Multi-frame dimensionality-reduction difference method for extracting key frames of video

Shuaipeng Cai
Automation School
Beijing University of Posts and Telecommunications
Beijing, China
17839938364@163.com

Qinyan Zhang
Automation School
Beijing University of Posts and Telecommunications
Beijing, China
zh_qinyan@163.com

Qing Wang
Research Institute of Information Technology
Tsinghua University
Beijing, China
qing.wang@tsinghua.edu.cn

Yi Lei
Beijing Dfusion Co.,Ltd.
Beijing, China
leiyi9345@163.com

Jijiang Yang
Research Institute of Information Technology
Tsinghua University
Beijing, China
yangjijiang@tsinghua.edu.cn

Corresponding author: yangjijiang@tsinghua.edu.cn

***Abstract*—Key frame extraction is very important for video data processing. This paper mainly studies how to extract the key frames of the video data when transforming the medical action video data into image data, so as to process and mine the medical action video data with image processing technology in the later period. Because of the difference of the doctor for some special action concerns, the existing algorithm of key frames can not extract the important frames selected by the doctors exactly, so this paper will review the traditional method to extract key frames, and use a new method of key frame extraction (multi-frame dimensionality-reduction difference method) to extract the key actions that clinical doctors pay attention to. Compared with the traditional method, this method can extract the frame image concerned by the doctor better.**

Keywords—Key frame extraction, medical action video data, multi-frame dimensionality-reduction, image processing

I. INTRODUCTION

With the increase of video data production, the mining and application of video data becomes the direction of some researchers. Image classification technology tends to be mature, so in the current video classification methods, the mainstream is based on the image method, the core idea is to convert video data into image data, and then use the method of deep learning to classify. This involves the extraction of video key frames. In the traditional methods, the extraction of video key frames is mostly based on ordinary mathematical methods: average value method and specific frame method. The average method is divided into frame average method and histogram average method. There are also extraction using the clustering method and the curve method.

This paper mainly studies the inter-frame pixel information: a new inter-frame difference method, multi-frame dimensionality-reduction difference method, is proposed. In the process of extracting key frames, the number of frames of dimensionality-reduction difference can be dynamically adjusted according to the number of frames of video data. The frame number of dimensionality-reduction difference is also referred to as a unit in this article. The purpose of dividing the unit is to split the action into different stages. Each stage can extract key frames, and the final key frame contains the whole process of an action. The advantages of this method are as follows: firstly, redundant frames can be eliminated as much as possible; secondly, key frames in video data can be better selected.

The method of multi-frame dimensionality-reduction difference method used in this paper is inspired by the two-frame difference method. The inter-frame difference method itself is used for moving target detection. Here, the gradient of the image is found by setting the threshold value, and the key frames are reserved. The core idea of this method is to preliminarily divide all the frame, making frames dimension is reduced, so the advantage of a section of the video data often contains a different action, or an action of multiple components, standing long jump movement, for example, to select the number of frames in dimensionality reduction can be specified, and thus makes the movement divided into several stages, each stage can extract key frames, finally get the whole complete action of key frames. Dimensionality reduction is in addition to the preliminary division, in the process of extraction, for every frames in each stage the difference between two frames, remove part of the frame by threshold value method, makes the number of frames at the stage reduced to reach dimensionality reduction, repeat the above process, until the key frames are selected. When differencing between two frames, the absolute value of the corresponding position of pixel gray value difference is computed, and the difference is compared with the threshold. The process is shown in the figure 1, the first step is that all the frames are divided, the frames can be set by yourself, and then use interframe difference method to each divided stage respectively, remove part of the frames, through the analysis of the threshold. In the subsequent steps, not dividing all the frames as the first step, repeat interframe difference method in the first step, until the key frames are selected in accordance with set threshold.

978-1-7281-7303-0/20/$31.00 ©2020 IEEE
DOI 10.1109/COMPSAC48688.2020.00-48

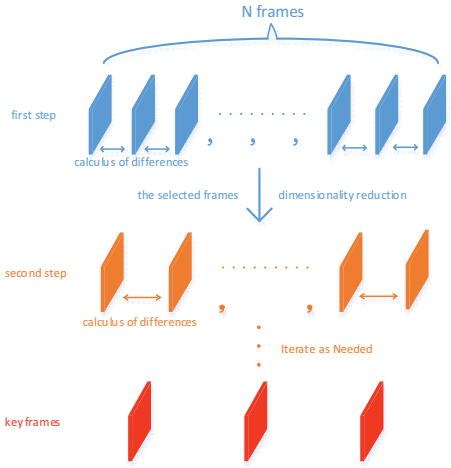


Fig. 1. The frame difference method to dimensionality-reduction of overall process.

