Hindawi
Mathematical Problems in Engineering
Volume 2019, Article ID 5217961, 10 pages
https://doi.org/10.1155/2019/5217961

Research Article

Key-Frame Extraction Based on HSV Histogram and Adaptive Clustering

Hong Zhao[1], **Wei-Jie Wang**[1], **Tao Wang**[1], **Zhao-Bin Chang**[1] and **Xiang-Yan Zeng**[2]

[1]*School of Computer and Communication, Lanzhou University of Technology, Lanzhou 730050, China*
[2]*Department of Mathematics and Computer Science, Fort Valley State University, Fort Valley, GA 31030, USA*

Correspondence should be addressed to Hong Zhao; 594286500@qq.com

Received 24 May 2019; Revised 12 August 2019; Accepted 23 August 2019; Published 22 September 2019

Guest Editor: Marco Perez-Cisneros

Along with the fast development of digital information technology and the application of Internet, video data begins to grow explosively. Some applications with high real-time requirements, such as object detection, require strong online video storage and analysis capabilities. Key-frame extraction is an important technique in video analysis, which provides an organizational framework for dealing with video content and reduces the amount of data required in video indexing. To address the problem, this study proposes a key-frame extraction method based on HSV (hue, saturation, value) histogram and adaptive clustering. The HSV histogram is used as color features for each frame, which reduces the amount of data. Furthermore, by using the transformed one-dimensional eigenvector, the fixed number of features can be extracted for images with different sizes. Then, a cluster validation technique, the silhouette coefficient, is employed to get the appropriate number of clusters without setting any clustering parameters. Finally, several algorithms are compared in the experiments. The density peak clustering algorithm (DPCA) model is shown to be more effective than the other four models in precision and *F*-measure.

1. Introduction

Advancements in digital storage, content distribution, and digital video recorders result in making the recording of the digital content procedure easy [1]. Handling such volume of content becomes a challenge for the implementation of the real-time applications, such as video surveillance, educational purposes, video lectures, and sports highlights [2]. The user might not have always adequate time to watch the entire video and the integral video content might not be the interest or important for the user [3]. Among all the media types (text, image, graphic, audio, and video), video is the most expressive one because it combines all the other media information together. However, video processing is a relatively time-consuming task due to the large and unstructured format of video data. Key frames provide a suitable abstraction and framework for video indexing, browsing, and retrieval. The use of key frames greatly reduces the quantity of data required in video browsing and provides an organizational framework for dealing with video content [4]. Key-frame extraction has been recognized as one of the important research issues in video information retrieval [5].

Clustering is a powerful technique for statistical data analysis, used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, data compression, and computer graphics. Traditional clustering algorithms, such as *K*-means, require some prior knowledge to determine the initial parameters, most of them need to be specified manually, and it would be a tough job to define the optimum parameter. Numerous research efforts and progresses have been done to extract video key frames in recent years, but the existing approaches have high computational complexity, and do not capture the main visual content effectively.

In this study, the cluster validation technique, silhouette coefficient [6], is employed to obtain the optimal cluster number. Regardless of video coding, video key-frame is a

relatively subjective concept, and there is no unified criterion for evaluating the quality of key frames so far. Extracting key frames with an unsupervised clustering algorithm can combine the characteristics of video content well. For video data, clustering algorithm can automatically classify video data according to their similarity, and three common clustering algorithms are used for experiments, respectively.

The main contributions of this paper lie in the following aspects: (1) The HSV histogram is used to transform high-dimensional abstract video image data into quantifiable low-dimensional data, which reduces the computational complexity while capturing image features. (2) The silhouette coefficient (SC) index (discussed in Section 2.2) is implemented to find the best prior *k-value*, which reduces time for computation. (3) Key-frame extraction is converted into clustering problem, and each cluster centroid or nearby centroid frame is declared as the key frame, and (4) the density peak clustering algorithm (DPCA) (discussed in Section 3.2.3) is proposed to extract key frames with better *F*-measure. Moreover, it only requires computing the distance between all the pairs of data points and does not need other additional prior parameters except the optimal cluster number k.

The rest of this paper is organized as follows. Section 2 provides a brief survey of the related work. In Section 3, the proposed key-frame extraction algorithms are described. The experimental results and analysis are given in Section 4. Section 5 concludes this paper.

2. Related Works

Lu et al. [7] divide the existing multimedia content research into key-frame based and video-skim based approaches as video summaries. Video summarization can be of two categories either be a sequence of frozen images which are also called storyboard or moving images called skimming [8]. The video-storyboard is defined as a group of stationary key frames, which summarizes the important video content with minimal data [9]. This class of video summaries is well explored using numerous clustering algorithms where different clusters are formed on the basis of similarity between the frames [10–13]. On the other hand, the video skimming retains the important information without losing the semantics of video sequences [14]. Generally speaking, the video-storyboard mainly analyzes the visual content rather than audio information. Its construction and expression are relatively simple, and it is often flexibly organized for browsing and indexing. Dynamic video summary makes a comprehensive consideration on multimedia information flow. It usually contains rich audio, action, and even text information, which can express the content of original video more clearly. Therefore, dynamic video summary is more entertaining and ornamental, but it is difficult to realize [15].

