J. Vis. Commun. Image R. 65 (2019) 102678

Contents lists available at ScienceDirect

J. Vis. Commun. Image R.

journal homepage: www.elsevier.com/locate/jvci

Quality-guided key frames selection from video stream based on object detection

Mingju Chen [a,b], Xiaofeng Han [c,*], Hua Zhang [a], Guojun Lin [b], M.M. Kamruzzaman [d]

[a] *Robot Technology Used for Special Environment Key Laboratory of Sichuan Province, Southwest University of Science and Technology, Mianyang 621010, China*
[b] *Artificial Intelligence Key Laboratory of Sichuan Province, Sichuan University of Science & Engineering, Zigong 643000, China*
[c] *College of Mathematics and Systems Science, Shandong University of Science and Technology, Qingdao, China*
[d] *Department of Computer and Information Science, Jouf University, Sakaka, Al Jouf, Saudi Arabia*

ARTICLE INFO

Article history:
Received 9 August 2019
Revised 7 October 2019
Accepted 7 October 2019
Available online 11 October 2019

Keywords:
Convolutional neural network
Key frame
Object detection
SIFT characteristics
Quality assessment model

ABSTRACT

Object detection technique is widely applied in modern intelligent systems, such as pedestrian tracking, video surveillance. Key frames selection aims to select more informative frames and reduce amount of redundant information frames. Traditional methods leveraged SIFT feature, which have high key frame selection error rate. In this paper, we propose a novel key frames selection method based on object detection and image quality. Specifically, we first leverage object detector to detect object, such as pedestrian, vehicles. Then, each training frame will be assigned with a quality score, where frames contain objects have high quality score. Afterwards, we leverage CNN based AlexNet architecture for deep feature representation extraction. Our algorithm combines mutual information entropy and SURF image local features to extract key frames. Comprehensive experiments verify the feasibility of practicing the key frame extractor based on convolutional neural network by training the model, and conduct a quality assessment model study.

1. Introduction

In recent years, video data has been widely used in many fields, especially with the rapid development of video surveillance technology and platform, video surveillance has become the main technical means of emergency command [1]. However, due to the large overlap between adjacent frames of video images and the large amount of data redundancy, using all video data as processing objects will significantly increase the workload of data processing, thereby reducing the processing efficiency of data [2–4,40–43]. Therefore, it is an important technical means to reduce the processing amount of data and improve the calculation efficiency by using a partial frame (referred to as a key frame) satisfying a certain degree of overlap requirement in a video image. At present, most of the key frame extraction methods for video images are based on the retrieval of video content [5–8], that is, the feature (cluster) space of the current key frame is first established, and then the feature distance between the next frame and the current key frame is calculated, and By setting the feature distance threshold to filter and extract key frames, the calculation efficiency is high. However, if the video image is used for mapping, the overlapping degree requirements must be met between adjacent key frames, and the overlap degree between the key frame images obtained by the above algorithm is large, which cannot meet the requirements of mapping accuracy. In response to the above problems, many scholars have proposed new key frame extraction algorithms, such as Chasanis et al. [9], Wattanarachothai [10], Hannane et al. [11], Thepade, etc. [12] using the same name point matching method, through frame-by-frame calculation and screening Key frames accurately obtain key frames that meet the overlap requirements. However, these algorithms use frame-by-frame calculations, which have low processing efficiency and are difficult to meet the application fields with high timeliness requirements. Jing et al. proposed a key frame extraction algorithm for layered adaptive frame sampling, which performs secondary screening on key frames and improves the efficiency of splicing [13]. Momin et al. [14] predicted the trajectory of four corners of the image by Kalman filtering, and realized the overlap between adjacent frames. However, due to the instability of flight attitude, this method is difficult to predict accurately by mathematical model. Luo et al. [15] proposed a key frame selection algorithm based on POS, but it is only suitable for video images with relatively stable flight attitude. According to the prior knowledge of the camera's specification

[*] Corresponding author.
E-mail addresses: chenmingju@suse.edu.cn (M. Chen), hanxiaofeng@sdust.edu.cn (X. Han), zzhh839@163.com (H. Zhang), tiemujian123400@163.com (G. Lin), mmkamruzzaman@ju.edu.sa (M.M. Kamruzzaman).

https://doi.org/10.1016/j.jvcir.2019.102678
1047-3203/© 2019 Elsevier Inc. All rights reserved.

parameters, flight platform speed, ground relative elevation, etc., Liu calculates the theoretical overlap between adjacent frames and filters the key frames according to the fixed frame interval, but the method is more undulating in terrain. Large areas are difficult to guarantee the requirements for overlap. Thepade [16] proposed a method to simplify the overlap of SIFT calculations and improve the efficiency of key frame extraction. Shi et al. [17] proposed a fast extraction algorithm based on geographic location of image space. The key frame extraction algorithm extracts key frames by calculating the degree of overlap between the current key frame and the previous frame, that is, setting the threshold of the initial key frame and the overlap degree. If the current frame meets the overlap threshold requirement, it is regarded as Is a key frame; otherwise, recursively proceeds to the next frame to continue the calculation. Although the above algorithms can extract key frames more accurately, it is necessary to cyclically calculate the heading overlap of the previous key frame image and each subsequent frame image until the next key frame meeting the overlap requirement. Since the degree of overlap cannot be directly measured, it can only be calculated by the same-named point, and extracting the same-named point and matching requires a lot of calculation work, thereby reducing the efficiency of data processing.

