Research on Video Information extraction algorithm based on Deep Learning

Qichen Li

International School, Beijing University of Posts and Telecommunications, Beijing, 100000

Keywords: video abstract; hierarchical clustering; multi-feature similarity; video segmentation; key frame

Abstract: The emergence of a large number of video data shows a greater demand for video abstract, and the existing feature-based and shot-based video abstract extraction methods are difficult to meet the actual needs in terms of computation, accuracy and reliability. using multi-feature layering of video to segment shots, using the strategy of first coarsening and then fine, using simple feature segmentation and then clustering complex features. Accurate video clips and key frames are obtained, and then global features are extracted from each key frame, and the similarity is compared to generate the final video abstract. The video summary is generated adaptively without considering the weight of multiple features. The experimental results on public video data sets such as VSUMM show that the multi-feature layering method effectively improves the performance of video abstract extraction. The accuracy and recall rate are better than the traditional methods, and the computational complexity is obviously reduced.

1. Introduction

With the development of mobile phone, network and other video-related equipment and technology, video data increases rapidly, and a large number of video data also have a greater demand for video excerpt extraction [1]. Video abstract is a single frame or a group of a small number of video frames extracted from the original video data that can reflect the main information content of the original video, which can be used in video preview, video classification, video recognition and other applications [2].

The existing video summary generation algorithms are divided into two categories: feature-based extraction and video shot-based segmentation. in order to make effective use of the multi-feature information of multi-frame images, redundant frames can not be removed in advance. Each frame needs to calculate multiple eigenvalues, and then cluster multiple times, which requires a large amount of computation [3]. And it is difficult to set each feature weight effectively. The video abstract extraction method based on shot segmentation uses local features to divide the video into different video clips according to the content to obtain the structure information of the whole video. However, due to the time and space limitations of local features, when there are complex video content changes with multiple scenes and shots, it is easy to lose important frames or generate redundant frames. The video abstract generated by extracting video clips can not effectively represent the main content of the original video, and its performance is difficult to meet the actual needs [4].

In this paper, a video abstract extraction algorithm based on multi-feature layering is designed, which adopts the strategy of fast first and then slow. In each layer, video shots are segmented and segments are clustered with different features. Finally, accurate video shot segmentation and key frames with time sequence information are obtained. Then the similarity of local feature clustering of key frames is compared to get candidate video abstracts. According to the needs of practical application. Combined with the global features, the final video abstract is obtained from the candidate video abstract [5].

  DOI: 10.25236/cseem.2020.175

2. Video shot Segmentation based on Multi-feature

Before video abstract extraction, the original video is generally preprocessed to remove empty frames and error frames. Shot segmentation phase begins with feature selection from three directions: pixel segmentation feature, color feature and key point matching. using multiple features, we can better complete video shot segmentation according to the change of video content. Ensure the structural integrity of video shot segmentation [6].

The initial video RGB image obtained by the code can not directly represent the color information such as color light and shade, hue and freshness, so it is transformed into HSV. As most videos are shot by handheld devices, a variety of slight jitter will lead to a change in shooting angle and lead to a change in color brightness. In fact, the video content has not changed much. In this way, it is easy to segment the frames with the same content incorrectly. In the HSV color space, the chromaticity H and saturation S components are closely related to the color information accepted by the human eye, while the luminance V component is not directly related to the color information of the image image, so it has little effect on video segmentation, so this paper only quantizes the H and S components at 16 levels and synthesizes them into one-dimensional feature vectors according to the image formula (1).

$$\begin{matrix} h_0 \\ h_1 \\ \vdots \\ h_{14} \\ h_{15} \end{matrix} + \begin{matrix} s_0 \\ s_1 \\ \vdots \\ s_{14} \\ s_{15} \end{matrix} \rightarrow [h_0 \quad \cdots \quad h_{15}, \quad s_0 \quad \cdots \quad s_{15}]. \tag{1}$$

In terms of similarity discrimination, this paper uses Euclidean distance to calculate the similarity between two images. The smaller the Euclidean distance of Sim (foot /) and J is, the more similar it is. If the normalized Euclidean distance of the one-dimensional eigenvector calculated by the formula (1) of the two images is 0, their Euclidean distance can be calculated by formula (2).

$$\mathrm{Sim}(R, I) = (\sum_{i=0}^{L-1}(h_i^R - h_i^I)^2)^{\frac{1}{2}} \tag{2}$$

The segmentation of the video clustered by the Euclidean distance of HSV features is still affected by the interference of color features. often, some color features change between frames, but the video content does not change greatly, so the ORB algorithm is further used to determine clustering by the ratio of the number of ORB feature points matched between two images to the average number of feature points of the two images. The video clips are clustered again by similarity matching[7].