Data clustering is an unsupervised pattern classification method, which has been widely used in the field of video data analysis in recent years. According to the principle of minimizing the similarity between clusters and maximizing the similarity within clusters, this method clustered video data streams. Cluster centers are selected as class representations to eliminate redundancy. Jain et al. [16] classified clustering algorithms into two categories: partitioning and structuring. The former can divide data at one time to determine all classifications, while the latter need to recursively classify in a cohesive or split way. Amiri and Fathy [17] use an improved *K*-means algorithm to cluster shot-level key frames. Compared with the traditional *K*-means, the algorithm can obtain the number of clusters adaptively. Kumar et al. [8] propose a novel key-frame extraction technique to summarize the video lectures so that a reader can get the critical information in real time. Singh et al. [9] use the *k*-medoids algorithm to extract key frames and implement the Calinski–Harabasz- (CH-) based cluster validation technique to get the optimal cluster set. Similarly, Kumar et al. [18] employ Davies–Bouldin index to choose the desired number of key frames without incurring additional computational costs. Other common clustering methods include fuzzy clustering, spectral clustering, self-organizing map, and so on. [19]. However, there is no algorithm suitable for various data types and application backgrounds. Therefore, data clustering algorithms should be selected according to the characteristics of data in practical application.

2.1. Interframe Distance Metric. The distances between adjacent frames is defined as the difference of their visual content, where there can be a combination of color, texture, shape, or more [20]. Firstly, we extract total N frames with size of $W \times H$ from a test video, where W and H represent the width and height of a frame, respectively. The content of adjacent frames does not change much, but the data of three RGB channels need to be calculated separately [14]. In order to reduce the computation burden, an improved HSV histogram [21] method is used to reduce the dimensionality of data.

All N color frames of a video are converted from RGB color space into HSV color space. Then, considering the human visual resolution ability, the hue H component is divided into 12 parts, and the saturation S and value V components are divided into 5 equal parts. The quantitative formulas are as follows:

$$H = \begin{cases} 0, & \text{if } h \in [316, 20], \\ 1, & \text{if } h \in [21, 40], \\ 2, & \text{if } h \in [41, 75], \\ 3, & \text{if } h \in [76, 155], \\ 4, & \text{if } h \in [156, 190], \\ 5, & \text{if } h \in [191, 270], \\ 6, & \text{if } h \in [271, 295], \\ 7, & \text{if } h \in [296, 315], \end{cases} \qquad (1)$$

$$S, V = \begin{cases} 0, & \text{if } s, v \in [0, 0.2], \\ 1, & \text{if } s, v \in [0.2, 0.7], \\ 2, & \text{if } s, v \in [0.7, 1.0]. \end{cases}$$

Like RGB color space, any pixel is represented by three components of h, s, and v, where $h \in [0, 360]$, and

$s, v \in [0, 1]$, and these pixels are both quantified to $8 \times 3 \times 3$ color space using equation (1). The sensitivity of human eyes to the H component is greater than the S component, and the sensitivity to the S component is greater than the V component. Finally, these three color components are merged into one-dimensional feature vectors, as shown in the following equation:

$$F = HQ_S Q_V + SQ_V + V = 9H + 3S + V, \tag{2}$$

where Q_S and Q_V are the quantization parameters of S and V, respectively, and both of them are set to 3. Therefore, $F \in [0, 71]$, and each frame is converted into a one-dimensional vector with 72 attributes, which are independent of the size of the video frame. The HSV histogram is depicted in Figure 1, in which the horizontal axis represents 72 attributes of the one-dimensional feature vector F, and the vertical axis represents the number of pixels appearing on the scale of F in an image.

The interframe difference approach [3] is used to estimate the difference/changes between two consecutive frames. $\mathrm{DS}_{i,j}$ represents the difference between any two frames f_i and f_j using the Euclidean distance:

$$\mathrm{DS}_{i,j} = \left\| f_i - f_j \right\|_2 = \left(\sum_{k=1}^{n} \left| e_{k,i} - e_{k,j} \right| \right)^{1/2}. \tag{3}$$

Each frame f_i is represented by a one-dimensional vector with n attribute values, where $e_{k,i}$ is the kth element of f_i.

2.2. Optimal Cluster Number. Obtaining the optimal cluster number is a challenge. As we all know, a priori method is a more suitable technique than posteriori or determined technique to select the size of the key frame within the abstraction process [22]. To address this problem, the internal cluster evaluation method is used for the validation of clustering, namely, the silhouette coefficient (SC). The SC index value is a measure of how similar an object is to its own cluster (cohesion) compared to other clusters (separation), which ranges from −1 to +1. A high value of the SC index indicates that the object is well matched to its own cluster and poorly matched to neighboring clusters. If most objects have a high value, then the clustering configuration is appropriate. If many points have a low or negative value, then the clustering configuration may have too many or too few clusters [23]. During the experiment, a series of possible k-values are used to calculate the silhouette coefficient to obtain the optimal one, and the optimal cluster number usually appears at the maximum of the SC index. The SC index of each sample point $s(i)$ is defined in the following equation:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}, \quad -1 \le s(i) \le 1, \tag{4}$$

where $a(i)$ denotes the average distance between i and all other data within the same cluster, and $b(i)$ indicates the smallest average distance of i to all sample points in any other cluster, of which i is not a member. The average $s(i)$ over all the points of the entire dataset is the SC index, which reflects how appropriately the data have been clustered.

