Video Shot Boundary Detection Based on Feature Fusion and Clustering Technique

FENG-FENG DUAN AND FEI MENG
School of Journalism and Communication, Hunan Normal University, Changsha 410081, China
Hunan Social Public Opinion Monitoring and Network Public Opinion Research Center, Changsha 410081, China
Corresponding author: Feng-Feng Duan (dffeng2010@126.com)

This work was supported by the National Social Science Fund Project of China “Research on intelligent acquisition, analysis, and processing technology and application of cross media network public opinion big data” under Grant 18BXW109.

ABSTRACT For the problems of low accuracy and high complexity in detection of gradual shot boundary and long shot, a new video shot boundary detection algorithm based on feature fusion and clustering technique (FFCT) is proposed. In the algorithm, the interval frames of video sequence are selected, converted to gray images and scaled by sampling. With the frames, the speed-up robust features (SURF) and fingerprint features are extracted from non-compressed domain and compressed domain, and then the extracted features are fused. Next, K-means method is used to cluster the fused features, and linear discriminant analysis (LDA) is introduced to map the clusters to realize cohesion within classes and looseness among classes. Finally, the correlation of the feature classes between frames is calculated, and the features in each class are selected through density calculation and matched to realize the coarse detection and fine detection of video shot boundary. In the experiment, compared with the latest representative algorithms, it has the highest accuracy for the proposed algorithm. In particular, the detection of gradual shot boundary and long shot are also more accurate. Meanwhile, the average time consumption is also reduced. The experimental results show that the proposed algorithm has high accuracy and time efficiency, especially for gradual shot boundary and long shot detection.

INDEX TERMS Shot boundary detection, feature fusion, clustering, mapping.

I. INTRODUCTION

Video shot boundary detection, which is also known as video shot detection and temporal video segmentation [1], aims to partition a video into its basic units (shots) by detecting shot boundaries, and the frames within each unit have greater similarity, while the frames among units have greater difference. Therefore, shot boundary detection is mainly based on the differences of video sequence frames in time dimension. The features showing these differences are mainly for visual attributes [2] and other types of features such as coding information [3]. There are two types of shot boundary detection, the one is abrupt transition, and the other is gradual transition [4]. In addition, there maybe also exist long shots in video [5]. Long shot refers to a long video clip with smooth changes in content produced by the camera’s push, pull, shake, move and other actions. In the clip of long shot the previous sequence frames have great changes in content with those rear sequence frames or some sub-clip sequence frames, but there are no scene changes or switching effects. The difficulties of video shot boundary detection mainly lie in the inaccurate detection of gradual shot boundary and the fact that the long shot is mistakenly detected as multiple shot segments and contains several boundaries [5], [6]. So, in most cases the errors and low accuracy in shot boundary detection are mainly in the inaccurate detection of gradual transitions and long shots, as the reasons of object motion, camera motion, flashlights and illumination change. Video shot boundary detection is the key and foundation of content-based video analysis, index and retrieval. With the growth of video big data on the Internet and the expansion of video application range, the video content analysis and understanding, object and scene recognition, as well as content-based index and retrieval have received more attention from the researchers [7]. Therefore, the research and application of video shot boundary detection have always been concerned, and it is also a hot area in the field of video. The performance of shot boundary detection is mainly measured by accuracy and time complexity [8].

The rest of this article is organized as follows. Section II introduces the related work. Section III presents the feature fusion and clustering technique algorithm. Section IV shows
Video shot boundary detection is usually realized according to the differences of adjacent video content, so it needs to detect and extract the features of video frames for comparison. In the application, feature detection and extraction usually can be divided into two directions based on processing compressed video domain and non-compressed video domain. The non-compressed domain method refers to the algorithms based on visual features, such as histogram [4], [9]–[12], pixel [13]–[16], edge shape [17], motion [18] as well as orthogonal polynomial [19]–[22]. While the compressed domain method refers to the algorithms based on compression coding, such as entropy coding including discrete cosine transform (DCT) and discrete Fourier transform (DFT) [12], macroblock coding [23], motion vector coding [24]. In addition, there are also some algorithms that fuse several features both from compressed domain and non-compressed domain [24]–[26], which is very common in studies of shot boundary detection. With the extracted features of video frames, the methods of shot boundary detection mainly include distance similarity-based method, model-based method such as statistical model and graph segmentation model [27], machine learning-based method.