Therefore, how to extract key frame information in railway video images quickly and effectively is the focus of this paper. Considering that the repetition rate of adjacent frames in the video is generally high, the extraction of key frames can reduce the number of frames, thereby improving the image feature point detection and matching efficiency, and also providing an organizational framework for image stitching. In response to this key technology, it has received extensive attention from researchers and has achieved certain research results. Literature [18] proposed a method based on video content from the degree of change of color or texture information between adjacent frames. In [19], by calculating the distance between the current frame and the eigenvalue between the centroids, all the frames in the video are clustered and analyzed, and the analysis method based on video clustering is obtained. In [20], an algorithm based on motion feature analysis is proposed. The basic principle is to use optical flow analysis to minimize the amount of motion in the video as a key frame. The above three types of traditional algorithms mainly select key frames based on changes in the overall information of the image, which easily cause problems such as key frame selection errors, large computational complexity, and poor real-time performance [21–24].

Therefore, this paper proposes a key frame fast extraction algorithm based on the correlation coefficient of overlapping regions. The algorithm calculates the correlation coefficient based on the high correlation of adjacent key frame overlap regions, and uses the polynomial fitting method to fit the correlation coefficient of video images. The trend of change enables accurate positioning of key frame frames to quickly and accurately extract key frames. In this paper, the deep learning-based target detection method is used to classify the key frames in the video by establishing a convolutional neural network model, which makes the target detection based on deep learning possible in the application of key frame extraction [25].

2. Proposed method

2.1. Overview of convolutional neural networks

As a key technology for deep learning in the field of computer vision, convolutional neural network is an artificial neural network that simulates the cerebral cortex by designing a bionic structure, which can realize the training of multi-layer network structure. Compared with traditional image processing algorithms, convolutional neural networks can use local receptive fields to obtain self-learning ability to cope with large-scale image processing data, while weight sharing and pooling function design reduce the dimensionality of image feature points [26]. The complexity of parameter adjustment is reduced, the sparse connection improves the stability of the network structure, and finally produces high-level semantic features for classification, so it is widely used in the field of target detection and image classification [27]. Using the autonomous learning ability of the convolutional neural network model, target detection can be achieved. This section mainly uses a typical convolutional neural network structure, as shown in Fig. 1 [28].

Convolutional neural network is a multi-layer deep network structure, which is mainly composed of input layer, convolution layer, pooling layer, fully connected layer and output layer. The input layer is an image that needs to be processed, which the computer can understand as a number of matrices [29]. The convolutional layer is an important part of the convolutional neural network. The convolution operation of the matrix between the input layer and the convolutional layer extracts the features of the input image. The pooling layer is also an important part of the convolutional neural network [30]. It is usually placed after the convolutional layer. The function is to average or maximize the target area pixel of the input image, that is, downsampling processing and reducing the feature. Avoid image over-fitting while maintaining image resolution. The fully connected layer is located between the last pooled layer and the output layer, and is composed of 0 or more, each of which is connected with all the neurons

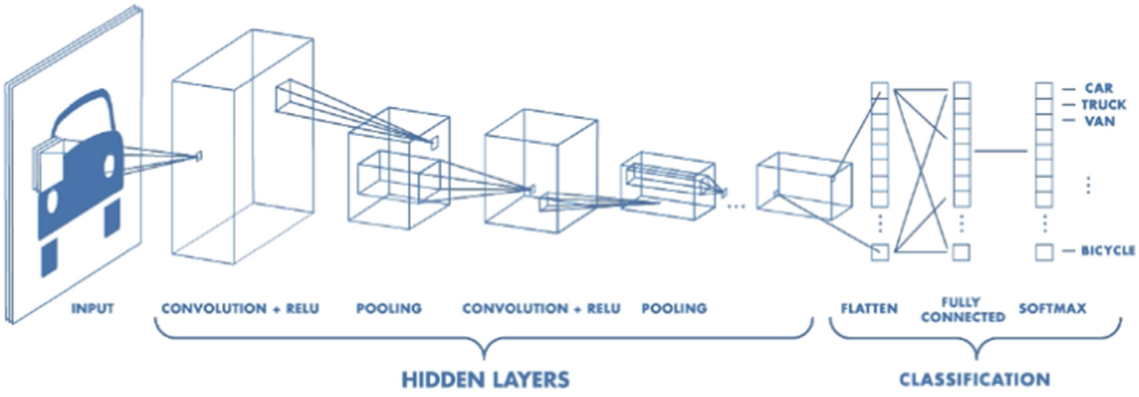

Fig. 1. The basic structure of the convolutional neural network.