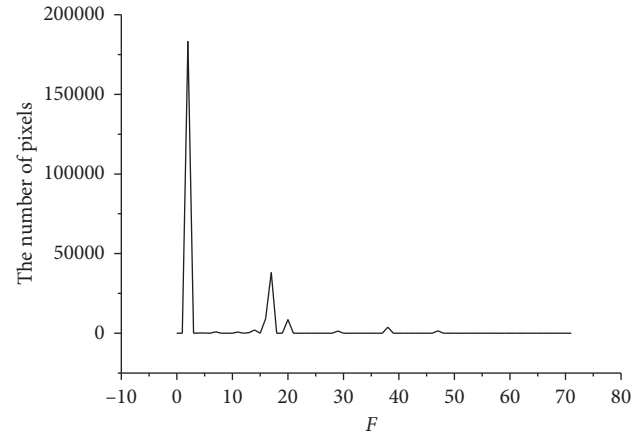

Figure 1: HSV histogram.

3. Key-Frame Extraction Algorithm

3.1. Key-Frame Extraction Algorithm in Compressed Domain. From the perspective of video coding, I-frame [24] is a complete image and the first frame of each GOP (group of pictures, a video compression technology used by MPEG), which is moderately compressed as a reference point for random access. I-frames do not require other video frames to decode, and they are encoded without reference to any other frames except themselves [25]. Therefore, we apply the I-frame method to extract the key frame. The I-frame method directly utilizes some characteristics of compressed video data to analyze and process video. A simple and feasible method of key-frame extraction in compressed domain is to extract I-frame with FFmpeg [26], and then take the extracted I-frame as the key frame.

3.2. Improved Key-Frame Extraction Algorithm in Uncompressed Domain. The uncompressed domain key frame extraction algorithm requires the decompression of videos, which takes a certain amount of time. The key frames are mainly extracted from the views of video content, so the characteristics of the video itself can be fully utilized.

When the clustering algorithm is applied to extract key frames, some frames with high similarity are clustered into a class, and the cluster center is regarded as a key frame of video. The classic clustering algorithm mainly has the following 3 categories.

3.2.1. Partition-Based Clustering Algorithm. The partition-based method generally adopts mutually exclusive partition on the dataset, which means every single object belongs to only one cluster. Most of partition-based methods are based on distance. Among them, the most classic partition-based clustering algorithm is K-means [27], whose goal is to divide the given data points into k clusters by minimizing the absolute distance between the data points and the selected cluster centers. Data point is assigned to the cluster closest to it, and the clustering center is recalculated based on the existing data points in the cluster. The cluster center and the

data points assigned to them represent a cluster. The process of the improved K-means clustering algorithm can be summarized as Algorithm 1.

3.2.2. Hierarchical Clustering Algorithm. Hierarchical clustering (also known as hierarchical clustering analysis or tree clustering) is a clustering analysis method, which seeks to establish the hierarchical structure of clustering. The algorithm does not require prespecified number of clusters. Strategies for hierarchical clustering are generally classified into two categories: AGNES (Agglomerative Nesting) and DIANA (Divisive Analysis). The agglomerative method (also called the bottom-up method) starts with each object as a separate cluster, and then merges the nearest objects until all samples are merged into the same cluster. The divisive method is also known as the top-down approach. Initially, all samples are in the same cluster, and the largest cluster is split until each object is separated [28]. Video data is a highly abstract and complex data, so the splitting of hierarchical clustering algorithms is not feasible. An improved agglomerative hierarchical clustering algorithm (AGNES) is considered, which is implemented in Algorithm 2.

3.2.3. Density-Based Clustering Algorithm. Different from the previous two clustering methods, the density-based clustering algorithm defines clusters as areas with higher density than the rest of the dataset. Clusters consist of all density-connected objects and all objects that are within these objects' range. Objects in some sparse areas are used to separate clusters, usually considered as noise and boundary points. In density-based spatial clustering of applications with noise (DBSCAN), clusters with an arbitrary shape are easily detected. However, the DBSCAN algorithm has to set a density threshold, discards points in the region where the density is lower than the threshold as noise, and assigns points with higher density than the threshold to different clusters for the disconnected region. The threshold directly affects the results of the algorithm, and choosing an appropriate one is a difficult task.

The density peak clustering algorithm (DPCA) (clustering by fast search and finding of density peaks) [29] is a new density-based clustering algorithm, which can find clusters with different densities by the visualized method, quickly find the density peak points (i.e. cluster centers) of datasets, and efficiently allot sample points and eliminate outliers [30]. It requires that each cluster has a maximum density point as the cluster center, each cluster center attracts and connects the points with lower density around it, and different cluster centers are relatively far away [31]. That is, the density peak clustering algorithm is based on two assumptions: (1) the density of cluster center is greater than that of their neighbors within the same cluster, and (2) the distance between different cluster centers and the higher density point is relatively large. Therefore, there are two main quantities that need to be calculated: local density ρ_i and distance from higher density points δ_i, which are defined as follows respectively:

$$\rho_i = \sum_{x_j \in U} \chi\left(\mathrm{DS}_{i,j} - d_c\right), \tag{5}$$

where ρ_i denotes the local density of each datum x_i, $\mathrm{DS}_{i,j}$ represents the interframe distance between the sample point x_i and x_j, and d_c indicates the cutoff distance, which depends on the value range of the empirical parameter $t \in [1\% \sim 2\%]$ in the literature [29]. And the method is robust with respect to changes in the metric that do not significantly affect the distances d_c.