### A. SHOT BOUNDARY DETECTION BASED ON DISTANCE SIMILARITY

Shot boundary detection based on distance similarity of features is to determine whether the content change of adjacent frames is the shot boundary according to the similarity calculation and difference evaluation. This method is most commonly used in the field of shot boundary detection and segmentation. Huang et al. [13] propose a method of using local key point feature matching to detect abrupt and gradual shot boundary without modeling. Lakshmi Priya et al. [19] construct a linear expression of base vector features, with which the feature difference of adjacent frames is measured by the city block distance, so as to estimate shot changes. Tippaya et al. [25] use cumulative moving average of visual discontinuity signals to identify the shot boundaries based on cosine similarity, which can achieve the shot boundary detection of abrupt, gradual without setting the threshold. Shen et al. [26] propose a shot boundary detection algorithm based on hybrid high-level fuzzy Petri net (HLFPN) and key point matching (HLFPN-KM). With the HLFPN model and SURF, the algorithm performs pre-detection and figures out possible false shots and gradual transitions respectively. It reduces the computational complexity and improves the accuracy. Based on the point feature descriptors extracted by the ORB algorithm improved from SIFT and SURF, Liu et al. [16] use the structural similarity index (SSIM) method to calculate the similarity of feature distance to estimate the similarity of frames, so as to achieve the detection of abrupt and gradual shot boundaries. Dhiman et al. [12] implement the detection of abrupt and gradual shot boundary according to DCT feature matching and histogram feature matching respectively. Abdullussain et al. [22] realize the video segmentation and shot boundary detection by calculating the similarity of orthogonal polynomial features.

For the algorithms of shot boundary detection by feature distance similarity calculation, some researches adopt the method of adaptive threshold [9], [14], [28], but in many other cases, the method of threshold setting is also introduced [15]. De Klerk et al. [15] propose the algorithm of shot boundary detection by threshold setting according to the calculation of the pixel value differences using Jensen Shannon divergence (JSD). In the field of shot boundary detection, there still exists difficulties in choosing suitable thresholds for different videos, and empirical thresholds usually lead to low precision [29].

### B. SHOT BOUNDARY DETECTION BASED ON MODEL

The shot boundary detection algorithm based on model should design the detection model firstly according to the common gradual type of shot, and then match the shots in videos to realize detection. Han et al. [24] integrate the motion vector and regional histogram features into a unified model which employs support vector machines to realize the detection of abrupt and gradual shot boundaries for sports videos. The algorithm has a good effect on football videos. With the extracted MPEG macroblock features in compressed domain, Ren et al. [23] classify the candidate cuts into five sub-spaces via pre-filtering and rule-based decision making. Then, the phase correlation of DC images is used to estimate the shot boundary to realize the detection. Mohanta et al. [17] propose a unified model for detecting different types of video shot transitions. The algorithm uses edge strength features and motion matrix to divide frames into three types, including no change, abrupt change and gradual change. Without using threshold and sliding window, it can realize shot segmentation according to frame transition model and frame model. Bi et al. [10] propose a new framework based on dynamic mode decomposition (DMD) for shot boundary detection by dividing frames into several temporal foreground modes and one temporal background mode. This algorithm is dependent on a mode setting of background changing in different shots and unchanging in the same shot, which is obviously not suitable for gradual transitions and long shots. The shot boundary detection method based on model usually has strong pertinence. It needs to design models for all kinds of gradual types, and it usually has better detection effects for certain shot transitions, but it is more sensitive to motion in videos. Moreover, modeling needs some prior knowledge and also has high complexity, which is suitable for the professional field but not universal.
C. SHOT BOUNDARY DETECTION BASED ON MACHINE LEARNING