The main purpose of using multi-frame dimensionality-reduction difference method to extract the key frames is to use deep learning methods to study children's developmental coordination disorders later, and to evaluate school-age children through image video classification technology. The method in this paper draws the following conclusions through experiments: When the specified unit frame number is twice the desired number of frames, the extracted frames can be closer to the frames selected by clinicians, and the redundancy is smaller. The most ideal result obtained in the experiment of this paper is that the accuracy is 83.3%, and the redundancy is 30% (redundant frames are different from the number of frames extracted by the doctor).

The following is the framework of this paper: The second part introduces the research related to key frame extraction. The third part mainly describes the method of key frame extraction in detail. The fourth part gives the results of some traditional methods and the method in this paper, and compares them with the frame images selected by the doctor. Finally, the conclusion and future work are discussed.

II. RELATED WORKS

The history of the studies of video key frame extraction has long, its problems is mainly because the development of the computer makes the processing of image data becomes easier and easier, and at present for image data processing method is more mature. Converting the video data to image data for further mining and the application is a reliable method, video is composed of many frame image, the implicit information among a few adjacent frames is often close, in the video classification study, if extracting all the frames to do processing, the content close to the frame to extract the feature information is similar and it is not much helpful with further classification, Therefore extracting the key frame of video data becomes an important task. In addition, compared with image data, video data occupies more storage space and requires more computation. There are many solutions to the key frame extraction problem, and there are three kinds of methods: basic way, clustering, curve method, etc.

Basic way: The basic methods are divided into average value method and fixed interval frame method [1-3].Frame average value method is to calculate the average value of all pixels at a specific pixel position in the image of all frames of video data, and then select the frame image that is closest to the average value at that position. Histogram averaging method is to calculate the average value of statistical histogram for all video frame images, and select the frame image closest to the average value in the histogram of all frame images of video data as the key frame. However, the frame average method has an obvious disadvantage, because it can only extract one frame as the key frame. For a complete action, it is easy to miss other key actions. Specific frame method, also known as fixed interval frame method, extracts frames with fixed interval as key frames, but its disadvantage is that the extracted frames may miss key actions.

Clustering approach: The frames with similar attitude in the video data are grouped and the problem is transformed into a clustering problem [4]. This method was tested by Liu et al. (2003) [5]. Moir et al. (2013) improved the method and proposed a clustering algorithm based on adaptive threshold [6]. The method was to add adaptive threshold algorithm to the original IOSDATA algorithm [7] to extract key frames.

Curve method: This method is similar to the boundary detection algorithm, which extracts the content information boundary of the front and rear frames, and compares the curve difference and threshold between frames to extract the key frame information. The curve simplification method was first proposed by Lowe [8], and was later applied in practice by Lim and Thalmann [9], who regarded the motion sequence as the trajectory curve in the high-dimensional feature space and extracted key frames according to the motion attitude changes. This method was constantly improved in the later researches.

Through the analysis of the above methods, it can be seen that no matter what method, there is a common place, to find the difference between the frame information to extract the key frame, to find the appropriate threshold to do some operations before and after the frame, which is similar to the method of inter-frame difference. In contrast to complex algorithms, using simple logic to innovate can sometimes yield good results. The multi-frame dimensionality-reduction difference method is an innovation of the inter-frame difference method, which is used to extract the important movements that the clinician pays more attention to. In the assessment of children's developmental coordination disorder, clinicians often pay more attention to the coordination of children's movements. If the video data is taken by the camera and the key frame extraction algorithm is used to obtain the important frame images in the video, it can help the doctor to evaluate the development status of the child.

III. KEY FRAME EXTRACTION ALGORITHM

For the assessment of children's developmental coordination disorder, we aim to use the method of deep learning to collect the video data of the motion used in the assessment through the medical camera, and then classify the video through the image method, which involves the extraction of key frames. Before the key frame extraction, we use the CMU laboratory testing human body skeleton algorithm [10] to deal with video data, the purpose is to make the original data points of the human body

be marked out, when the clinical doctors select key frames the bending angle of each joint can be seen clearly, which allows doctors to find its some frames of the concern. The key frame image selected by the doctor will be given as the control group in the next section.

Next, we will recall the traditional key frame extraction algorithm. Average method is a relatively traditional method, which is introduced by histogram average method. It is mainly used to calculate the histogram of each frame of the image, calculate the average of the histogram of all frames of the image, and find out which frame of the image is closest to the average of the histogram.

