency and Mel frequency can be approximated by the following formula:

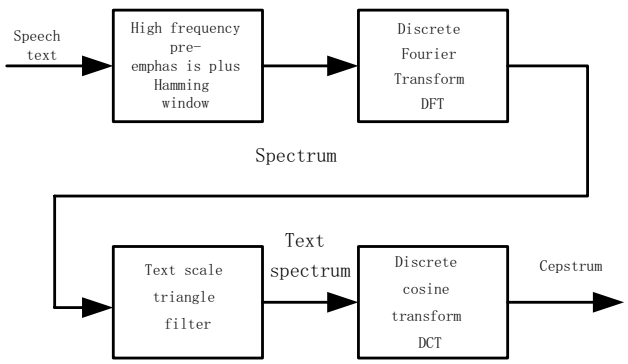

Fig. 3 Calculation process of text frequency cepstrum

$$\text{Mel}(f) = 2595lg\left(1 + \frac{f}{700}\right) \tag{6}$$

2.2.3 Construction of Gaussian mixture model based on voiceprint

The Gaussian Mixture Model (GMM) is derived from the classical speech model: Hidden Markov Model (HMM). The basic idea of GMM is to establish a statistical model for the speaker's speech characteristics. In GMM, the Gaussian mixture probability density is the sum of the probability densities of M Gaussian components, which can be expressed by the formula:

$$p(\vec{x}|\lambda) = \sum_{i=1}^{M} w_i b_i(\vec{x}) \tag{7}$$

where $\vec{x}$ is a D-dimensional random variable. $b_i(\vec{x})$ is the probability density function of the i th Gaussian component, w_i is the weight of the i th Gaussian component. The probability density of each Gaussian component is actually an D -dimensional Gaussian function, that is

$$b_i(\vec{x}) = \frac{1}{(2\pi)^{D/2}\left|\sum_i\right|^{1/2}} \exp\left\{-\frac{1}{2}(\vec{x} - \vec{\mu})^T \sum_i^{-1}(\vec{x} - \vec{\mu}_i)\right\} \tag{8}$$

where $\vec{\mu}$ and $\sum_i$ represent the mean and covariance matrix of the i-th Gaussian component. The sum of Gaussian component weights satisfies the normalization condition $\sum_{i=1}^{M} w_i = 1$. A Gaussian mixture probability density can be represented by three sets of parameters: the weight, mean and covariance matrix of each component, which can be recorded as a parametric model:

$$\lambda = \left\{w_i, \vec{\mu}_i, \sum_i\right\} i = 1, \ldots, M \tag{9}$$

Each voiceprint feature in the multimedia video is represented by the GMM model or by the parameter model λ .GMM model is shown in Fig. 4.

3 Independent video subtitle base ICA feature extraction method

3.1 Video subtitle independent component feature extraction

Independent component analysis is generated on the basis of blind source separation (BSS). Independent component analysis (ICA), principal component analysis (PCA) and singular value decomposition (SVD) together form a linear