$\chi(x)$ is an indicator function, which is defined as follows:

$$\chi(x) = \begin{cases} 1, & x \le 0, \\ 0, & x > 0. \end{cases} \tag{6}$$

The distance between any two points in the dataset U is calculated and sorted in ascending order. Then, the value of d_c takes the numeric value at the position t in the incremental sequence. ρ_i tells how many points are within the distance d_c. Next, the distance from higher density points δ_i is defined as follows:

$$\delta_i = \begin{cases} \max\limits_{x_j \in U, j \ne i} \left(\mathrm{DS}_{i,j}\right), & \rho_i \ge \forall \rho_j, \\ \min\limits_{j:\rho_j > \rho_i} \left(\mathrm{DS}_{i,j}\right), & \text{otherwise}, \end{cases} \tag{7}$$

where ρ_i is the global maximum value and item δ_i is the maximum distance between any other point x_j and x_i. Otherwise, item δ_i is the minimum distance between any other sample x_j and x_i, where the local density of x_j is greater than that of x_i. Therefore, DPCA aims to find data objects with high local density and large relative distance to be the centroid of the cluster. Meanwhile, these cluster centers attract and connect the points with low density around them, and they are relatively far away from each other. According to the calculation results of ρ_i and δ_i, the two-dimensional decision graph is generated to show the plot of δ_i as a function of ρ_i for each point, where transverse axis represents ρ_i and longitudinal axis represents δ_i. Some data points in the upper-right corner of the decision graph can represent different cluster centers because of their high local density and relatively high relative distance from other clusters. The process of improved density peak clustering algorithm (DPCA) is shown as Algorithm 3.

4. Experimental Result

In this section, two test datasets are utilized to verify the proposed algorithm. We use the optimal cluster number k to get two test datasets, each containing 50 videos from the Open Video Project (OVP) [32]. Each video lasts on an average of two minutes. In order to get the optimal cluster number k, the SC validation technique is applied to two test videos with a range from 3 to 15 and 3 to 20, respectively. The maximum value of SC index well describes the optimum k-value. According to the analysis of the experimental result, $k = 6$ and $k = 17$ in Figures 2 and 3 are the finest values of the SC index for two datasets, respectively. The 1st dataset contains 50 videos which the optimal cluster number k is 6,

(1) Initialization: convert each video frame into a one-dimensional vector by HSV histogram
(2) Optimal cluster number: calculate the maximum SC index for the value of k
(3) Perform K-means cluster algorithm for the given k-value
(4) Select k cluster centers for each cluster, and the centroid or nearby centroid frame is regarded as the key frame

Algorithm 1: K-means clustering algorithm for key-frame extraction.

(1) Initialization: each video frame is put into an initial cluster and converted into a one-dimensional vector
(2) Optimal cluster number: calculate the maximum SC index for the value of k
(3) Repeat
(4) Calculate the distance between any two adjacent clusters
(5) Merge two nearest clusters to generate a new cluster
(6) Until the expected k clusters are generated or other termination conditions are satisfied

Algorithm 2: AGNES clustering algorithm for key-frame extraction.

(1) Initialization: convert each video frame into a one-dimensional vector by HSV histogram
(2) Optimal cluster number: calculate the maximum SC index for the value of k
(3) Calculate the distance between any two points and sort them in ascending order
(4) Take the value at $t = 2\%$ of the incremental sequence as the cutoff distance d_c
(5) Calculate the ρ_i and δ_i for each frame and generate the decision graph
(6) Combining the decision graph to select the k cluster centers in descending order of $\gamma_i = \rho_i \delta_i$

Algorithm 3: DPCA clustering algorithm for key-frame extraction.

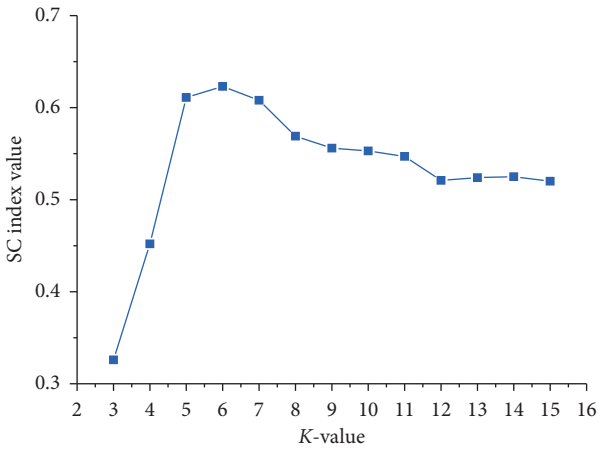

Figure 2: Selecting the maximum SC index value to get an optimal k-value on the 1st dataset.

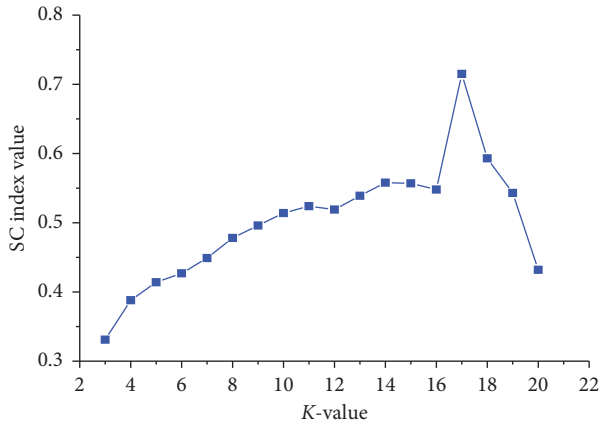

Figure 3: Selecting the maximum SC index value to get an optimal k-value on the 2nd dataset.

and the 2nd dataset contains 50 videos which the optimal cluster number k is 17.

4.1. Key-Frame Extraction Algorithm in Uncompressed Domain. Here, several experiments are being discussed to approve the efficiency and performance of the clustering-based key-frame extraction algorithm.