In recent years, with the rise and application of artificial intelligence deep learning technology, the researches on shot boundary detection by machine learning method are gradually increasing. In the shot boundary detection based on machine learning, the multi-layer convolution network is usually used for comprehensive learning to the video content, rather than only using one or several features. Tong et al. [14] use convolutional neural network (CNN) to extract features for training, classifying and tagging for frames. Then, the detection of abrupt shot boundary is realized according to pixel differences and adaptive threshold, and the detection of gradual shot boundary is also realized according to the correlation of tags. Liang et al. [30] propose a video shot boundary detection method based on convolutional neural network. With GPU acceleration this method takes local frame similarity and dual-threshold sliding window similarity into consideration, which improves the accuracy and speed of shot boundary detection. Jung et al. [31] propose a method of shot classification by using convolutional neural network, and prove that CNN is efficient for supervised learning in shot detection. Gygli [32] proposes a learning and training method of end-to-end feature from pixels to final shot boundaries by using deep learning technology to estimate the shot boundary. The method also optimizes the detection time efficiency with convolutional neural network. Wu et al. [11] use the convolutional neural network to realize the shot boundary detection based on the fusion of color histogram and depth features in two steps, including the abrupt shot boundary detection and the gradual shot boundary detection respectively. Chakraborty and Thounaojam [4] propose a shot boundary detection method by optimizing the weights of Feed-Forward Neural Network (FNN), in which the frames are classified into possible transition frames and normal frames to determine abrupt and gradual transitions according to threshold setting. The method achieves high accuracy and reduces the complexity.

In addition, some researches optimize the time complexity of shot boundary detection and improve the time efficiency. Lu and Shi [9] use candidate segment selection and singular value decomposition (SVD) to accelerate shot boundary detection, in which the SVD is introduced to reduce the feature dimension to improve efficiency. Li et al. [28] propose an algorithm to improve the speed of shot boundary detection through preprocessing technology, in which the preprocessing and bisection-based comparisons are used to achieve a balance between detection accuracy and speed.

In the research and practice of shot boundary detection, many algorithms usually have a good effect on abrupt transitions, but are not good enough on gradual transitions and long shots, and the complexity is also usually high. Especially, for the machine learning technology based on convolution neural network (CNN) which is popular in recent years, there are two significant limitations in shot boundary detection: running on expensive GPU and constructing the large-scale training dataset. In other words, the efficiency and accuracy of machine learning-based methods are largely determined by the performance of graphics hardware and the size of the training dataset [16]. In order to overcome these problems existing in current related algorithms, while without high performance of graphics hardware and large-scale training dataset, we propose a video shot boundary detection algorithm based on feature fusion and clustering technique (FFCT) in this article. The contributions are that: 1) With the global and local features of compressed domain and non-compressed domain, as well as the selection of frames, the accuracy and efficiency of shot boundary detection are improved. 2) With the clustering and mapping of features, the cohesion within class and the looseness among classes are achieved, so the transition of gradual shot is clearer, and the accurate detection of gradual shot boundary is realized. 3) With the calculation of correlation and similarity in coarse and fine detection, the accurate detection of long shot is realized. However, long shot is not considered and mistakenly detected as multiple shot segments and contains several boundaries in existing algorithms.

III. THE FEATURE FUSION AND CLUSTERING TECHNIQUE ALGORITHM

The video shot boundary detection algorithm based on feature fusion and clustering technique (FFCT) proposed in this article is mainly achieved by clustering and matching of fused features, including speed-up robust features (SURF) and fingerprint features. For the selected interval frames, firstly, the frames are converted to gray images and are scaled by sampling, and the SURF in non-compressed domain and the fingerprint features in compressed domain are extracted and fused. Then, clustering and mapping are implemented for the fusion features to achieve effective aggregation of similar features. Finally, correlation calculation is executed for the feature clusters between adjacent frames to realize coarse detection, and the main features of the adjacent frames are matched to realize fine detection. The flow chart of the proposed algorithm in this article is shown in Fig. 1.