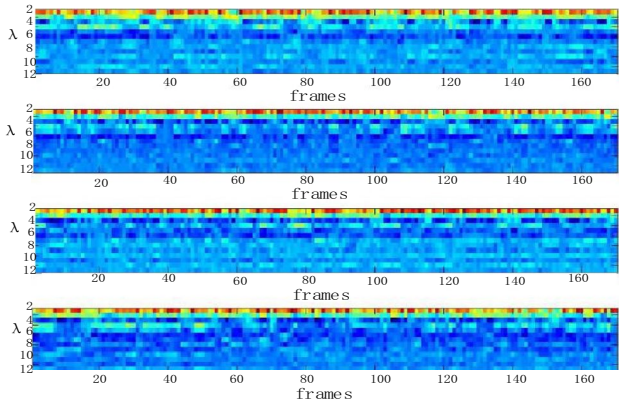

Fig. 4 GMM model

transformation technique. Principal component analysis and singular value decomposition can only cover the second order correlation of data based on energy attribute decomposition of data (Liu et al. 2017), and ICA can cover high-order correlation of data. When classifying resources such as video, image, sound, etc., many related features can be extracted, and these features are generally hidden in high-order related features. ICA ensures mutually independent higher-order features on the basis of reducing feature dimensions (Wang and Pan (2017a, b)). It replaces principal component analysis (PCA) and singular value decomposition (SVD), and it's widely used in video, image, sound and other resource identification and classification operations to reduce the identification bias and the classification bias. Suppose the number of multimedia video image block training samples is represented by L, and N represents the number of pixel points in a single training sample. And there are $N = n_1 \times n_2$ (row n_1 and column n_2 of an individual sample). y is the reconstruction coefficient of ICA, and X represents the training sample database, then:

$$y = WX \tag{10}$$

W is represented the decomposition square matrix, and the decomposition matrix is obtained by the ICA calculation. Bartlett uses the independent video subtitle ICA feature to characterize multimedia video (Lu et al. 2016), and the correct rate of ICA reconstruction coefficients is higher than PCV.

The ICA calculation can acquire mutually independent multimedia video image bases, and map the image bases in the image bases, and the generated coefficients are independent video caption base ICA features of the image.

3.2 Independent video subtitle base ICA features

When extracting the independent video caption base ICA feature, set $n_1 \times n_2$ to represent the row vector of the multimedia video image block. N represents its dimension, then $N = n_1 \times n_2$. L represents the number of samples, when $L \geq N$,it will generate $L \times N$ matrix x.

The process of extracting independent video subtitle base ICA features of multimedia video is as follows:

1. Refer to standard PCA calculation method, take C as X^T covariance matrix. Find the eigenvectors and eigenvalues of C, and arrange the feature data from large to small. Suppose q_i represents the eigenvector corresponding to the first m feature data ($i = 1, \ldots, m$. q_i is a $N \times 1$ column vector), q_m represents the matrix of row N and column M, then:

$$q_i = \left[q_1, q_2, \ldots, q_m\right] \tag{11}$$

2. In the previous step, q_m is the eigenvector containing the largest feature data m. In the previous step, q_m is the eigenvector containing the largest feature data of m, so it can be seen that the matrix x composed of the original training samples contains the most energy. Assuming that q_m^T is the transposed matrix of q_m, replace X in Eq. (10) with q_m^T. Calculated using the fast fixed-point ICA method:

$$y = Wq_m^T \Rightarrow q_m^T = W^{-1}y \tag{12}$$

Each line of y in Eq. (12) is an independent video subtitle base. $m \times m$ represents the feature matrix W, and the value of W can be obtained in training;

3. The eigenvector base coordinates represent individual training library samples, which can be expressed as $R_m = xq_m$. R_m means that, in the matrix $L \times m$, the first line represents the coordinates of the first sample pair of m eigenvector bases, the last line represents the coordinates of the m eigenvector bases of the L sample. Suppose x_{mse} is the minimum close value of x, than the least square error method can be used to obtain x_{mse}. Bring Eq. (11) into the formula to get:

$$x_{mse} = R_m q_m^T = xq_m q_m^T = xq_m W^{-1}y \tag{13}$$

4. Equation (13) shows that, Line i of $xq_m W^{-1}$ is the linear combination coefficient of the number i training sample relative to the independent video subtitle base in y. So assuming $I_{1\times N}$ represents any test sample, the multimedia video independent video caption base ICA feature can be expressed as:

$$a = Iq_m W^{-1} \tag{14}$$

Linear combination of mutually independent image bases to form independent video caption base ICA features. The essence of the image subtitle block is composed of some horizontal and vertical strokes or stroke segments. The independent subtitle base ICA feature is similar to the stroke segment that constitutes the image caption base, which is the main factor for the formation of multimedia subtitles (Ghosh et al. 2018). Therefore, the independent subtitle ICA feature is an important basis for classifying multimedia video text subtitle blocks (Tao et al. 2016).