of the previous layer, and is targeted according to the needs of the target detection. Map the feature vector to the output layer for easy classification. The output layer is located at the end of the neural network and is set to Softmax regression. The main function is to classify the input vectors mapped by the fully connected layer [31], and finally output the one-dimensional prediction vector, and the dimension is equal to the number of classifications. The combination of the convolutional layer + the pooling layer in the convolutional neural network can be repeated multiple times according to the actual task needs. Through the feature extraction of multi-layer neural network, the feature expression ability of different depths is obtained from spatial features to deep semantic features [32]. Finally, the results of target detection are obtained through the fully connected layer and the output layer. Therefore, according to the function of each layer, the convolutional neural network can be divided into two parts – a feature extractor composed of an input layer, a convolutional layer, and a pooled layer, and a classifier composed of a fully connected layer and an output layer [33]. After nearly two decades of rapid development, many convolutional neural network model structures have emerged, from the early LeNet model for handwritten digit recognition to the recent deep residual learning ResNet model. The number and depth of convolutional neural networks are constantly the accuracy of increasing image recognition is also increasing. In 1998, LeCun et al. designed the earliest representative convolutional neural network LeNet, which consists of two convolutional layers, two downsampling layers (pooling layers), and three fully connected layers. In 2012, the AlexNet model proposed by Krizhevsky et al. greatly improved the accuracy of image classification in the ILSVRC competition. The model consists of 5 convolutional layers, 3 pooling layers and 3 fully connected layers, through long time and big data. The training (about 60 million training parameters) shows the great potential of convolutional neural networks in the field of image classification. The VGG-Net model was designed by Simonyan et al. [34]. in 2014. There are six different network structures, all of which have a size of 3 × 3, which reflects the "simple, deep" characteristics of Szegedy et al. The proposed GoogleNet model, unlike the Alex-Net and VGG-Net models, increases the accuracy by increasing the number of layers in the network structure. Instead, it introduces the inception module to change the size of the receptive field and extract more features. In 2016, He et al. proposed the ResNet model, which used residual learning to alleviate the problem of degraded accuracy after the number of layers of the network structure increased to a certain extent, and achieved the best image recognition effect [35].

2.2. Research on key frame extraction method based on target detection

Based on the lens boundary method: This extraction method is mainly because there should be a small change between adjacent image frames in the same lens. Therefore, the boundary change of the entire lens image should also be small, and once the boundary appears in the lens. Changes, you can analyze the variables involved in these shots, so the first frame and the last frame of the lens boundary change can be selected as key frames. The advantage of this method is that it can be easily found in the video. The change scene, but the disadvantage is that when the lens shakes or moves, the selected picture will have errors, and for some longer events, it is difficult to find only the first and last two frames, if the lens split is impossible. Information screening. Image-based feature method: This method mainly uses the feature change of each frame in the image to obtain the key frame. First, the first data frame in the image is set as the key frame, and then the current data frame and the previous one. The key frames are compared and the difference between the two is obtained. If the difference exceeds a predetermined threshold, the data frame is a key frame. Under the premise of this data frame change, the key frame method can more flexibly judge the key frame and improve the flexibility of the operation, but he is not sensitive to motion, and there is no way to effectively select the change of information, so there are also Certain defects. Based on motion analysis method: This motion-based analysis method mainly uses camera motion to identify image information [36]. The motion of the camera includes two kinds of zoom motion and rotary motion. When the analysis is based on zoom motion, select the first two frames of zoom. The picture is used as a key frame, and based on the motion of the rotation, it is necessary to use a picture that overlaps the previous frame by less than 30° as a key frame. On the basis of the exercise method, when the research related motion is carried out, the amount of data that needs to be calculated is large, and the time consumed is also large, and more importantly, the local minimum value in the WOL method. It is also relatively vague. Cluster-based approach: With the development of clustering science and technology, it is widely used in data information processing. For some established data samples, we have not clearly defined the data samples from the beginning, so we hope that we can use some methods to make the sample more reasonable, so that we can put together similar elements. Therefore, it is more distinguished from the more different elements. This is the main method of the clustering. From the initial state of the clustering method, any element in the sample is reasonably allocated to a specific cluster, thereby satisfying the customer's needs and system requirements. However, this calculation method is not simple. It is still difficult to achieve, and the number of calculations in this implementation process is quite large [37]. After conducting a comprehensive study on the lens data, it is more desirable to have a video of the lens. The main factors are related to the following clustering: First, in the related video information, any video clip usually includes many shots. If these lenses are distinguished one by one, it is not only time-consuming and laborious, but also The analysis results are also more complicated. More importantly, only one lens is processed, which makes it difficult to save the time characteristics and motion specificity of the video clip, and the video will not appear continuous enough. Second, the video is length-compressed by clustering. From the current point of view, the key frames have been successfully used to represent specific shots, and these videos can be effectively processed by related technologies. However, such data access is still relatively large [38]. For example, a video clip is divided into 600–1500 related shots, which means key frames with the same data, for example, 3000–750,000 key frames in a video, then if we enlarge the subject to the entire video Monitoring the library, the amount of the entire data will be an order of magnitude, if a single processing of this data will be a more complex project, so the clustering process can achieve the simplification of these data, that is, will be different The key frames are differentiated and classified one by one, which not only reduces the scope of the related search, but also effectively reduces the time taken for the index. Extracting key frames based on MPEG compressed stream: In the current process of processing many video information, it is basically saved by using a compressed file format of MPEG. If you want to extract, you must first decompress the compressed file first, then the extraction of key frames is very large, and it will take up a lot of storage space. Therefore, key frame extraction based on MPEG format has become a key technology for video information analysis. It was first proposed in compressed video streams. Keyframe extraction is performed by scholars such as Janko who use macroblock detection in compressed streams to extract key frames from compressed video streams. In addition, there is a simpler method for extracting compressed video keyframes. The compressed video is divided into several video segments according to the equal time period, and then the first frame images of the videos are respectively presented, and the comparison and change of the videos can be effectively found by comparing the images of the first frames of the segments, thereby Extract to keyframes. However, the compressed video stream extraction key frame technology also has certain defects, which are prone to errors.