OVP official websites provides the storyboards for each video, and these storyboards are selected by experts and used as ground truth. Several key-frame extraction algorithms have been proposed in recent years, but there is no unified criterion to assess various models. A relatively mature approach is based on the following three metrics to assess the performance of each algorithm: precision, recall, and F-measure [33].

There are four definitions based on the ground truth and results of the algorithm:

(i) True positive (TP), a frame that belongs to both the ground truth and the output of the algorithm

(ii) False positive (FP), a frame that is selected by the algorithm but do not belong to the ground truth

(iii) True negative (TN), a frame is neither selected by the algorithm nor the key frame from the ground truth

(iv) False negative (FN), a frame that is not selected by the algorithm but belongs to the ground truth

The precision, recall, and F-measure are defined as follows:

$$\begin{aligned} \text{precision} &= \frac{\text{TP}}{\text{TP} + \text{FP}}, \\ \text{recall} &= \frac{\text{TP}}{\text{TP} + \text{FN}}, \\ F\text{-measure}\left(F_\beta\right) &= \frac{\left(1 + \beta^2\right) \times \text{precision} \times \text{recall}}{\beta^2 \times \text{precision} + \text{recall}}, \end{aligned} \tag{8}$$

where Precision indicates the ability to remove useless frames, Recall represents the ability to keep import information, and F-measure is about the harmonic mean of precision and recall ($\beta = 1$). In all, the higher value of F-measure represents more accurate algorithm.

4.2. Experiment Results of K-Means. The computation process of the K-means algorithm has been discussed in Section 3.2.1. The optimum k-values of two test videos are determined to be 6 and 17, respectively. The comparison between the key frames extracted by the K-means algorithm and the ground truth of the two test videos are shown in Figures 4 and 5.

4.3. Experiment Results of AGNES. The algorithm flow of AGNES has been discussed in Section 3.2.2. Unlike K-means, AGNES provides four methods to calculate the distance between two clusters, namely, the linkage criterion, which specifies the distance to be used between sets of observation: (1) ward-linkage minimizes the variance of the clusters being merged; (2) average-linkage uses the average of the distances of each observation of the two sets; (3) complete-linkage uses the maximum distances between all observations of the two sets; (4) single-linkage uses the minimum of the distances between all observations of the two sets. Different linkage criteria lead to different experimental results. Take the ward-linkage as an example, as shown in Figures 6 and 7. From Figure 7, the key frame extracted from test video 2 reaches 21 frames. Here, the number of key frames extracted is allowed to exceed the k-value of 17 because there are more than one discontinuous frame sequence in several clusters.

4.4. Experiment Results of DPCA. A higher t-value leads to a larger cutoff distance d_c, which makes ρ_i larger. In order to make ρ_i large, $t = 2\%$ is adopted as the experimental scheme. As discussed, the only points with high ρ and relatively high δ are likely to be the cluster centers. A hint for choosing cluster centers is provided by the plot of $\gamma_i = \rho_i \delta_i$ sorted in decreasing order, those data points with high γ_i values are most likely to become clustering centers. Then, the decision graph on the first test video is depicted in Figure 8, where points 35, 390, 450, 540, 943, and 1480 have the first six large γ_i, and they can be considered as cluster centers. In other words, f_{35}, f_{390}, f_{450}, f_{540}, f_{943}, and f_{1480} are six key frames extracted by the DPCA model. The comparison between the key frames extracted by the DPCA model and the ground truth of the two test videos are shown in Figures 9 and 10.

4.5. Experiment Results of I-Frame Method. The comparison between the average number of key frames extracted by the I-frame method and true values are shown in Table 1. Obviously, using the I-frame method to extract key frames is not effective. The I-frame method does not consider the video content, only from the perspective of video coding extracting key frame, so that redundant key frames appear. In order to further confirm the above conclusion, we add comparative experiments. The test video is a white-screen video of 20 seconds. We extracted key frames with the I-frame method. The experimental results are shown in Table 2. A large number of redundant key frames are obtained. However, the white-screen video should have only 1 key frame in theory. The shortcoming is that the video content information features used in the process of key-frame extraction are less, which leads to poor key frame quality.

4.6. Experimental Comparison. The qualitative analysis of the above experimental results shows that key frames extracted by the DPCA model are almost the ground truth. Tables 3 and 4 show the quantitative evaluation of these proposed models on two test datasets, with the best results shown in bold. Based on the above results and experimental data, we can draw a conclusion that the two key-frame evaluation indexes, precision and F-measure of the DPCA model, are larger than the other four models in this paper. Therefore, the DPCA model achieves a relatively better performance as compared with the other models.

5. Conclusion

Cluster analysis has shown promising prospects due to its successful application in many fields. In addition to classical clustering algorithms, various clustering algorithms are being put forward continuously. In practical applications, various clustering algorithms have their own appropriate scenarios, and the performance of the same algorithm varies greatly on different datasets. The DPCA algorithm only requires computing the distance between all the pairs of data points and does not need other additional prior parameters except the optimal cluster number. Its two important quantities are of great practical significance, and those frames with higher ρ means that the scenes of events they describe account for a larger proportion of the whole shot or video. Moreover, the higher δ means that the distance between different key frames is large enough to get rid of redundant key frames. The results of

Figure 4: Comparison of key frames extracted by the *K*-means model on the 1st test video with the ground truth.