4 Automatic detection and positioning of video subtitles

4.1 Automatic positioning of candidate text lines

Preprocessing video images is divided into two steps: First, the multimedia video image is segmented to form $N \times N$ sub-blocks, and each sub-block is named using a subtitle block (+1) or a non-subtitle block (−1); Then, using the independent video subtitle ICA feature extraction method, the m ICA image features of the independent multimedia video text sub-block are extracted.

To enhance viewing, text with rich strokes essentially enhances the contrast between text and background. Video subtitle position is usually set at the edge of the image. The distinction between lines of text is obtained by laterally projecting the edge density of the image to obtain the demand threshold. After the image pre-processing process divides the multimedia video image into different blocks, the block image is detected by the vertical edge of the Sobel operator. Set Sobel operator as $\begin{bmatrix} 1 & 0 & -1 \\ 2 & 0 & -2 \\ 1 & 0 & -1 \end{bmatrix}$, then a longitudinal edge detection map is obtained.

The horizontal pattern expansion of the edge detection density map can expand the edge density contrast of non-text areas and text areas. Setting structural elements as $SE = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}$, it will get an expanded picture.

The lateral distribution of edge density was obtained by transverse projection of the image. In order to ensure the smooth positioning of text lines in the image acquisition process, non-text interference needs to be covered. The traditional coverage method is usually achieved by setting a threshold value, but the threshold value is sometimes unable to adapt to the image and background changes, and this leads to low accuracy in positioning results. In this paper, a new adaptive iterative localization algorithm is proposed based on the characteristics of lateral distribution of edge density. The mean edge density is the iteration step size, and the height of the text line is no less than 10 pixels of the basic premise of non-text coverage. The edge density of the remaining text lines in this cycle is no less than 50%, or the number of candidate text lines obtained through multiple cycles has been stable, which are the condition of the loop iteration stopping. The process of candidate text line positioning by adaptive iterative positioning algorithm is as follows:

1. X represents the edge density sequence of the longitudinal edge detection map, and X_{mean} represents the mean value of the edge density sequence. Negative sequence values are generated, resulting in a new sequence denoted by X_{new}. That is $X_{new} = X - X_{mean}$;

2. In order to obtain the point of the candidate character line *TextLine*, the positioning of the adjacent rising 0 starting point and falling 0 starting point are monitored;

3. Determine the height of candidate text lines. If it is less than 10, the candidate text line does not meet the coverage requirement for exclusion, and the sequence value of the new sequence X_{new} interval is set to 0;

4. It is necessary to judge whether the new sequence X_{new} edge density filling rate is greater than 50%. If it is greater than 50%, the cycle condition *FlagA* can be set to 0, and if it is less than 50%, *FlagA* is set to 1;

5. Express the number of candidate lines for this acquisition as *TextLineNum*, and the previous time was expressed as *LastTextLineNum*. Determine whether the current time is the same as the previous one. If it is consistent, set *FlagB* to 0. If the inconsistency setting *FlagB* is 1;

6. Set the loop variable value as $X = X_{new}$, $LastTextLineNum = TextLineNum$;

7. When one of *FlagA* and *FlagB* equals to 0, the iteration stops, and if neither is 0, the iteration continues;

Above steps show that the coverage of non-text lines affects the iteration step length and accuracy during the positioning of candidate text lines. After completing the above process loop, text candidate lines of multimedia video text can be obtained.

4.2 Automatic positioning of candidate text blocks

The process of automatically locating candidate text blocks in the multimedia video text is to position the text regions of the candidate text lines from beginning to end. This process can review the flaws in the detection of the character candidate lines, thereby detecting the corresponding character candidate regions in the horizontally expanded picture as the main monitoring area (Liu et al. 2019a, b).