A. SELECTION OF INTERVAL FRAMES AND FEATURE EXTRACTION

Video shot is composed of a number of sequence frames and usually there is a shot boundary only after a number of frames. So, in order to reduce the computational complexity and improve the efficiency of shot boundary detection, some of the video frames can be selected for matching calculation to accelerate the detection speed. In the proposed algorithm, the effective frames are selected by using non-overlapping interval method for feature extraction and matching detection. In other words, selecting one frame every few frames as a candidate frame for calculation. The number of the interval frames is $\omega$, and the selection of interval frames is shown in Fig. 2.

In this article, for each video, starting from the first frame ($n = 1$), and then selecting frames successively with the
Gaussian filtering can be carried out first. In this study, box filtering is used to replace Gaussian filtering approximately to improve the operation speed, and then the SURF algorithm is executed to detect interest feature points.

For the detection of feature points, Hessian matrix should be constructed to generate stable edge points or mutation points. So, for the function of $f(x, y)$, which corresponding to image $I(x, y)$, the Hessian matrix of each pixel can be constructed as follows:

$$H(f(x, y)) = \begin{bmatrix}
\frac{\partial^2 f}{\partial x^2} & \frac{\partial^2 f}{\partial x \partial y} \\
\frac{\partial^2 f}{\partial y \partial x} & \frac{\partial^2 f}{\partial y^2}
\end{bmatrix},$$

(1)

in which the matrix elements are the second partial derivatives of the function $f(x, y)$.

The discriminant of Hessian matrix is

$$\det(H) = \frac{\partial^2 f}{\partial x^2} \frac{\partial^2 f}{\partial y^2} - \left(\frac{\partial^2 f}{\partial x \partial y}\right)^2.$$

(2)

When the discriminant of Hessian matrix obtains the local maximum or minimum, it can be determined that the current point is brighter or darker than the other points in the neighborhood. Then through the construction of scale space and the location of feature points, the stable feature points are finally selected and determined [33].

The 4 × 4 rectangular area blocks are divided in the neighborhood of the feature point, so the 16 sub areas in the main direction of the feature point can be formed, and the direction of each sub area is along the main direction. For each sub area, 25 Haar wavelet features are counted in the horizontal and vertical directions of the pixels relative to the main direction. In this way, the sum of the features can be calculated, including the horizontal direction, the absolute value of the horizontal direction, the vertical direction and the absolute value of the vertical direction, and they can be expressed as $\sum dx$, $\sum |dx|$, $\sum dy$, $\sum |dy|$ respectively. They are the four feature vectors formed by each sub region, so that each feature point can form $4 \times 4 \times 4 = 64$ dimensional feature vectors, and finally the feature vectors of each point can be expressed as one-dimensional linear feature vector, which is represented by:

$$F_s(I) = (x_1^s, x_2^s, \ldots, x_n^s),$$

(3)

where $I$ represents the frame image, $F_s(I)$ is the point feature function, $x_i^s$ is one feature, $i \in \{1, 2, \ldots, n\}$ and $n = 64$.

2) FINGERPRINT FEATURE EXTRACTION

Fingerprint feature extraction refers to use a specific algorithm to calculate features for the selected interval frames and generate a set of feature values as the fingerprint [34]. At present, there are many algorithms about fingerprint feature extraction, such as Average Hash (aHash), Perceptual Hash (pHash) [34]. These algorithms usually need binary calculation and hash processing. In order to obtain more accurate and robust fingerprint features quickly, the Discrete

1) SURF EXTRACTION

For the selected interval sequence frames, the SURF point detection algorithm is used for feature extraction. As an excellent algorithm of local feature extraction, SURF has rotation and scale invariance. In order to reduce the complexity of feature point detection, improve the efficiency and consistency of detection, and obtain scale independence,
Cosine Transform (DCT) is used in the proposed algorithm of this article to obtain DCT coefficients. Then some of them are selected as the global features of frame image while without hash processing. The algorithm flow is as follows:

Step 1: Preprocessing. The input image I is preprocessed by Gaussian filtering to remove high frequency noise and then the image is divided into \( N \times N \) matrix blocks.