The traditional key frame extraction method usually achieves good results on the railway surveillance video by means of the change of the overall information of the image, but there are still some problems. First of all, it is necessary to manually design the first and last frames of a selected train, which affects the real-time and efficiency of key frame extraction. Secondly, without deepening the characteristics of the image, the selected target key frame error rate is higher and the generalization ability is poor. The successful application of convolutional neural networks to target detection can be seen as identifying whether the image is a target or a non-target. Therefore, this chapter designs a key frame extractor based on convolutional neural network. Firstly, a key point detection algorithm based on SIFT feature is used to extract a small number of regions as candidates on railway video images, and the depth features of candidate regions are extracted by Alex-Net model. Finally, the features are input into the SVM classifier for classification, and the results of the train head, tail and body are obtained, and the feasibility of the key frame extractor based on convolutional neural network is verified [39]. The key frame extractor flow chart proposed in this paper is shown in Fig. 2.

Key frame coarse recognition based on SIFT features: SIFT full scale scale invariant feature transform is an algorithm for detecting and describing local features in images. It was proposed by DavidLowe in 1999. The principle is to find extreme points in different scale spaces (features) Point), calculate position, scale, direction and other information, and describe the feature points with a 128-dimensional feature vector. Since the SIFT feature can better describe the local characteristics of the target and keep the transformations such as translation and rotation unchanged, it can be applied to the key frame feature region detection of the video image. The SIFT feature point algorithm mainly includes four parts: firstly, based on the establishment of the scale space, the Gaussian difference pyramid is used to find the local key points; then the scale space is fitted to obtain the exact value of the key point position and scale; further use the gradient The direction histogram assigns direction information to the key points; finally, a description vector of the feature points is generated. Through the rough recognition of key frame feature regions based on SIFT, a large number of regions with large differences from target detection are rejected. Only a few feature points similar to target detection enter the convolutional neural network as candidate regions, which reduces the work for the next fine recognition. The amount. Key frame fine recognition based on convolutional neural network: In the rough recognition stage of the key frame extractor, a large number of non-target areas have been screened, leaving a few candidate areas. The image feature points of the candidate regions have similarities, and the depth mining image features are required to be distinguished. Therefore, in the fine identification stage of the key frame extractor, the Alex-Net convolutional neural network model is used to extract the features of the deep convolution network from the candidate target regions, generate feature vectors, and use the SVM classifier of the key frame extractor to extract the feature vectors. Classification to get the final extraction results. In this paper, the Alex-Net convolutional neural network model is used, which consists of 5 convolutional layers, 3 pooling layers and 3 fully connected layers, of which 60 M parameters and 650 K neurons can be classified into 1000 categories. The SIFT feature vector of the candidate target region enters the model as an input image, and undergoes a five-layer convolution operation and a corresponding maximum pooling process. At the same time, the nonlinear activation function ReLU is used to accelerate the convergence speed, and the GPU parallel architecture realizes efficient convolution calculation.

2.3. Image feature extraction

After the key frame is extracted, the key frame shots are identified and the feature quantity is extracted, so that the corresponding search rules can be established. The image feature quantity extraction of the key frames involves the number of lens color features in the image, and the lens shape. And the number of features of the texture, all of which together constitute the spatial characteristics of the lens data, so this feature will also be applied to the relevant basis of video data clustering search. Color feature extraction: the color feature of the video image is the most basic feature of the video information. The color information processing process for the key frame image is: firstly, different color spaces are set for different video scenes, and then for any component. The value is determined, so that the characteristics of the scene color can be transformed into mathematical features. Finally, under the premise of mathematical science, the similarity between different image vectors is specified, and the similarity of different colors is judged by mathematical expression. Its main structure includes RGB structure, HSV structure and HIS structure. RGB is actually similar to our vision. However, the disadvantage is that it cannot effectively separate the brightness of the color, the saturation of the color, the hue of the color, etc., but only using the different components of the base color for comparison, so Handling some images that have changes in chromaticity and brightness tends to be inadequate. The HSV color space is based on the color space established by the human visual perception system to identify the chroma, saturation and brightness of the picture. This recognition

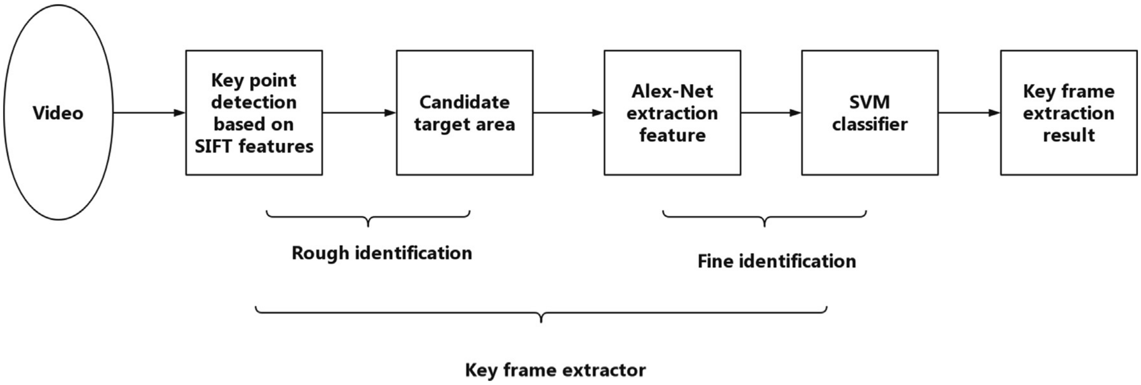

Fig. 2. Key frame extractor flow chart based on target detection.