Edge density distribution map was obtained by longitudinal projection of monitoring area. When the vertical projection distribution is uniform, there is no need to use the iterative method (Wei et al. 2017). It is assumed that the mean value, maximum value and minimum value of edge density projection are respectively expressed as *Mean*,*Max* and min. The following threshold calculations can be used to obtain candidate text blocks:

$$Thresh = \min\left[Mean, (Max - Mean)/2\right] \tag{15}$$

The line width is regarded as the block height and the block width, and the blocks of the adjacent distance less than one word width are integrated to form a complete line. Constraint rules, such as fill rate constraints, text line minimum

Table 1 Different feature representation of video subtitles

Characteristic categories	Original grayscale feature	Key grayscale features	PCA feature	ICA feature
Feature representation	Original pixel	Selected pixel	PCA	ICA
Dimension	121	43	36	36

Table 2 Comparison of different characteristics of the test samples

Group	Subtitle characteristics	Background characteristics	Similar to the sample
Test sample 1	Less subtitle and concentrated	Less complicated and some interference	Similar
Test sample 2	More subtitles and concentrated	More complexity and interference	Similar
Test sample 3	More subtitles and no concentrated	Normal complexity and interference	No similar

and maximum constraints, etc., can be used to check the text candidate row detection to ensure that the obtained text block positioning results are more accurate and reflected in a constrained form (Wang et al. 2018).

5 Experimental analysis

1200 frames of multimedia video images were selected from multiple programs of a TV station for experiment. Among them, 300 frames were selected as training samples and the remaining 900 frames as test samples. Simulation experiments were conducted under the environment of Windows2000 and Matlab5.3.

During the experiment, the grayscale value of each large sub-block of 11 × 11 pixels is referred to as the original grayscale feature; the gray value of each 11 × 11 pixel size sub-block is called the original gray level feature during the experiment; the 41 gray values (4 × N-3) obtained by the "米" diagonal structure of each 11 × 11 pixel size sub-block are referred to as key gray values. When extracting ICA features of independent video subtitling basis of multimedia video text, PCA algorithm is firstly used to conduct mean removal and whitening processing of feature data. Therefore, PCA can be regarded as a component of ICA; then use the fast fixed-point ICA method to obtain the feature matrix W.

The comparison results of different video subtitle features and recognition performance are shown in Table 1:

Table 1 shows that the dimensions of the ICA feature and the PCA feature are lower than the other two features.

In the process of comparing the recognition effect of four characteristics, relevant settings are required if you want to obtain the recognition accuracy and error rate of different methods effectively: (1) Support vector algorithms are the same for all features. In order to ensure the accuracy of experimental results, support vector machines should be used for all features. (2) There is no need to use pyramid model in subtitle location and recognition of original image. (3) The result statistics are performed after the support vector output is completed, and no post processing is required.

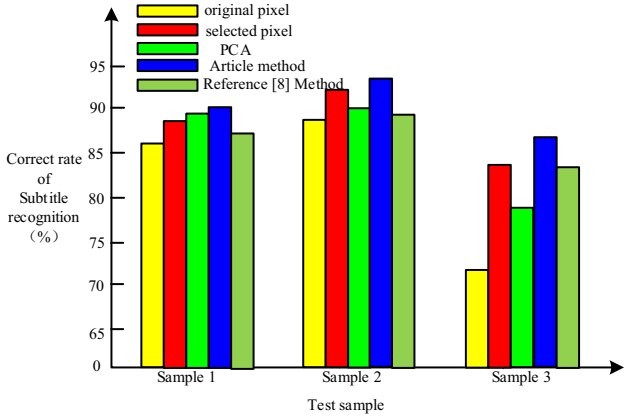

Fig. 5 Test the comparison of subtitle recognition accuracy of sample

The 900-frame test sample is divided into three types of test samples according to the difference in characteristics between background and subtitles. The 300-frame subtitle images that are very different from the training samples are randomly selected for applicability analysis of various features, such as the third type of test samples in Table 2.

