Key frame extraction for Human Action Videos in dynamic spatio-temporal slice clustering

Mingjun Sima
College of Computer Science, Sichuan University
Department, University, Chengdu, Sichuan, 610065, China
*Corresponding author’s e-mail: 2015141462192@stu.scu.edu.cn

Abstract. Detecting representative frames in videos based on human actions is quite challenging because of the combined factors of human pose in action and the background. This paper proposed a key frame extraction algorithm based on dynamic spatio-temporal slice clustering. This algorithm firstly uses the dynamic spatio-temporal slice position selection method based on the human mask heat map to calculate the position of slice to realize the dynamic selection of slice positions, then complete the extraction of spatio-temporal slice images. After clustering the spatio-temporal slice images, this method extracts key frames according to the clustering results. The experimental results prove the validity of spatio-temporal slice location selection method, the proposed algorithm can effectively solve the problems of information redundancy and key information missing in existing methods. We conduct experiments on a challenging human action dataset UCF101 and show that our method can detect key frames with high accuracy.

1. Introduction
In the modern era, video has become an important way for people to spread information on the Internet. In recent years, the number of videos on the Internet has increased significantly. How to quickly and effectively find the required video clips in a large number of videos has become a hot topic. In particular, a large proportion of these videos depict events about people such as their activities and behavior.

Video is composed of continuously changing image frames, and people call the frames that can effectively represent the main content of the video in this continuous image frame as key frames. Through the extraction of video key frames, it can effectively reduce redundant information and facilitate subsequent processing. Key frame extraction technology plays an important role in the field of video processing.

This paper is organized into five sections. The following section discusses some of the related important earlier work. The third section presents the proposed method. This is followed by Sect. 4, comprising the outcome of our experimentation and comparison with associated results of other state-of-the-art methods. In Sect. 5, a general conclusion and some future perspectives are proposed.

2. Related work
The common key frame extraction technology mainly includes:(1) Key frame extraction based on video motion information[10]; (2) Key frame extraction based on video content information[11]; (3) Key frame extraction based on video shots information[8]; (4) Key frame extraction based on clustering[9].

The first method extracts and analyzes the motion information in the video to achieve the key frame extraction. Wlof W[2] uses optical flow analysis to measure the motion in a shot and select key frames
at the local minima of motion and proposed a hierarchical key frame selection methodology in which first classify shots into categories and then apply key frame selection algorithms appropriate to each category.

The second method uses the texture, color, shape and other underlying features of each frame in the video frame sequence as the basis for key frame extraction. The size of feature difference between adjacent frames can effectively reflect the change of video content, so the key frame can be extracted according to this feature. Li et al. [3] present a content-based multi-keyframe abstraction to make an effort through exploring the temporal and semantic feature correlations among videos with overlapping views, and selecting the most representative frames for condensed abstraction using weighted correlation map.

In the third method, video is divided into independent shots, and the image frames with fixed positions in each lens are selected as key frames.

Last method makes full use of the relevance of information between video shots and within video shots, clusters videos according to the distance information between video frames, and selects one frame from each cluster as the key frame. Common clustering methods include k-means algorithm, hierarchical clustering algorithm [4], etc. This kind of method is widely used because of its good extraction performance. So in this paper, the clustering algorithm is improved. Liu et al. [1] proposed a Pyramid Motion Feature (PMF) for each frame of an action sequence, and use the AdaBoost learning algorithm to select key frames for each action sequence. Yan et al. [7] proposed a novel deep two-stream ConvNets based key frame detection model, which combines the advantages of CNN model and LDA. They also proposed a novel automatically generating label method to train the deep key frame detection model. Gharbi et al. [12] proposed a method called key frame extraction by graph clustering (KEGC) and use local description computed around interest points. They use local frames description to have significant and robust features representing the different objects in the frame.

3. Proposed Approach

3.1. Spatio-temporal slice
Spatio-temporal slicing is a video analysis method, which extracts a certain pixel from each frame of the video sequence and synthesizes a two-dimensional image containing spatio-temporal information to express the motion information in the video [13]. The implementation details of the algorithm are as follows:

A row (column) of pixels is extracted at a fixed position of each frame, and then a pair of two-dimensional images with ordinate (abscissa) as pixels and ordinate (abscissa) as time are synthesized in time order. As a result, the whole video is condensed with one image. The synthesized image is equal to slicing a video sequence, so it can be called spatio-temporal slice image, as shown in Figure 1.

Figure 1 Schematic diagram of spatio-temporal section position selection.
Spatio-temporal slicing only processes part of the pixel information in the video, so the computational complexity of this method is lower than other video analysis methods. At the same time, it does not need to consider the contour information of the moving object when analyzing the related information of the moving object in the video through the spatio-temporal slice, so this method has
strong anti-interference ability. Even if the target is partially or completely occluded by other objects in a short time, it will not have much impact on the results.