method is also in line with human's handling of color space. Shape feature extraction: For the extraction of shape features, the main method includes the description of the shape and the region, or the combination of the two to achieve feature extraction. The boundary feature mentioned here is mainly the outer boundary of the object in the image. The outer boundary of the object in the image is related to the relevant position of the actual region of the image, but the description of the boundary shape feature is any element in the image information. Shape data can be performed using related vectors. The implementation method is various, and the typical one is the Fourier descriptor, which mainly transforms the boundary of the object and regards it as the characteristic of the target shape. Description, this method has good stability for the translation and rotation of the target. Based on the description of the feature, it is required to describe the shape data of any target in detail, such as the relevant area data of the area, the target center of gravity data, etc. Sometimes, in order to solve the related occlusion problem of the target, relevant local features are also applied. For example, the characteristics of the line, the characteristics of the circular arc, the characteristics of the corner point and the high curvature point of the target, etc., the relatively complex description methods such as the shape of the target shape invariant moment, mainly represent the seventh-order invariant moment, and its use universal. The main advantage of this target shape feature is that it can help you directly distinguish between the target and the background, focus more on the internal features of the object, and pay less attention to the external content. In fact, this and the characteristics of the color and related texture still have Larger difference. Texture feature extraction: The texture feature of an image is mainly to define the specific local correlation property of the image, that is, the correlation between the pixels within the local region of the target image. Under normal circumstances, the texture feature is more Realization is the irregular change of a certain area. The description of the characteristics of all target images is flawed. The characteristics of the texture are also the key characteristic indicators of its search. Based on the description of the texture features, many experts and scholars pointed out the application. Image feature analysis methods such as spatial gray level co-occurrence matrix method and wavelet texture analysis method.

3. Experiments

The experimental environment of this paper is: Intel (R) Core (TM) i5-4210MCPU@2.60GHZ, 8 GB memory, NVIDIA GTX850M graphics card. Use development tools based Python3.6.5 version of PyCharm, Alex-Net convolution neural network using TensorFlow framework. In the experiment, we monitored the railway prerecorded video, introduced as an input image to the key frame extractor, and SIFT features extracted by the convolutional neural network feature extraction depth, automatically adjust the network parameters and weights Based Learning depth The target detection extracts the key frame of the video image, and extracts the key frame image of the video as shown in Fig. 4. The HSV color space can accurately reflect the gray level change and color change of the image than the RGB space. According to the definition of the HS V color space, for the component values of any pixel points R, G, and B, the corresponding HSV color space is H, S. The component values of V and V are calculated by the following formulas.

$$H = a\ cos\left\{\frac{[(R-G)+(R-B)]/2}{\sqrt{(R-G)^2+(R-B)}*(R-G)}\right\} \tag{1}$$

$$S = 1 - \frac{3}{(R+B+G)}[min(R,G,B)] \tag{2}$$

$$V\ = \frac{1}{3}(R+G+B) \tag{3}$$

In 1948, American mathematician Shannon proposed the concept of entropy on the basis of thermodynamic entropy. Information entropy is a quantitative description of the information contained in an object, that is, the uncertainty of the object information is described by entropy. For an image, it can be assumed that the grayscale distribution of its pixels is independent and uncorrelated, and the grayscale distribution of the image is expressed as $p = \{p_1, p_2, \ldots\ldots p_n\}$. For the two images X and Y, their information entropy and joint information entropy can be defined as Eqs. (4)–(60 below.

$$\mathrm{H(X)} = -\sum p_x(i) log\, p_x(i) \tag{4}$$

$$\mathrm{H(Y)} = -\sum_j p_y(j) log\, p_y(j) \tag{5}$$

$$\mathrm{H(X,Y)} = -\sum_{i,j} p_{x,y}(i,j) log\, p_{x,y}(i,j) \tag{6}$$

where $p_x(i)$ Represents the probability density function of image X. $p_y(j)$ A probability density function representing the image Y $p_{x,y}(i,j)$ Represents the joint probability density function of images X and Y. The probability density function of an image can be obtained by regularizing the gray histogram of the image. The joint probability density function of the image can be obtained by regularizing the joint grayscale histogram of the two images. In the image, the mutual information indicates how much the two images contain information about each other. Generally, the larger the mutual information value I(X, Y), the higher the correlation between the two images. To accurately extract the video key frame, the key frame proposed herein based MI-SIFT feature extraction algorithm. The algorithm first converts the video V{f1, f2,... fn} into the HSV color space. Next, calculate the mutual information of two adjacent frame image entropy, a measure of similarity of two images, if the similarity is greater than a set threshold value) obtained new video subset, resulting in a series so that the final subset of the set of video V. And then calculate the standard deviation of mutual information matching algorithm according to the selected keys SURF feature points, or selecting a first subset of video frames as part of a key frame.

4. Discussion

Experimental selected video at 20 different test scenarios, and determine the value of each parameter. The S parameter is an empirical value. In this experiment, S = 7.3 is set, and the T parameter is the average of the mutual information entropy standard deviation of all video subsets vk. To verify the effectiveness of the proposed algorithm, the algorithm and key based on the K-means clustering frame extraction algorithms are compared, the comparison results are shown 3–4 in FIG. Fig. 3 detects 3 keyframes and Fig. 4 detects6 keyframes. Among them, Fig. 3 has 1 frame redundancy. By analyzing the original video, two frames of useless scenes in the video were found, one of which was two people skiing.