For the three types of test samples in Table 2 and the four characteristics in Table 1, the accuracy and error rate of multimedia video subtitle location recognition are judged by the method of this paper. The results are shown in Figs. 5 and 6.

It can be seen from Figs. 5 to 6 that the ICA feature among the four features used in this paper has the highest accuracy rate for subtitle recognition, with the average accuracy rate higher than 88% and the lowest error rate for subtitle recognition, indicating that the ICA feature used in this paper can achieve effective identification of subtitle features.

In order to further verify the recognition accuracy of the proposed method, the method of this paper is compared with the Yin's method (2019) and Abdulrahim's method (Abdulrahim and Salam 2016), and the specific comparison is shown in Fig. 7.

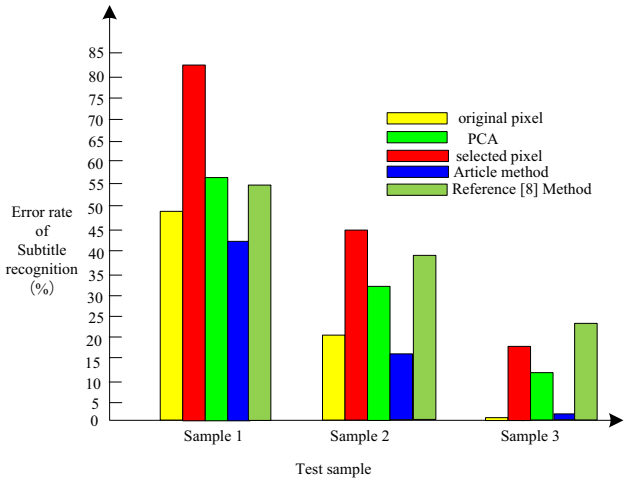

Fig. 6 Test sample subtitle recognition error rate comparison

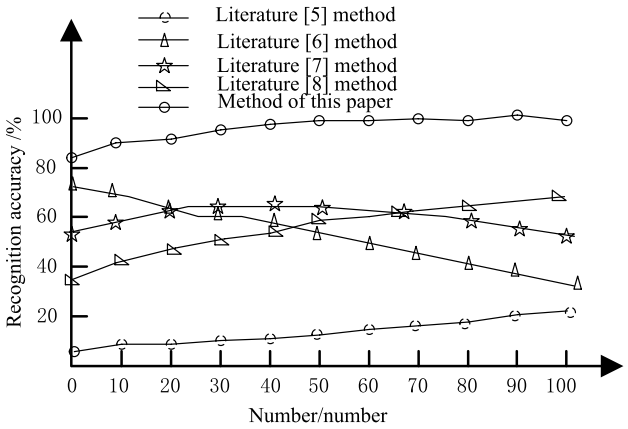

Fig. 7 Comparison of recognition accuracy of different methods

According to Fig. 7, the identification accuracy of Yin's method (Yin et al. 2019) is about 10%, that of Abdulrahim's method (Abdulrahim and Salam 2016) is about 50%, that of Yan's method (Yan et al. 2018) is about 48%, that of Liu's method (Liu et al. 2017) is about 57%. The identification accuracy of the method proposed in this paper is about 95%. This fully proves the superiority and high accuracy of this method.

120 different images including news and advertisements are selected from different TV programs for simulation experiment. The size of 120 images must be 288*352. Then Tables 3 and 4 are the detection results of subtitle lines and blocks of these video images by the method in this paper respectively:

$$row\ recall\ ratio = \frac{\mathrm{Line}A}{\mathrm{Total\ Line}} \times 100\% = 90.1\%$$

Table 3 Subtitle line detection results

Subtitle total line	Correct subtitle line (line A)	Error subtitle line (line B)	Undetected subtitle line (line C)
242	218	11	24

Table 4 Subtitle box detection results

Subtitle total box	Correct subtitle box (box A)	Error subtitle box (box B)	Undetected subtitle box (box C)
271	239	18	32