The calculation process of this method is as follows:

The average value of the histogram is:

$$A = \frac{1}{N\sum_{n=0}^{N-1}\sum_{x=0}^{k-1} H[f(i,j,n),x]} \quad (1)$$

Among them, the difference between the histogram of an image and the average value of the histogram is as follows:

$$D(f_n, A) = \sum_{x=0}^{k-1} H[f(i,j,n),x] - A \quad (2)$$

We take: $$Z = \min\{D(f_n, A)\} \quad (3)$$

In the above formula, $H[f(i,j,n),x]$ is the histogram of frame n, k is the grayscale of the image, n is the total number of n frames of the video data, and Z is the smallest difference between the average value of the histogram and the average value of the histogram of a frame among all the frames, which is regarded as the key frame. In this process, we can obtain an image from the video data as the extracted key frame, its disadvantage is that the extraction of a few key frames, cannot fully highlight a complete action.

There are many methods to improve the key frame extraction algorithm based on clustering and curve. We will not go into details here. The following will be given the operation process of inter-frame difference method is given, and the gray image is used for calculation. Then, the difference between two adjacent frames of $f_t(i,j)$ and $f_{t+1}(i,j)$ is as follows:

$$D = \sqrt{\sum_{i=1}^{h}\sum_{j=1}^{w} | f_t(i,j) - f_{t+1}(i,j) |^2} \quad (4)$$

In the above formula, $f_t(i,j)$ represents the grayscale value of the image at frame t and the difference value between two adjacent frames. h and w represent the width and height of the image respectively. If the difference between two adjacent frames is greater than the threshold value, the motion state in the image is considered to have changed, and the image of the subsequent frame is considered to be the key frame.

IV. Experiments

In this section, we will use the commonly used key frame extraction algorithm and the method in this paper to process the key frame images selected by clinicians as the basis (as the control group), and conduct a comprehensive analysis with the results of the control group. As shown in figure 2, the doctor selects the key frame of standing long jump:

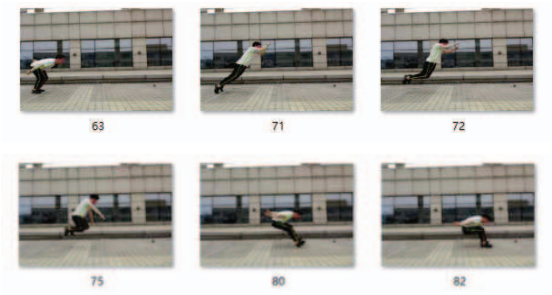


Fig. 2. The key frame selected by the clinician.

As shown in Figure 2, we preliminarily processed the video data and displayed the human skeleton using the openpose algorithm, so as to facilitate the doctor to observe the joint angles formed during the movement and select the frames considered important.

When extracting key frames based on average value, we use histogram mean value method, that is, we select the image which is closest to the average value of histogram among all frames. Next, we provide the key frame image obtained by histogram mean method:

Fig. 3. The key frame image extracted by the histogram average method.

As shown in Figure 3, the key frame selected by the average method is frame 60.The video data used in this paper is the standing long jump motion specified by the doctor, which was collected by us using the medical camera. The data duration is 6 seconds, and there are 134 images in total.

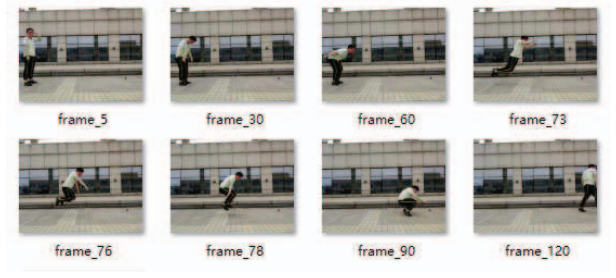
Multi-frame dimensionality-reduction difference method is to convert all frames to grayscale, several continuous frames are selected as a unit, make all the frames divided into several units, these units may contain different action, or the same action of several stages. The division makes the number of the frames in each unit reduced, to achieve the purpose of dimensionality reduction. In each unit to all adjacent frame the finite difference method is used to calculate, 'and' operation is used between two adjacent results, and through the threshold comparison, to eliminate part of the frames, again to achieve the purpose of dimensionality reduction, By repeating the process, the key frames of different stages can be extracted.