3. Generation of Video Abstract

3.1 Key frame selection algorithm based on Local Features

The next step is to extract key frames from the segmented video clips as candidate video abstracts. Firstly, when matching the third ORB feature points in video segmentation clustering, the boundary frames of each video clip with ORB eigenvalues have been calculated to form a set 0. Where y + is the number of video clips after video segmentation. In this way, the number of images in set 0 is twice the number of video clips finally segmented, that is, 2 video clips.

For each video clip, according to the eigenvector X of each image in the clip. To calculate the distance between frames between adjacent images is to combine the two categories with the closest distance between frames into one category, and calculate the clustering centers of each category:

$$C_{\text{num}} = \frac{1}{l_{\text{num}}}\sum_{i=1}^{2h_j} \quad x_i, i = 1,2,3,\cdots, l_{\text{num}} \tag{3}$$

3.2 Generate final video summary based on global features

The candidate video abstracts are obtained through section 2.1, which are all video abstracts based on video shot segmentation to maximize the complete reflection of the original video content. when the candidate video abstract is output directly, a better recall rate will be obtained, but the accuracy rate is relatively low, so it can be applied to applications with high recall rate, which is called method one in this paper.

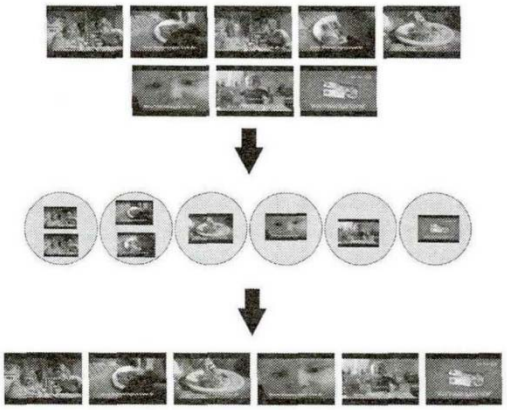

Figure 1. Re-cluster the candidate video abstracts according to the global characteristics

Since there are redundant video summaries for similar video clips in method 1, in order to further obtain video abstracts that not only retain timing information but also accurately describe video content, Euclidean distance is used to extract global features of candidate video abstracts for clustering, and similar candidate video abstracts are re-clustered into one class. Then the most representative frame in the same class is selected as a static video summary, as shown in figure 1. This method is called method 2 in this paper, which will achieve higher accuracy.

The specific algorithm is as follows:

Calculate the set of candidate video summary keyframes $E = \{e_1, e_2, \cdots, e_K\}$, the two candidate video abstracts e_i, e_j, the Euclidean distance of the video Sim, the Euclidean distance of the video, the Euclidean distance $\mathrm{Sim}(e_i, e_j) = \left(\sum_{k=0}^{L-1} \left(e_{ik} - e_{jk}\right)^2\right)^{\frac{1}{2}}$.

Set the threshold ϕ if $\mathrm{sim}(e_i, e_j) < \phi$, group it into one category

The selection of key frames after the class: according to the number of key frames in each class, the current frame is output directly when nail2, the frame in front of the time sequence is output as the summary when n is greater than or equal to 3, the key frame closest to the average value in each class is output as the summary.

4. Generation of Video Abstract

Table 1 is a comparison of the performance of video summaries generated on VSUMM datasets and WY-316 data sets using two different feature selection classifiers. Table 1 uses feature extraction methods with different channels on VSUMM datasets and WY-316 datasets except that the number of image components is different. The other parameters "data set" H, S, V "H,:" 3 "are the same.

On the VSUMM dataset, the precision and recall rate of the H, S two-component method of the WY-316 dataset are improved compared with the H, S, V triple "VSUMM", "WY-316", "0.82" 0.61 "0.84", 0.82 "0.86" 0.82, "0.62", "0.90" component methods, and on the WY-316 data set, the accuracy and recall rate of the two-component method are higher than those of the H, S, V three-component method, and the recall rate of the two-component method is improved. The recall rate is improved more obviously. in terms of time consumption, the two-component method reduces the time consumption by 7.7% compared with the three-component method. thus it can be seen that the improved method of HSV color feature is better than the traditional method. The experimental results on the WY-316 data set show that this method is more effective on the new network video

data. Compared with H, S and V components, H, S and V components can avoid the influence of jitter interference caused by human factors and show better robustness by using H and S color histograms.