3.2. Dynamic spatiotemporal slice location selection method

Spatio-temporal slice image is the concentration of video information, and its result is directly related to slice location. If the selection is not correct, even if the best clustering algorithm is used, the accurate key frame cannot be obtained. The common horizontal or vertical slicing is to select the horizontal or vertical extension line of the video center as the temporal and spatial slicing location, which may not be able to achieve the most effective extraction of information in the video.

In order to solve this problem, this paper extracts the human mask from the video image and calculates the corresponding human mask heat map, so as to achieve a more accurate expression of the human motion region, and select the row and column position with the highest intensity as the Spatio-temporal slice location. The algorithm details of human mask heat map is as follows:

First, Yolact++ is used to extract the human mask. Yolact++ [14] can detect objects and segment instances at the same time. Yolact++ is improved on the basis of Yolact. Yolact has two parallel network branches. In the first branch, a set of "prototype masks" with the same size as the input image are generated by using full convolution network; The second branch is to predict the "mask coefficient" of each target detection frame on the basis of target detection. Finally, the prototype mask and mask coefficient are combined linearly to complete the task of instance segmentation.

Second, only the pixel values of the original image region corresponding to the human mask region in each frame are retained to obtain the human contour image, and the image is saved.

Then, the human body contour is binarized. Considering that there may be a small amount of edge information in the human body contour, the binarization threshold is set to 5 to get the binarized human body contour. A frame of the body contour map and Binary body contour map in test video “v_Shotput_g19_c03” is shown in Figure 2.

![Figure 2](image_url)

Figure 2  A frame of the body contour map and Binary body contour map in test video “v_Shotput_g19_c03”.

Finally, the binary human contour image is accumulated frame by frame, and the values of each pixel are normalized to between 0 and 1. In this paper, the image is named human mask heat map (HMMH). The formula is as follows:

\[ I_{\text{mask}}(x,y) = \frac{\sum I_i(x,y)}{\max \sum I_i(x,y)} \]  \hspace{1cm} (1)

\( I_{\text{mask}}(x,y) \) is the final human mask heat map. A human mask heat map of test video “v_Shotput_g19_c03” is shown in Figure 3. We can see that areas with higher motion intensity have brighter colors.
The human mask heat map can well reflect the area and intensity of human movement in the whole video, whether in static or dynamic background. Therefore, the location of dynamic spatio-temporal slice obtained by this method can make the extracted image contain more abundant information.

3.3. Clustering and key frame determination

After obtaining the slice location, the spatio-temporal slice image corresponding to the video can be obtained. It is necessary to cluster the image according to the time axis direction, so as to obtain the clustering results divided according to the time axis. After merging some clustering results, the image frames corresponding to the clustering centers of each category will constitute key frames.

In clustering the image, this paper chooses the K-means clustering method, but considering the continuity of the time dimension of the spatio-temporal slice image, the following modifications are made:

1) The updating method of cluster center: When each category has a new sample added, the cluster center of this category will change. In K-means algorithm, the mean value of each class sample is taken as the new cluster center. However, since the spatio-temporal slice image is clustered along the time direction, the clustering center of each class should be a certain frame. The new clustering center obtained by K-means algorithm may not correspond to the frame sample. In order to solve this problem, when changing the clustering center of each class of samples, it is necessary to calculate the average value of all samples of each class, The nearest sample from each cluster is selected as the new cluster center.

2) New distance calculation method: the distance between frames in video data is not only related to the sample data, but also related to the time gap between samples. Therefore, these two factors should be taken into account in the calculation of inter frame distance. Finally, the calculation formula of the distance between samples is as follows:

$$\text{dis}(i,j) = \frac{\| \text{slice}(i) - \text{slice}(j) \|}{\sum_{f} d(f)}$$

The smaller the distance between two frames, the more similar the two images are.

4. Experimental analysis

4.1. Dataset

Since there is no benchmark dataset for key frame detection in human action videos, we conduct our key frame detection experiments on a challenging dataset for video based human action recognition, namely UCF101[15]. This dataset consists of 101 actions classed with 13,320 video clips. It has at least 100 video clips for each action category. Besides its 101 categories, UCF101 has coarse definitions which divide the videos into human-object interaction, human-human interaction, body movement, musical instrument playing and sports. In these five kinds of videos, three videos are selected for each class, and a total of 15 videos are used as experimental data.
4.2. Evaluation of Experiments

At present, there are many commonly used evaluation criteria in the field of video key frame extraction, precision ratio and recall ratio are widely used in the field of key frame extraction. This paper selects these two indicators to evaluate the performance of this method.

(1) recall ratio

Recall ratio can be used to measure the degree of missing key frames. The calculation formula is as follows:

\[ R = \frac{N_c}{N_c + N_m} \times 100\% \]

(3)

where \( N_c \) is the number of correct keyframes to be extracted, and \( N_m \) is the number of keyframes missed.

(2) precision ratio

Precision ratio can be used to measure the accuracy of key frame extraction. The calculation formula is as follows:

\[ P = \frac{N_f}{N_c + N_f} \times 100\% \]

(4)

where \( N_f \) is the number of misdetected keyframes.