Step 2: The DCT. DCT is executed for input image I to obtain DCT matrix of \( N \times N \).

Step 3: Feature extraction. For the DCT matrix, selecting the region of \( 8 \times 8 \) in the upper left corner, which is the low frequency area of the image and has large of energy.

Step 4: Feature normalization. In zigzag order, the obtained 64 feature values are arranged into one-dimensional linear fingerprint features can be expressed as:

\[
F_d(I) = (x_1^d, x_2^d, \ldots, x_n^d),
\]

where \( I \) represents the frame image, \( F_d(I) \) represents the function of extracting fingerprint features, \( x_i^d \) is one feature, \( i \in \{1, 2, \ldots, n\} \) and \( n = 64 \).

3) FEATURE FUSION

In order to achieve the consistency of feature contribution ratio and component weight in multi feature fusion, Gaussian normalization is improved and applied to the point and fingerprint feature vectors which are extracted from non-compressed domain and compressed domain. In the implementation, the specific method is to combine the linear feature vectors of each point with the linear fingerprint feature vectors to form \( 2p \) dimensional linear feature vectors. While taking the \( 2p \) dimensional feature vector sequence as a row, a matrix can be formed by combining multiple feature sequences. The normalization expression of each feature vector can be defined as:

\[
g(i, j) = \frac{1}{2} \left( \frac{f(i, j) - M_j}{3\sigma_j} + 1 \right),
\]

where \( f(i, j) \) and \( g(i, j) \) are the feature vector values of the column \( j \) in the row \( i \) before and after normalization, \( M_j \) and \( \sigma_j \) are the mean value and standard deviation of the feature vector in the column \( j \) respectively, if \( g(i, j) > 1 \), then \( g(i, j) = 1 \), and if \( g(i, j) < 0 \), then \( g(i, j) = 0 \).

The consistency optimization of different types of features can be realized according to the method of improved Gaussian normalization. Finally, the normalized point features and fingerprint features are fused to obtain the feature vectors of video frame images and any one feature element in frame image can be represented as \( x = (x_1, x_2, \ldots, x_k) \), where \( k = 2p \), and \( p = 64 \) in this article.

B. FEATURE CLUSTERING AND OPTIMIZATION

The features of different regions in frame image often have certain similarity, especially in adjacent or repeated regions. Therefore, K-means algorithm can be used to cluster the

fusion features in intra frame. For the K-means algorithm, the key is to determine the number of clusters. If the number of clusters is large, it will inevitably increase the complexity of correlation calculation of clusters between adjacent frames, while if the number of clusters is small and there are too many elements in each cluster, it will increase the complexity of intra cluster calculation. In this article, the number of clusters \( K \) is set as \( \log_2 N \), where \( N \) is the number of fusion feature elements. In the method for determining the number of clusters, with the increase of \( N \) in the detection and extraction of features, the clusters indicate a slight increase, which is nonlinear, while the elements in the cluster indicate a greater increase. When the correlation of clusters between adjacent frames is small, the similarity of elements in the cluster does not need to be calculated and the complexity of similarity matching of features is greatly reduced. So, the efficiency of shot boundary detection can be improved, which meets the needs of practical application. After feature element clustering, \( K \) classes can be formed for each frame image, which are represented as \( C = (C_1, C_2, \ldots, C_K) \), the number of elements in each class is \( N_1, N_2, \ldots, N_K \) respectively, and the number of elements in any class is expressed as \( N_k \), where \( k \in K \).

In order to achieve better clustering, make the separation among classes clearer, and achieve the cohesion within class and the looseness among classes, in this article the clustering is optimized based on linear discriminant analysis (LDA). LDA was first proposed by Fisher and then widely introduced into pattern recognition, artificial intelligence and other related fields [35]. Its basic idea is to map high-dimensional feature samples to the best low-dimensional identification vector space, so as to achieve the maximum inter class distance and the minimum intra class distance in the new feature space. In this way, it can realize the clustering of the same kind of samples and the dispersion of the different kinds of samples, so as to realize the optimization of clustering and separation. At present, most of the researches on classification using LDA are for two classes in two-dimension, but for our research in this article it is innovated to achieve multi classification in two-dimension.