Finally, the key frame of the entire video data is obtained. Because the number of the frames can be specified when the units are divided, we experimented with several different number of frames, and compared, Each action can be split into different stages, when the specified unit frames are too close, the key frames extracted are also closer. So the interval of the experimental frames between two units should not be too small. We experimented with three frame interval and the unit frame number is roughly one to three times that specified by the doctor. Followings are the key frames extracted from each unit frame in the experiment:

(a) every 6 frames

(b) every 9 frames

(c) every 13 frames



(d) every 16 frames

Fig. 4. The key frames extracted from each unit

We used the key frames extracted by the doctor as the control, and analyzed the accuracy and redundancy of the frames in different units. Because the actions of the adjacent frames in a certain stage is often closer, so in terms of accuracy, with doctors to extract the difference before and after the number of frames as accurate values within a frame (If two frames appear to be close to a certain frame extracted by the doctor, it is regarded as one frame, but it is removed as the exact value when calculating redundancy). All six frames are accurate and are deemed to be 100% accuracy; In terms of redundancy, the ratio between the number of extra frames and the number of extracted frames is considered as the redundancy after removing the number of frames with the exact value.

The calculation formula of accuracy and redundancy is:

$$Fa = \frac{n}{m}, \quad Fr = \frac{N-n}{N}$$

In the above formula, Fa is the accuracy, n is the accurate value, m is the number of frames extracted by the clinician; Fr is the redundancy, and N is the number of extracted frames.

The analysis results are shown in the following table:

TABLE I. ACCURACY—REDUNDANT ANALYSIS

Specifies the Number of Cell Frames	6	9	13	16
Fa (%)	50	33.3	83.3	33.3
Fr (%)	90	78.6	30	75

The experimental results show that after selecting the appropriate frame for unit division, it can basically extract the complete flow of the entire action, it is closer to the frames selected by the doctor, not only can the whole process of standing long jump be well retained, but also as many redundant frames are eliminated as possible, which reduces a large amount of computation for subsequent processing. According to the analysis results of table I, it is appropriate to select the specified frame number about twice as many as the desired frame number, and the extracted frame can be closer to the frame selected by the doctor, with a small degree of redundancy.

V. CONCLUSION

The multi-frame dimensionality-reduction difference method realizes the use of computer to automatically extract the key frames of medical action video data, which is a step forward for the use of deep learning method to evaluate children's developmental coordination because the 13 movements used in the evaluation were similar to the standing long jump. Compared to other methods, it can better obtain the key frame selected by the clinician. However, the disadvantage is that the computer cannot find a complete motion by itself with several different stages in the experiment, so it cannot select the unit frame number by itself. What the algorithm wants to improve in the future is the selection of threshold. By adding adaptive threshold analysis, we can find the inflection points in different stages of the action, and let the computer determine the selection of frame number of an action unit by itself, so as to make the algorithm better adaptive. As far as the results are concerned, this method has been able to meet the need of this

project to automatically extract key frames of medical action video data by computer.

REFERENCES

[1] Zhong Qu, Lidan Lin, Tengfei Gao, et al. An Improved Key-frame Extraction Method Based on HSV Colour Space [J]. journal of software, 2013, 8(7).

[2] Zhang J. Key frames extraction for video [J]. 2016.

[3] Huamin F, Wei F, Sen L, et al. Framework for shot boundary detection and key-frame extraction [J]. Journal of Tsinghua University, 2005, 8(2):121-126.

[4] Zhao H, Wang T, Zeng X. A Clustering Algorithm for Key Frame Extraction Based on Density Peak [J]. Journal of Computer & Communications, 2018, 06(12):118-128.

[5] Liu F, Zhuang YT, Wu F, Pan YH. 3D Motion Retrieval with Motion Index Tree. *Comput Vis Image Und*, 2003; 92(2-3): 265-284.

[6] Moir G L, Graham B W, Davis S E, et al. An Efficient Method of Key-Frame Extraction Based on a Cluster Algorithm [J]. Journal of Human Kinetics, 2013, 39(1):5-13.

[7] Wu L, Li X, Yong S. A New Method for Bad Data Identification of Integrated Power System in Warship Based on Fuzzy ISODATA Clustering Analysis. *Electr Eng*, 2011; 97: 101-108.

[8] Lowe D G. Three-dimensional object recognition from single two dimensional images. *Artifl Intell*, 1987; 31(3): 355-395.

[9] Lim I S, Thalmann D. *Key-Posture Extraction out of Human Motion Data by Curve Simplification*. Istanbul:23rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 1167-1169; 2001.

[10] Cao Z, Hidalgo G, Simon T, et al. OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields [J]. 2018.