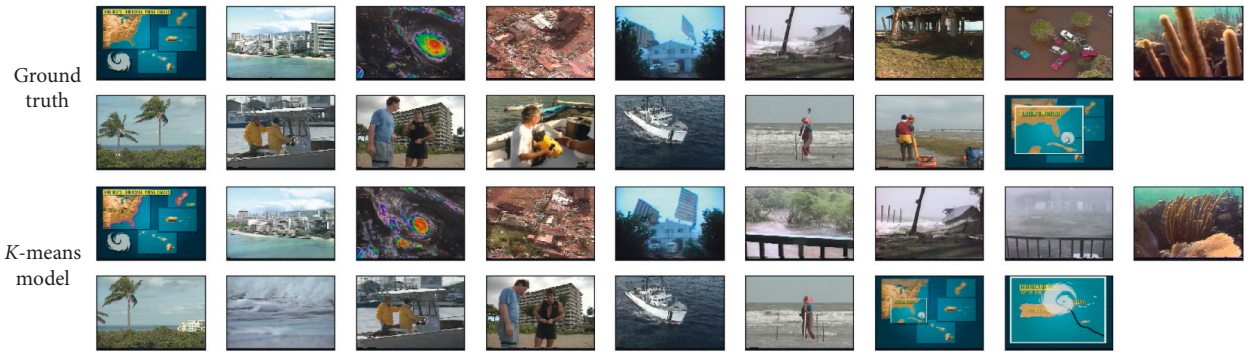

Figure 5: Comparison of key frames extracted by the *K*-means model on the 2nd test video with the ground truth.

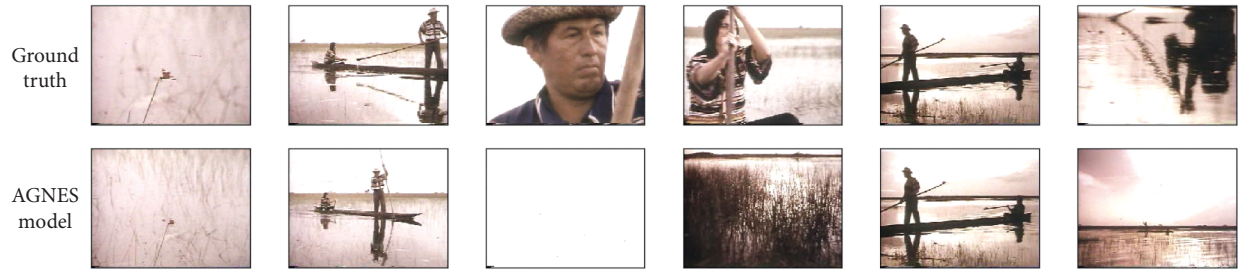

Figure 6: Comparison of key frames extracted by the AGNES model on the 1st test video with the ground truth.

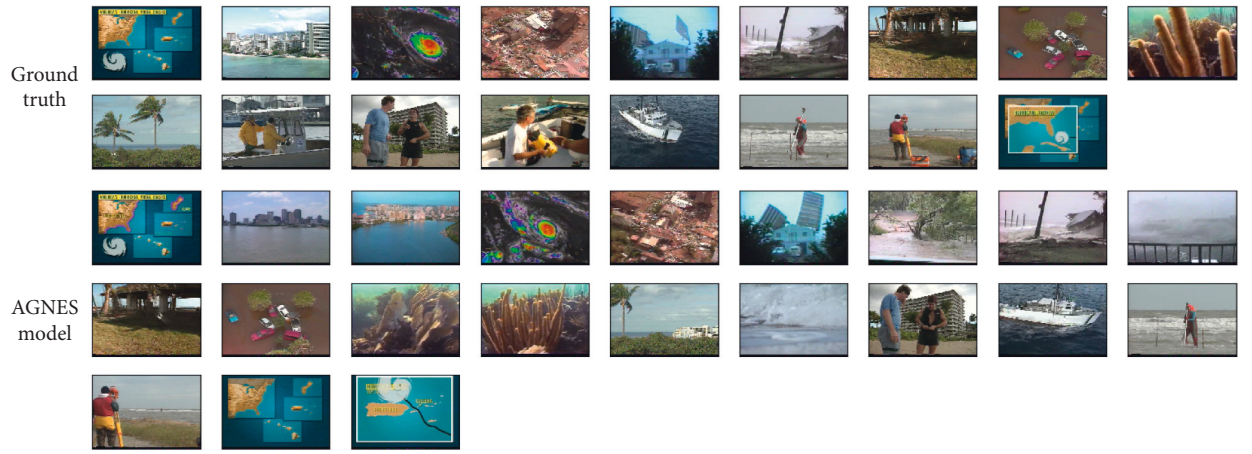

Figure 7: Comparison of key frames extracted by the AGNES model on the 2nd test video with the ground truth.

experiment show that the DPCA model has the best performance among several key frame extraction algorithms involved in this paper. Applied to environment perception module of automatic cars, the improved key-frame extraction algorithm can effectively capture the salient visual content of car-mounted cameras. Because the environment perception

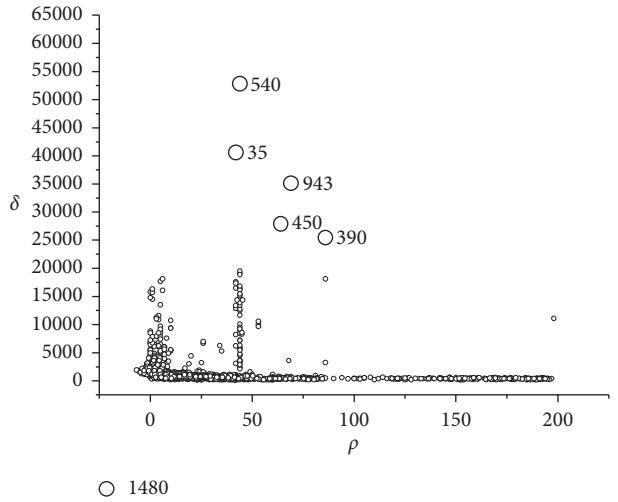

Figure 8: Decision graph for the 1st videos.