Table 1. Comparison of feature extraction methods using different channels on VSUMM datasets and WY-316 datasets

Data set	Evaluation index			
	H、S、V		H、S、	
	Precision rate	Recall rate	Precision rate	Recall rate
VSUMM	0.82	0.61	0.84	0.62
WY-316	0.82	0.86	0.82	0.90

According to the method 1 and method 2 proposed in section 2.2 of this paper, compared with the existing traditional video abstract algorithms, the experimental data extracted from the VSUMM data set is shown in Table 2, and the data is the average of all the experiments in the data set.

Table 2. Experimental results of this algorithm and other traditional algorithms on VSUMM data sets

Method	Precision item rate	Recall rate	F-score
VSUMM	0.64	0.72	0.67
VISCOM	0.65	0.70	0.67
VRHDPS	0.77	0.61	0.68
CSEA	0.66	0.87	0.72
Proposed1	0.75	0.89	0.81
Proposed20	0.84	0.62	0.71

As can be seen from Table 2, the accuracy of method 2 is much higher than that of VSUMM algorithm, VISCOM algorithm, incremental VRHDPS algorithm and literature algorithm on VSUMM data sets. Method 2 in this paper, after proposing the candidate video abstract, considers the global features again to reduce the candidate video summary, which can maximize the video summary to describe the video content accurately. Method 1 is superior to other algorithms in recall rate, and the accuracy of precision is better than VSUMM algorithm, VISCOM algorithm and literature algorithm. Method 1 selects video abstracts from cut video shots, which can retain the time sequence information to the maximum extent and express the original video content completely. the F-score of method 1 is much higher than that of other algorithms. The F-score of the two algorithms is slightly lower than that of the literature, but higher than that of other algorithms. The overall performance of the algorithm proposed in this paper is better.

5. Conclusion

In this paper, a video abstract extraction algorithm based on multi-feature layering is designed. firstly, the complexity of the original video data is calculated according to the video features, respectively, using pixel feature differences, color feature differences and feature point matching differences for multi-feature video shot segmentation, and then according to the sudden change of video frequency content, the number of clusters is determined adaptively, and the candidate video abstracts are obtained.

In the final video generation stage, global features are introduced, structural information is retained, and two different methods are adopted to meet different application requirements with emphasis on accuracy and recall rates. experimental results on the public dataset show that because of its hierarchical structure and the utilization of multi-feature information, the proposed algorithm not only speeds up the processing speed, but also improves the quality of video summary.

Each layer is processed separately to avoid the problem of multi-feature weight assignment, and the improved HSV color feature extraction algorithm is faster and better than the traditional HSV

color feature extraction algorithm. Video frames are clustered in the time sequence order through multi-feature similarity, and the timing information is retained.

Combined with the global feature extraction method, it makes up for the deficiency of the local feature method. How to combine the time series information and video segments to generate dynamic video abstract, and how to introduce a more effective feature similarity calculation method to improve the quality of video summary and speed up the extraction speed will be the focus of further research.

References

[1] Xiaowei Gu, Lu Lu, Shaojian Qiu, Quanyi Zou, Zhanyu Yang. Sentiment key frame extraction in user-generated micro-videos via low-rank and sparse representation[J]. Neurocomputing,2020,410.

[2] Huimin Yang, Qiuhong Tian, Qiaoli Zhuang, Linye Li, Qinglong Liang. Fast and robust key frame extraction method for gesture video based on high-level feature representation[J]. Signal, Image and Video Processing,2020.

[3] Mobile Communications; Researchers from Kamaraj College of Engineering & Technology Report Details of New Studies and Findings in the Area of Mobile Communications (Eagle Eye Cbvr Based On Unique Key Frame Extraction and Deep Belief Neural Network) [J]. News of Science,2020.

[4] Engineering; Study Data from National Institute of Advanced Industrial Science and Technology (NIAIST) Update Knowledge of Engineering (Multi-sensor Integration for Key-frame Extraction From First-person Videos) [J]. Journal of Engineering,2020.

[5] T. Prathiba, R. Shantha Selva Kumari. Eagle Eye CBVR Based on Unique Key Frame Extraction and Deep Belief Neural Network[J]. Wireless Personal Communications,2020(prepublish).

[6] Hoang Nguyen Ngoc, Lee Guee Sang, Kim Soo Hyung, Yang Hyung Jeong. Effective Hand Gesture Recognition by Key Frame Selection and 3D Neural Network[J]. Smart Media Journal,2020,9(1).

[7] Wenbin Xie, Zhen Zhang, Yuefei Wang, Yuanyuan Zhang, Liucun Zhu. The adaptive key frames extraction method based on representativeness and independence features[J]. International Journal of Information and Communication Technology,2020,16(4).