Fig. 3 only detects a frame with a particularly informative picture. According to this, it can be judged that the expression of the video content in Fig. 4 is more accurate. Key From the above analysis, based MI-SURF key feature extraction algorithm proposed

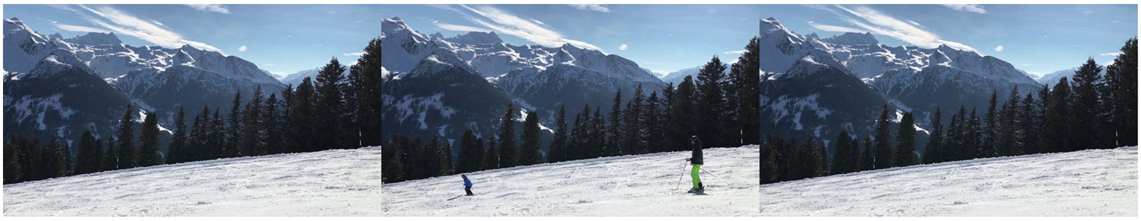

Fig. 3. Key frame extracted based on traditional algorithm.

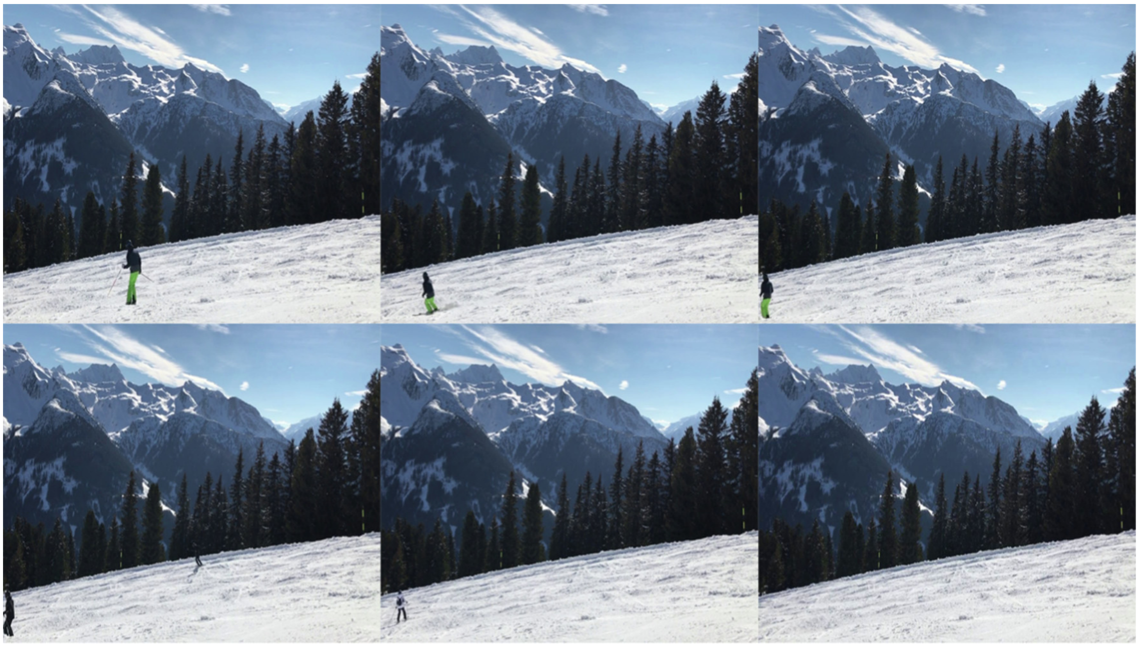

Fig. 4. Based on the key frame extracted by the algorithm of this paper.

frame than the frame K-means clustering extraction algorithm can fully convey the main content than the original video. The key frame extraction K-means clustering algorithm and the proposed algorithm comparison, there are more missed frame, expressed in the video content is not complete.

The polynomial fitting accuracy depends on whether the fitted curve can approximately reflect the actual trend of the video stream. 5 is a curve fit of discrete points (in accordance with the set sampling interval) obtained by fitting the formula, determining a sample interval is therefore an important prerequisite for the accuracy of the polynomial fit. In order to determine the optimal sampling interval Δ, the polynomial fitting accuracy of different sampling intervals is analyzed. In this paper, a series of sequence images with a total of 200 frames are used for example verification. The first video frame is provided as the key frame, the key frame 100 is the next preselected theory, by the frame bit to the left and right sides, respectively, $\Delta = 5$, 10, 15, 20, 30 to obtain the sampling interval of each frame 5 image, correlation coefficients are calculated for each video frame and key frame, a frame to determine the best position of the key frames of different sampling intervals. Distribution of discrete points resulting fitted curve, different sampling intervals selected key frames bits are shown in Table 1.

It can be seen from Table 1 that when the sampling interval Δ is 5, 10, and 15, respectively, the key frame frame difference is small; when the sampling interval Δ is 20 and 30 respectively, the key frame frame changes significantly. When the polynomial fits, the smaller sampling interval works best for the local variation trend. Since the camera is difficult to maintain steady during shooting, sudden jitter or tilt causes a sharp increase or decrease in the overlap of image sequences, which will affect the fitting results of smaller sampling intervals, such as peak forward, backward or even loss. Situation, so while ensuring the accuracy of the fit, try to increase the sampling interval. It can be seen from Table 1 that the fitting results of the sampling interval Δ of 5 and 10 are similar. In this paper, the polynomial fitting is performed with $\Delta = 10$, and the fitted key frame image and the initial key frame image overlap area can be seen, polynomial fitting the selected frame image and

Table 1
Curve fitting frame bit correspondence table.