If the fusion feature element is expressed as \( W = (w_1, w_2) \), in which \( w_1 \) and \( w_2 \) are each one-dimensional point feature and fingerprint feature respectively, that are the linear point feature sequence and the linear fingerprint feature sequence in each line of the feature fusion matrix. The mapping result based on LDA is \( Y = (y_1, y_2) \), then \( y_i = w_i^T x \), \( Y = W^T x \). The incompactness within the class can be defined by:

\[
S_w = \sum_{i=1}^{P} \sum_{x \in w_i} (x - \mu_i)(x - \mu_i)^T,
\]

where \( \mu_i \) is the mean of each class. When \( \mu_i \) is the given sample element, the incompactness among the classes can be defined by:

\[
S_B = \sum_{i=1}^{P} Q_i(\mu_i - \mu)(\mu_i - \mu)^T,
\]
where $Q_i$ represents the weight of the $i$th class and the value is determined by the percentage of the sample number in each class. The more samples, the greater the weight. $\mu$ is the mean of all samples and is defined by:

$$\mu = \frac{1}{p} \sum_{x} \frac{1}{x} \sum_{x \in W_i} Q_i \mu_j.$$ (8)

After the mapping, the compactness within the class is $S^V = W^T S_B W$, and the compactness among the classes is $S^B = W^T S_B W$. Then the key expression can be defined by:

$$J(w) = \max \left( \frac{|S^V|}{|S^W|} \right) = \max \left( \frac{|W^T S_B W|}{|W^T S_W W|} \right).$$ (9)

The aim of expression (9) is to find the maximum value of $J(w)$ in order to achieve the compactness as large as possible among classes and as small as possible within class after mapping. According to the classical Fisher linear discriminant analysis proposed by Duda et al. [36], the solution and optimal mapping direction can be obtained. Then the different clusters can be separated effectively.

C. COARSE AND FINE DETECTION OF SHOT BOUNDARIES

1) CORRELATION CALCULATION

According to the feature clustering and optimization, the high cohesion for intra class features and the high looseness for inter class features of each frame image can be achieved. Then, by calculating the correlation of classes between frames, the correlation of adjacent frames can be estimated. So, the coarse detection of video shot boundary according to the correlation can be realized. For the correlation calculation of classes, for example, the center point $p_0$ of any one $C_p$ of the $K$ classes formed in the current frame $t$, and its corresponding fusion feature is $X_{t(p_0)}$, which is the average value of each element in this class. Similarly, the center point feature of a class in the adjacent previous frame $t-1$ is $X_{t-1(q_0)}$, while there is no previous frame, it indicates the beginning of video. Then, the correlation measure of adjacent frames based on class similarity can be defined as:

$$FrameRel(X_t, X_{t-1}) = \frac{1}{\sigma(p_0)^K} \sum_{i=1}^{K} \text{sim}(X_{t(p_0)}, X_{t-1(q_0)}),$$ (10)

where $\sigma(p_0)$ is the standard deviation of the feature vector of $X_{t(p_0)}$ relative to $X_{t-1(q_0)}$, $\text{sim}(X_{t(p_0)}, X_{t-1(q_0)})$ represents the similarity value of $X_{t(p_0)}$ and $X_{t-1(q_0)}$ for each class and is determined by calculating the similarity of the feature vectors according to the Euclidean distance.

As the value of Euclidean distance and the value of similarity are in inverse relationship, the smaller the correlation measure value of $FrameRel(X_t, X_{t-1})$ is, the greater the correlation between two adjacent frames is. When the correlation is large, it means that it may not be shot change, while the correlation is small, it may be shot change. So, the coarse detection of shot boundaries can be realized. The high cohesion for intra class and high looseness for inter classes based on LDA in clustering mapping can make the frames with correlation more relevant, while the frames with non-correlation less relevant. Accordingly, the dissimilarity at the shot boundary will greatly increase, but at the non-shot boundary will greatly reduce. The effects of coarse detection according to correlation are that it can effectively detect the abrupt and some long shot boundaries, and the gradual shot boundary can be well discretized. But at the same time, there may also be a problem that due to the constant change in content of long shot, it may be divided and discretized, and considered as multiple gradual shot boundaries.