Table 5 Subtitle line detection results

Subtitle total line	Correct subtitle line (line A)	Error subtitle line (line B)	Undetected subtitle line (line C)
242	186	28	20

Table 6 Subtitle box detection results

Subtitle total box	Correct subtitle box (box A)	Error subtitle box (box B)	Undetected subtitle box (box C)
271	202	36	33

$$row\ precision = \frac{\mathrm{Line}A}{\mathrm{Line}A + \mathrm{Line}B} \times 100\% = 95.2\%$$

$$block\ recall\ ratio = \frac{\mathrm{Box}A}{\mathrm{Total}\mathrm{Box}} \times 100\% = 89.5\%$$

$$block\ precision = \frac{\mathrm{Box}A}{\mathrm{Box}A + \mathrm{Box}B} \times 100\% = 93.0\%$$

Then, Tables 5 and 6 respectively report the subtitle lines and blocks of the multimedia video text automatic positioning method based on multi-modal fusion for these images:

$$row\ recall\ ratio = \frac{\mathrm{Line}A}{\mathrm{Total}\mathrm{Line}} \times 100\% = 76.9\%$$

$$row\ precision = \frac{\mathrm{Line}A}{\mathrm{Line}A + \mathrm{Line}B} \times 100\% = 86.9\%$$

$$block\ recall\ ratio = \frac{\mathrm{Box}A}{\mathrm{Total}\mathrm{Box}} \times 100\% = 74.5\%$$

$$block\ precision = \frac{BoxA}{BoxA + BoxB} \times 100\% = 84.9\%$$

A comprehensive analysis of the analysis results of the above four tables, as well as the calculation of the recall rate and precision of subtitle lines and blocks for different methods, it can be seen that the recall and precision of the multimedia subtitle text are automatically located in this method higher than multimodal fusion method. This shows that the method in this paper works well for the line and block positioning of multimedia video text.

6 Conclusion

In order to realize the automatic positioning of multimedia video text and improve the detection precision and efficiency of multimedia video information, this paper proposes a multimedia video text automatic positioning method for subtitle frames. The outstanding advantages of this method are mainly in two aspects:

1. The basis of the method to realize the automatic positioning of multimedia video text is the multimedia independent video caption base ICA feature. It is a video subtitle independent component feature that can cover the high-order correlation of multimedia video data. It is a stroke segment that constitutes the image caption base, which is the main factor for the formation of multimedia subtitles, which can improve the automatic positioning accuracy of subsequent multimedia video texts;
2. In the process of multimedia video text localization, the method adopts adaptive iterative localization algorithm according to the characteristics of edge density horizontal distribution map. This method is simple and easy to understand, has strong adaptability to complex changes of video frames, and has faster and more accurate detection and positioning, and has the best effect on automatic positioning of multimedia video lines and blocks.

Although the method in this paper has made some achievements at the present stage, there are still some shortcomings. In the aspect of video subtitle independent component feature extraction, the method used for feature extraction has a lack of completeness of feature extraction; When the image candidate character lines are automatically positioned, the accuracy of selection and positioning of the candidate regions needs to be improved. In the future work, we will solve the shortcomings of the method.

References

Abdulrahim K, Salam RA (2016) Traffic surveillance: a review of vision based vehicle detection, recognition and tracking. Int J Appl Eng Res 11(1):713–726

Armstrong M (2017) Automatic recovery and verification of subtitles for large collections of video clips. SMPTE Motion Imaging J 126(8):1–7

Bakas S, Makris D, Hunter GJ, Fang C, Sidhu PS, Chatzimichail K (2017) Automatic identification of the optimal reference frame for segmentation and quantification of focal liver lesions in contrast-enhanced ultrasound. Ultrasound Med Biol 43(10):2438–2451

Gao Z, Lu G, Lyu C, Yan P (2018) Key-frame selection for automatic summarization of surveillance videos: a method of multiple change-point detection. Mach Vis Appl 29(7):1101–1117