Figure 9: Comparison of key frames extracted by the DPCA model on the 1st test video with the ground truth.

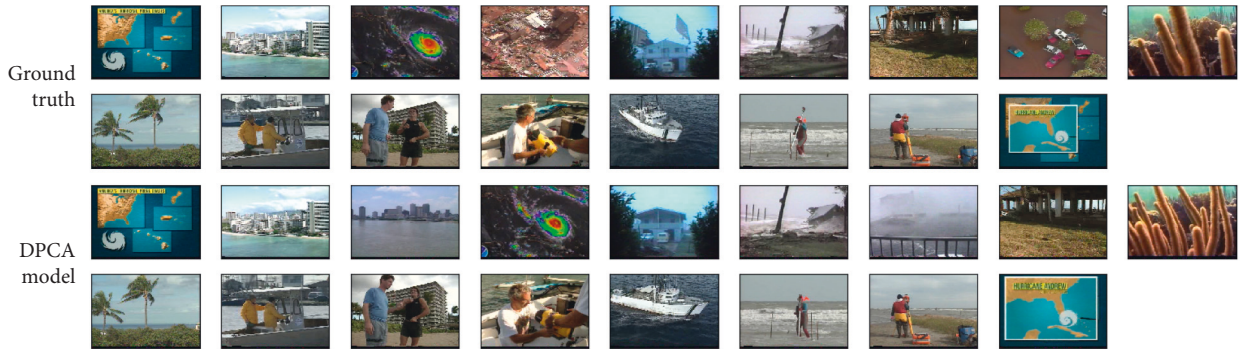

Figure 10: Comparison of key frames extracted by the DPCA model on the 2nd test video with the ground truth.

Table 1: Performance of the *I*-frame method on datasets.

Dataset	Number of key frames (true value)	The average number of key frames extracted by *I*-frame method
1st dataset	6	42
2nd dataset	17	105

Table 2: Extracting key frames of white-screen video by *I*-frame method.

Methods	Total frames	*K*-frames	Quality	Proportion (%)
I-frame method	500	42	Poor	8.40
Ideal method	500	1	Ideal	0.20

Table 3: Performance of these algorithms on the 1st dataset.

Algorithm	Precision (%)	Recall (%)	*F*-measure (%)
EVS [18]	65.9	58.8	62.1
I-frame	14.3	12.1	13.1
K-means + HSV	60.0	50.0	54.5
AGNES + HSV	43.0	64.0	51.4
DPCA + HSV	**66.6**	**60.6**	**63.4**

Table 4: Performance of these algorithms on the 2nd dataset.

Algorithm	Precision (%)	Recall (%)	*F*-measure (%)
EVS [18]	62.7	56.3	59.3
I-frame	16.2	10.7	12.9
K-means + HSV	57.6	45.3	50.7
AGNES + HSV	50.4	67.1	57.6
DPCA + HSV	**74.4**	**64.0**	**68.8**

module can only depend on a few key frames and does not require every single frame to be used to detect its surroundings environment, it is very suitable for those applications that require high real-time content analysis, for example, object detection.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This research work was supported by the National Science Foundation of China under Grant nos. 51668043 and 61262016, the CERNET Innovation Project under Grant nos. NGII20160311 and NGII20160112, and the Gansu Science Foundation of China under Grant no. 18JR3RA156.

References

[1] N. Singh, R. Arya, and R. K Agrawal, "Performance enhancement of salient object detection using superpixel based Gaussian mixture model," *Multimedia Tools and Applications*, vol. 77, no. 7, pp. 8511–8529, 2018.

[2] J. Meng, H. Wang, J. Yuan, and Y.-P. Tan, "From keyframes to key objects: video summarization by representative object proposal selection," in *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1039–1048, IEEE, Las Vegas, NV, USA, June 2016.

[3] J. Meng, S. Wang, H. Wang, J. Yuan, and Y.-P. Tan, "Video summarization via multiview representative selection," *IEEE Transactions on Image Processing*, vol. 27, no. 5, pp. 2134–2145, 2018.

[4] S. Yang and X. Lin, "Key frame extraction using unsupervised clustering based on a statistical model," *Tsinghua Science and Technology*, vol. 10, no. 2, pp. 169–173, 2005.

[5] M. Huang, L. Xia, J. Zhang, and H. Dong, "An integrated scheme for video key frame extraction," in *Proceedings of the 2nd International Symposium on Computer, Communication, Control and Automation*, pp. 258–261, Singapore, April 2013.

[6] P. J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65, 1987.

[7] S. Lu, Z. Wang, T. Mei, G. Guan, and D. D. Feng, "A bag-of-importance model with locality-constrained coding based feature learning," *IEEE Transactions on Multimedia*, vol. 16, no. 6, pp. 1497–1509, 2014.

[8] K. Kumar and D. D. Shrimankar, "F-DES: fast and deep event summarization," *IEEE Transactions on Multimedia*, vol. 20, no. 2, pp. 323–334, 2018.