2) MAIN FEATURE SIMILARITY MATCHING

When there is small correlation and slight similarity between adjacent frames, it is usually the shot boundary and there is no need further calculation of feature matching. If the correlation as well as the similarity of adjacent frames is large, it will be more difficult to judge whether it is the shot boundary or not and needs a further similarity matching of features. There are two purposes for this, the one is to realize the detection and estimation of gradual shot boundary, and the other is to realize the combination of non-shot change in long shot.

Because of the clustering in frame, as well as the mapping of aggregation and hash based on LDA, the elements within class are less differentiated and those among classes are more differentiated. Therefore, for the similarity calculation of features, only some elements of each class need to be matched, which can greatly reduce the time complexity of feature matching. For the selection of matching elements in class, the concept of density is introduced. In a class, the higher the density of an element, the greater its contribution to the whole class, and the corresponding features will be closer to the interest features in the image. So, they can be considered as the main features of the image using for similar matching.

For the number of $K$ classes formed after initial clustering in a frame $t$, there are $N_p$ elements in any class $C_p (C_p \subset C)$, and the density measure of the feature vector $X_{p(x)}$ corresponding to the element $x$ can be defined as:

$$Density(X_{p(x)}) = \frac{1}{N_p} \sum_{i=1}^{N_p} e^{\text{sim}(X_{p(x)}, X_{p(y)})},$$ (11)

where $X_{p(y)}$ represents the feature vector of the element $y$ in class $C_p$, $\text{sim}(X_{p(x)}, X_{p(y)})$ is the similarity of the feature vectors of element $x$ and element $y$, and it is calculated based on Euclidean distance. For elements in a class, when the ratio of the density measure value to the maximum density measure value is less than 0.4, the feature element is selected as the main feature element.

The main feature elements of the adjacent frames are matched to realize the accurate detection of the gradual shot boundaries and long shots. The specific method is as follows: first, the matching distance of features of adjacent frames is calculated and in the matching calculation it is
obtained by Euclidean distance. If it is not the first for the current frame, it is matched with the fusion feature elements in the previous frame. If the ratio of matching distance to the maximum matching distance, that is the degree of non-matching, is less than 0.4, it is considered to find a pair of matching elements. Then the matching rate can be calculated by matching elements, which is represented as $\frac{M_p}{M_c}$ for two adjacent frames, where $M_p$ is the number of matching elements and $M_c$ is the number of feature elements of the current frame. If the matching rate of this frame is less than 30% and 10% lower than the previous frame, it is considered as shot change and judged as gradual shot boundary. Otherwise, it is judged as a non-shot boundary and merged as a long shot.

### IV. EXPERIMENTAL RESULTS AND ANALYSIS

#### A. EXPERIMENTAL SCHEME

In experiment, in order to prove the progressiveness and the superiority in practical application of the proposed algorithm, the performances are evaluated in two aspects, namely, the verification for both the public standard dataset and the actual application dataset.

#### TABLE 1. Experimental Video Sequences Information of TRECVID 2001 and 2007 Datasets.

| Video Source | Video | Number of Frames | Total of Transitions | Number of Abrupt Transitions | Number of Gradual Transitions |
|--------------|-------|------------------|----------------------|-----------------------------|------------------------------|
| TRECVID 2001| D2    | 16586            | 73                   | 42                          | 31                           |
|              | D3    | 12304            | 103                  | 39                          | 64                           |
|              | D4    | 31389            | 153                  | 98                          | 55                           |
|              | D5    | 12510            | 71                   | 45                          | 26                           |
|              | D6    | 13648            | 85                   | 40                          | 45                           |
|              | BG    | 49815            | 127                  | 126                         | 1                            |
|              | BG    | 44987            | 91                   | 91                          | -                            |
| TRECVID 2007| BG16336| 2462            | 20                   | 20                          | -                            |
|              | BG28476| 23238            | 177                  | 176                         | 1                            |
|              | BG36136| 29426            | 100                  | 88                          | 12                           |
|              | BG37309| 9639             | 19                   | 11                          | 8                            |
|              | BG57770| 15836            | 35                   | 8                           | 27                           |