4.3. Experiments Analysis

The recall ratio and precision ratio of different algorithms are shown in Table 1. It can be seen from the table that recall ratio and precision ratio of our method is higher than the most of the existing classical algorithms, indicating that the proposed method is effective. The main reason of such result is that Spatio-temporal slicing only processes part of the pixel information in the video and HMHM can achieve the most effective extraction of information in the video.

| Methods                           | Recall ratio | Precision ratio |
|-----------------------------------|--------------|-----------------|
| Histogram[6]                      | 62.92\%      | 59.07\%         |
| Traditioinal spatio-temporal slice[13]| 68.55\%    | 65.31\%         |
| Weighted multi-view[16]           | 87.28\%      | 81.53\%         |
| Deep Key Frame Extraction[17]     | 94.54\%      | 92.17\%         |
| **Our method**                    | **95.66\%**  | **92.68\%**     |

5. Conclusions

In this paper, we propose a key frame extraction method based on dynamic spatio-temporal slice clustering. At the same time, this paper proposes a dynamic spatiot-temporal slice location method based on the human mask heat map (HMHM). Through the dynamic spatio-temporal slice location selection method, we get the spatio-temporal slice location which is in line with more abundant information. After getting the spatio-temporal slice according to this location, the improved clustering algorithm is used to get the required key frame. By taking advantage of adaptive spatiot-temporal slice location, the effect of key frame extraction has a great improvement, which point was verified by a serious experimental results.

It is worth mentioning, some misdetection are observed in our method. Note that the proposed method did not remove noise to improve the performance. In the future, a noise removing algorithm can be utilized to address the misdetection issues of the proposed approach and thus for further improvement of the overall detection accuracy.

In our future work, we aim to build a deep model using more effective unsupervised clustering algorithm. Besides, the proposed algorithm is aimed at human action videos. We hope to carry out relevant research on all videos in future research.
References

[1] Liu L, Shao L, Rockett P. Boosted key-frame selection and correlated pyramidal motion-feature representation for human action recognition[J]. Pattern Recognition, 2013, 46(7):1810-1818.

[2] Wolfs W. Key frame selection by motion analysis[C]// Acoustics, Speech, & Signal Processing, on Conference, IEEE International Conference. IEEE, 1996.

[3] Li P, Guo Y, Sun H. Multi-keyframe abstraction from videos[C]// IEEE International Conference on Image Processing. IEEE, 2011.

[4] Elaheh, Rashidi, Abdolreza, et al. An information theoretic approach to hierarchical clustering combination[J]. Neurocomputing, 2015.

[5] Soomro K, Zamir A R, Shah M. UCF101: a dataset of 101 human actions classes from videos in the wild[J]. arXiv:1212.0402, 2012.

[6] Hu W, Xie N, Li L, et al. A Survey on Visual Content-Based Video Indexing and Retrieval[J]. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 2011, 41(6):797-819.

[7] Yan X, Gilani S Z, Qin H, et al. Deep keyframe detection in human action videos[J]. arXiv preprint arXiv:1804.10021, 2018.

[8] Y Y Zhu and D ZHOU, "An approach of key frame extraction based on video clustering [J]", Computer Engineering, vol. 30, no. 4, pp. 12-14, 2004.

[9] A Amiri and M. Fathy, "Hierarchical Keyframe-based Video Summarization Using QR-Decomposition and Modified-Means Clustering[J]", EURASIP Journal on Advances in Signal Processing, vol. 2010, no. 1, pp. 892124, 2010.

[10] N Ejaz, SW Baik, H Majeed et al., "Multi-scale contrast and relative motion-based key frame extraction[J]", EURASIP Journal on Image and Video Processing, vol. 2018, no. 1, pp. 40-51, 2018.

[11] KW Sze, KM Lam and G. Qiu, "A new key frame representation for video segment retrieval[J]", IEEE Transactions on Circuits & Systems for Video Technology, vol. 15, no. 9, pp. 1148-1155, 2005.

[12] Gharbi, H., Bahroun, S. & Zagrouba, E. Key frame extraction for video summarization using local description and repeatability graph clustering. SIViP 13, 507–515 (2019).

[13] Chen L H, Su C W. Video Caption Extraction Using Spatio-Temporal Slices[J]. International Journal of Image and Graphics, 2018, 18(02):1850009.

[14] Bolya D, Zhou C, Xiao F, et al. YOLACT++: Better Real-time Instance Segmentation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, PP(99):1-1.

[15] K. Soomro, A. R. Zamir, and M. Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:1212.0402, 2012.

[16] Ioannidis A, Chasanis V, Likas A. Weighted multi-view key-frame extraction[J]. Pattern Recognition Letters, 2016, 72(mar.1):52-61.

[17] Meng J, Shuai Z, Lifang W, et al. Deep Key Frame Extraction for Sport Training[J]. Neurocomputing, 2019, 328(FEB.7):147-156.