[9] G. Singh, N. Singh, and K. Kumar, "PICS: a novel technique for video summarization," in *Machine Intelligence and Signal Analysis*, vol. 748, pp. 411–421, Springer, Berlin, Germany, 2017.

[10] S. Mei, G. Guan, Z. Wang, S. Wan, M. He, and D. D. Feng, "Video summarization via minimum sparse reconstruction," *Pattern Recognition*, vol. 48, no. 2, pp. 522–533, 2015.

[11] Y. Fang, X. Zhang, and N. Imamoglu, "A novel superpixel-based saliency detection model for 360-degree images," *Signal Processing-Image Communication*, vol. 69, pp. 1–7, 2018.

[12] Y. Hadi, F. Essannouni, and R. O. J. Thami, "Unsupervised clustering by *k*-medoids for video summarization," in *Proceedings of the International Symposium on Communications, Control and Signal Processing*, pp. 13–16, Marrakech, Morocco, 2006.

[13] R. Anirudh, A. Masroor, and P. Turaga, "Diversity promoting online sampling for streaming video summarization," in *Proceedings of the 2016 IEEE International Conference on Image Processing (ICIP)*, pp. 3329–3333, IEEE, Phoenix, AZ, USA, September 2016.

[14] K. Kumar, D. D. Shrimankar, and N. Singh, "Equal partition based clustering approach for event summarization in videos," in *Proceedings of the 2016 12th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS)*, pp. 119–126, IEEE, Naples, Italy, December 2016.

[15] J. Wang, X. Jiang, and T. Sun, "Review of video abstraction," *Journal of Image and Graphics*, vol. 19, no. 12, pp. 1685–1695, 2014.

[16] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: a review," *ACM Computing Surveys*, vol. 31, no. 3, pp. 264–323, 1999.

[17] A. Amiri and M. Fathy, "Hierarchical keyframe-based video summarization using QR-decomposition and modified *K*-means clustering," *EURASIP Journal on Advances in Signal Processing*, vol. 2010, Article ID 892124, 16 pages, 2010.

[18] K. Kumar, D. D. Shrimankar, and N. Singh, "Eratosthenes sieve based key-frame extraction technique for event summarization in video," *Multimedia Tools and Applications*, vol. 77, no. 6, pp. 7383–7404, 2018.

[19] C. Yu and Y.-J. Lin, "LISA: image compression scheme based on an asymmetric hierarchical self-organizing map," in *Proceedings of the International Symposium on Neural Networks*, pp. 476–485, Wuhan, China, 2009.

[20] Q. Zhang, S.-P. Yu, D.-S. Zhou, and X.-P. Wei, "An efficient method of key-frame extraction based on a cluster algorithm," *Journal of Human Kinetics*, vol. 39, no. 1, pp. 5–14, 2013.

[21] C. Xiao, "Image retrieval based on 72 HSV histogram and moment invariant," *Silicon Valley*, vol. 58, no. 2, pp. 195-196, 2012.

[22] B. T. Truong and S. Venkatesh, "Video abstraction: a systematic review and classification," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 3, no. 1, pp. 1–21, 2007.

[23] R. C. de Amorim and C. Hennig, "Recovering the number of clusters in data sets with noise features using feature rescaling factors," *Information Sciences*, vol. 324, pp. 126–145, 2015.

[24] A. Vetro, T. Wiegand, and G. J. Sullivan, "Overview of the stereo and multiview video coding extensions of the H.264/MPEG-4 AVC standard," *Proceedings of the IEEE*, vol. 99, no. 4, pp. 626–642, 2011.

[25] H. J. Zhang, J. Wu, D. Zhong, and S. W. Smoliar, "An integrated system for content-based video retrieval and browsing," *Pattern Recognition*, vol. 30, no. 4, pp. 643–658, 1997.

[26] T. Mastilovic, L. Beloica, I. Jovancic, and N. Soskic, "Spherical video player implementation using FFmpeg and SDL program support," in *Proceedings of the 2017 25th Telecommunication Forum (TELFOR)*, pp. 788–791, Belgrade, Serbia, November 2017.

[27] K. A. Jain, "Data clustering: 50 years beyond *K*-means," *Pattern Recognition Letters*, vol. 31, no. 8, pp. 651–666, 2010.

[28] Y. Dong, Y. Zhuang, K. Chen, and X. Tai, "A hierarchical clustering algorithm based on fuzzy graph connectedness," *Fuzzy Sets and Systems*, vol. 157, no. 13, pp. 1760–1774, 2006.

[29] A. Rodriguez and A. Laio, "Clustering by fast search and find of density peaks," *Science*, vol. 344, no. 6191, pp. 1492–1496, 2014.

[30] Q. Zhang and H. Zhang, "Clustering by fast search and find of density peaks based on manifold distance," *Computer Knowledge and Technology*, vol. 13, pp. 179–182, 2017.

[31] H. Zhao, T. Wang, and X. Zeng, "A clustering algorithm for key frame extraction based on density peak," *Journal of Computer and Communications*, vol. 6, no. 12, pp. 118–128, 2018.

[32] OVP, Video Open Project Storyboard, 2019, https://open-video.org/results.php?size=extralarge/.

[33] H. Zhao, Z. Chang, G. Bao, and X. Zeng, "Malicious domain names detection algorithm based on *N*-gram," *Journal of Computer Networks and Communications*, vol. 2019, Article ID 4612474, 9 pages, 2019.