Ghosh T, Fattah SA, Wahid KA (2018) CHOBS: color histogram of block statistics for automatic bleeding detection in wireless capsule endoscopy video. IEEE J Transl Eng Health Med 6:1–12

Głomb P, Romaszewski M, Sochan A, Opozda S (2011) Unsupervised parameter selection for gesture recognition with vector quantization and hidden markov models. In: IFIP conference on human-computer interaction. Springer, Berlin, pp 170–177

Korzun DG, Kashevnik AM, Balandin SI, Smirnov AV (2015) The smart-M3 platform: experience of smart space application development for Internet of Things. In: Internet of Things, smart spaces, and next generation networks and systems. Springer, Cham, pp 56–67

Kurzhals K, John M, Heimerl F, Kuznecov P, Weiskopf D (2016) Visual movie analytics. IEEE Trans Multimed 18(11):2149–2160

Liu S, Pan Z, Cheng X (2017) A novel fast fractal image compression method based on distance clustering in high dimensional sphere surface. Fractals 25(04):1740004

Liu S, Bai W, Zeng N, Wang S (2019a) A fast fractal based compression for MRI images. IEEE Access 7:62412–62420

Liu S, Liu G, Zhou H (2019b) A robust parallel object tracking method for illumination variations. Mob Netw Appl 24(1):5–17

Lu M, Liu S (2019) Nucleosome positioning based on generalized relative entropy. Soft Comput 23(19):9175–9188

Lu Y, Shahabi C, Kim SH (2016) Efficient indexing and retrieval of large-scale geo-tagged video databases. GeoInformatica 20(4):829–857

Lu G, Zhou Y, Li X, Yan P (2017) Unsupervised, efficient and scalable key-frame selection for automatic summarization of surveillance videos. Multimed Tools Appl 76(5):6309–6331

Ren J, Zhang C, Zhang L, Wang N, Feng Y (2018) Automatic measurement of traffic state parameters based on computer vision for intelligent transportation surveillance. Int J Pattern Recognit Artif Intell 32(04):1855003

Sangaiah AK, Medhane DV, Han T, Hossain MS, Muhammad G (2019) Enforcing position-based confidentiality with machine learning paradigm through mobile edge computing in real-time industrial informatics. IEEE Trans Industr Inform 15(7):4189–4196. https://doi.org/10.1109/TII.2019.2898174

Sheng H, Li C, Wen Q, Xiong Z (2009) Real-time anti-interference location of vehicle license plates using high-definition video. IEEE Intell Transp Syst Mag 1(4):17–23

Shum HY, Liao M (2001) Advances in multimedia information processing-Pcm 2001: second IEEE pacific rim conference on multimedia, Beijing, China, October 24–26, 2001, proceedings, vol 2. Springer Science & Business Media

Tao J, Franke U, Klette R (2016) Context-based multi-target tracking with occlusion handling. Mach Vis Appl 27(8):1339–1349

Wang W, Pan M (2017) An keyword extraction approach from video associated text based on multiple features. J Zhejiang Univ Technol 1:4

Wang JH, Liu TW, Luo X, Wang L (2018) An LSTM approach to short text sentiment classification with word embeddings. In: Proceedings of the 30th conference on computational linguistics and speech processing (ROCLING 2018), pp 214–223

Wei Y, Zhang Z, Shen W, Zeng D, Fang M, Zhou S (2017) Text detection in scene images based on exhaustive segmentation. Signal Process Image Commun 50:1–8

Xu L, Kamat VR, Menassa CC (2018) Automatic extraction of 1D barcodes from video scans for drone-assisted inventory management in warehousing applications. Int J Logist Res Appl 21(3):243–258

Yan C, Xie H, Chen J, Zha Z, Hao X, Zhang Y, Dai Q (2018) A fast Uyghur text detector for complex background images. IEEE Trans Multimed 20(12):3389–3398

Yin F, Wu R, Yu X, Sun G (2019) Video text localization based on Adaboost. Multimed Tools Appl 78(5):5345–5354