Sampling interval	Δ key frame position	Fit value (peak)	Actual value	Overlap/%
5	97	0.9036	0.9124	61.1
10	98	0.894 4	0.9132	60.4
15	99	0.8927	0.9045	59.6
20	103	0.8937	0.8591	58.3
30	107	0.8981	0.8363	57.6

the key frame image overlap area have high similarity and high positioning accuracy.

5. Conclusion

Through the application of convolutional neural network in the field of target detection, this paper implements the design and implementation of video image key frame extractor, in which the traditional SIFT feature points are selected in the rough recognition stage, and the image depth feature is mined in the fine recognition stage using convolutional neural networks. Improve the precision of key frame extraction. Aiming at the problem of missed detection and redundancy in the original video key frame extraction, this paper proposes a key frame extraction algorithm based on MI-SURF feature. In the HSV color space, the video is segmented into different video subsets by mutual information entropy. In each video subset, the mutual information entropy standard deviation and SURF characteristics are utilized to extract key frames. Experiments show that the algorithm is good for the original video and can accurately express the original video content.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This research was supported by the Open Fund Project of the Artificial Intelligence Key Laboratory of Sichuan Province (Grant no. 2016RYY02), the Project of Sichuan Department of Science and Technology (2019YJ0476, 2018GZDZX0043) and Jouf University, Sakaka, Aljouf, KSA.

References

[1] K. Schoeffmann, M.D. Fabro, T. Szkaliczki, L. Böszörmenyi, J. Keckstein, Keyframe extraction in endoscopic video, Multimedia Tools Appl. 74 (24) (2015) 11187–11206.

[2] C. Dang, H. Radha, Rpca-kfe: key frame extraction for video using robust principal component analysis, IEEE Trans. Image Process. Publ. IEEE Signal Process. Soc. 24 (11) (2015) 3742.

[3] G.G.L. Priya, S. Domnic, Shot based keyframe extraction for ecological video indexing and retrieval, Ecol. Inf. 23 (9) (2014) 107–117.

[4] X. Luo, Q. Xu, M. Sbert, K. Schoeffmann, F-divergences driven video key frame extraction, in: International Conference on Multimedia and Expo, 2014, pp. 23–35.

[5] A.I. Ioannidis, V.T. Chasanis, A.C. Likas, Key-frame extraction using weighted multi-view convex mixture models and spectral clustering, in: International Conference on Pattern Recognition, 2014, pp. 1168–1178.

[6] H. Liu, Y. Liu, F. Sun, Video key-frame extraction for smart phones, Multimedia Tools Appl. 75 (4) (2014) 1–19.

[7] G. Lin, L. Wang, Y. Lei, J. Zhang, S.O. Technology, B.F. University, Development and experiment of blooming video monitoring system based on key frame extraction method, Trans. Chin. Soc. Agric. Eng. 30 (1) (2014) 121–128.

[8] S.D. Thepade, A.A. Tonge, An optimized key frame extraction for detection of near duplicates in content based video retrieval, in: International Conference on Communications & Signal Processing, 2014, pp. 975–980.

[9] V.T. Chasanis, A.I. Ioannidis, A.C. Likas, Efficient key-frame extraction based on unimodality of frame sequences, in: International Conference on Signal Processing, 2015, pp. 23–36.

[10] W. Wattanarachothai, K. Patanukhom, Key frame extraction for text based video retrieval using maximally stable extremal regions, in: International Conference on Industrial Networks & Intelligent Systems, 2015, pp. 336–347.

[11] R. Hannane, A. Elboushaki, K. Afdel, P. Naghabhushan, M. Javed, An efficient method for video shot boundary detection and keyframe extraction using sift-point distribution histogram, Int. J. Multimedia Inf. Retrieval 5 (2) (2016) 89–104.

[12] S.D. Thepade, P.H. Patil, Novel video keyframe extraction using KPE vector quantization with assorted similarity measures in RGB and LUV color spaces, in: International Conference on Industrial Instrumentation & Control, 2015, pp. 323–329.

[13] Y. Jing, W. Wei, W. Yang, M. Zhang, Keyframe extraction using AdaBoost, in: International Conference on Security, 2014, pp. 154–161.

[14] B.F. Momin, G.B. Rupnar, Keyframe extraction in surveillance video using correlation, International Conference on Advanced Communication Control & Computing Technologies, 2017, S0218001418550212-.

[15] X. Luo, Q. Xu, M. Sbert, K. Schoeffmann, F-divergences driven video key frame extraction, in: IEEE International Conference on Multimedia & Expo, 2014, pp. 15–16.

[16] D.S. Thepade, H.P. Patil, Novel keyframe extraction for video content summarization using lbg codebook generation technique of vector quantization, Int. J. Comput. Appl. 111 (9) (2015) 49–53.

[17] Y. Shi, H. Yang, G. Ming, L. Xiang, Y. Xia, A fast and robust keyframe extraction method for video copyright protection, J. Electr. Comput. Eng. 2017 (2017) 1–7.

[18] A. Ioannidis, V. Chasanis, A. Likas, Weighted multi-view key-frame extraction, Pattern Recogn. Lett. 72 (2016) 52–61.

[19] N.S. Kumar, G. Shobha, S. Balaji, Key frame extraction algorithm for video abstraction applications in underwater videos, Underwater Technol. (2015) 1–3.

[20] J. Valognes, M.A. Amer, N.S. Dastjerdi, Effective keyframe extraction from RGB and RGB-D video sequences, in: Seventh International Conference on Image Processing Theory, 2018, pp. 281–293.

[21] B. Yan, Y.J. Guo, Speech authentication by semi-fragile speech watermarking utilizing analysis by synthesis and spectral distortion optimization, Multimedia Tools Appl. 67 (2) (2013) 383–405.

[22] X. Huang, J. Jia, Y. Li, Z. Wang, Complex nonlinear dynamics in fractional and integer order memristor-based systems, Neurocomputing 218 (2016) 296–306.

[23] H. Wang, L.Y. Peng, H.H. Ju, Y.L. Wang, $H\infty$ state feedback controller design for continuous-time T-S fuzzy systems in finite frequency domain, Inf. Sci. 223 (2013) 221–235.

[24] H.M. Sun, R.S. Jia, Q.Q. Du, Y. Fu, Cross-correlation analysis and time delay estimation of a homologous micro-seismic signal based on the Hilbert-Huang transform, Comput. Geosci. 91 (2016) 98–104.

[25] Z. Tian, H. Xing, Y. Tan, S. Gu, S.D. Golding, Reactive transport LBM model for CO_2 injection in fractured reservoirs, Comput. Geosci. 86 (2016) 15–22.

[26] J.J. Wang, W.H. Liu, D. Chen, Y. Xu, L.Y. Zhang, A micro-machined thin film electro-acoustic biosensor for detection of pesticide residuals, J. Zhejiang Univ. Sci. C 15 (5) (2014) 383–389.

[27] F. Meng, S. Gao, Z. Song, Y. Niu, X. Li, Mesozoic high-Mg andesites from the Daohugou area, Inner Mongolia: Upper-crustal fractional crystallization of parental melt derived from metasomatized lithospheric mantle wedge, Lithos 302 (2018) 535–548.

[28] Z. Song, J. Li, Z. Gu, W. Tang, J. Yu, L. Gao, Characteristics of buried paleo-channels in the Western South Yellow Sea during the late Last Glaciation, Techn. Gazette 23 (2016), https://doi.org/10.17559/TV-20160216023828.

[29] F. Xu, B. Hu, Y. Dou, Z. Song, X. Liu, Prehistoric heavy metal pollution on the continental shelf off Hainan Island, South China Sea: From natural to anthropogenic impacts around 4.0 kyr BP, Holocene 28 (2018) 455–463.

[30] Y. Cai, G. Chen, Y. Wang, L. Yang, Impacts of land cover and seasonal variation on maximum air temperature estimation using MODIS imagery, Remote Sens. (2017) 9.

[31] J. Ning, J. Wang, T. Bu, S. Hu, X. Liu, An innovative support structure for gob-side entry retention in steep coal seam, Min. Miner. 7 (2017), n/a.

[32] P.F. Shan, X.P. Lai, Numerical simulation of the fluid-solid coupling process during the failure of a fractured coal-rock mass based on the regional geostress, Transp. Porous Media 124 (3) (2018) 1061–1079.

[33] R. Pramanik, S. Bag, Shape decomposition-based handwritten compound character recognition for bangla ocr, J. Vis. Commun. Image Represent. 50 (2018) 123–134.

[34] X. Zhou, X. Liang, X. Du, J. Zhao, Structure based user identification across social networks, IEEE Trans. Knowl. Data Eng. 30 (6) (2018) 1178–2119.

[35] Souad Chaabouni, Jenny Benois-Pineau, Chokri Ben Amar, ChaboNet: Design of a deep CNN for prediction of visual saliency in natural video, J. Visual Commun. Image Represent. 60 (2019) 79–93.

[36] S. Ding, F. Wu, J. Qian, H. Jia, F. Jin, Research on data stream clustering algorithms, Artif. Intell. Rev. 43 (4) (2015) 593–600.

[37] L. Liu, R. Ding, H. Liu, H. Liu, 3D hybrid-domain full waveform inversion on GPU, Comput. Geosci. 83 (2015) 27–36.

[38] L. Sheng, Z. Wang, L. Zou, F.E. Alsaadi, Event-based H_{∞} state estimation for time-varying stochastic dynamical networks with state-and disturbance-dependent noises, IEEE Trans. Neural Netw. Learn. Syst. 28 (10) (2017) 2382–2394.

[39] H. Duan, Q. Zeng, H. Wang, S.X. Sun, D. Xu, Classification and evaluation of timed running schemas for workflow based on process mining, J. Syst. Softw. 82 (3) (2009) 400–410.

[40] Yildiray Anagün, Sahin Isik, Erol Seke, SRLibrary: Comparing different loss function for super-resolution over various convolutional architectures, J. Visual Commun. Image Represent. 61 (2019) 178–187.

[41] Y.L. Zhang, M.X. Zhao, J.L. Su, X. Lu, K.B. Lv, Novel model for cascading failure based on degree strength and its application in directed gene logic networks, Comput. Math. Methods Med. (2018).

[42] B. Song, Z. Wang, L. Zou, On global smooth path planning for mobile robots using a novel multimodal delayed PSO algorithm, Cogn. Comput. 9 (1) (2017) 5–17.

[43] S.P. Rana, M. Dey, P. Siarry, Boosting content based image retrieval performance through integration of parametric & nonparametric approaches, J. Vis. Commun. Image Represent. 58 (2019) 205–219.