On the other hand, in order to compare the accuracy of the algorithms in actual application, especially for the detection effect of high-resolution videos which are widely used at present, as the resolution of videos in datasets of TRECVID 2001 and TRECVID 2007 is low, it is more necessary to test the actual application video dataset. So, the paper adopts the video sequences about national culture, which are produced and established by China Art Science and Technology Institute and National Center for Public Cultural Development of China, and the total length of the videos is about 4318 seconds [38], [39]. There are 5 categories, including literature repertoire (for example, literature Zaju ‘Civet for Prince’ and Xi Opera ‘Pearl Tower’), drama opera (for example, local opera ‘Tale of White Serpent’, Beijing Opera ‘Wu Jiapo’ and Yang Opera ‘Duan Taihou’), heritage sites (for example, Beijing Opera ‘Facebook’, cultural site ‘Memorial Archway’ and national costume ‘Cheongsam’), cultural relics, culture education (for example, cultural publicity, cultural documentary, etc.). These videos are used as input test sequences for result evaluation, and there are various of abrupt and gradual transitions as well as long shots for each category video sequences. The time length, resolution, frame number of each category video sequences, as well as the total of transitions, the number of abrupt, gradual transitions and long shots are shown in TABLE 2. Because the FFCT algorithm proposed in this article and hybrid approach of HLFPN and key point matching (HLFPN-KM) proposed in reference [26] both adopt the fusion feature including SURF key points, and HLFPN-KM is also the latest excellent algorithm, so the experimental results of FFCT algorithm are compared with the HLFPN-KM algorithm. In addition, for ablation studies, the experimental results are also compared with the shot segmentation algorithm based on SURF (SSA-SURF) proposed in another paper [40], which does not consider the issues of gradual transition and long shot, and without feature clustering and mapping.

The experimental system is Windows 7 dual core 32-bit processor with a main frequency of 3.3GHz. The algorithm is implemented in the environment of Visual Studio 2010 with Visual C++ programming language. According to the related studies on shot boundary detection, to verify the effectiveness of the proposed algorithm, the recall, precision and F-Measure are introduced as evaluation parameters and they are defined by:

\[
\text{Recall} = \frac{\text{Correct}}{\text{Correct} + \text{Missing}} \times 100\%, \quad (12)
\]

\[
\text{Precision} = \frac{\text{Correct}}{\text{Correct} + \text{Error}} \times 100\%, \quad (13)
\]

\[
F1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \times 100\%, \quad (14)
\]

in which ‘Correct’ is the number of correctly detected transitions, ‘Missing’ is the number of missed transitions, ‘Error’ is the number of incorrectly detected transitions and $F1$ is the overall performance.
TABLE 2. Experimental Video Sequences Information of Actual Application Dataset.

| Video            | Time (S) | Resolution       | Number of Frames | Total of Transitions | Number of Abrupt Transitions | Number of Gradual Transitions | Number of Long shots |
|------------------|----------|------------------|------------------|----------------------|-----------------------------|-----------------------------|---------------------|
| Literature repertoire | 743      | 1920×1080, 1280×720 | 18622            | 108                  | 83                          | 25                          | 16                  |
| Drama opera      | 1130     | 1024×576, 720×576 | 28284            | 168                  | 122                         | 46                          | 22                  |
| Heritage sites   | 1164     | 1920×1080, 720×576| 29145            | 78                   | 17                          | 61                          | 18                  |
| Cultural relics  | 677      | 768×576           | 16940            | 60                   | 45                          | 15                          | 15                  |
| Culture education| 604      | 768×576           | 15100            | 197                  | 185                         | 12                          | 14                  |

B. SETTING OF EXPERIMENTAL PARAMETERS

1) SELECTION OF INTERVAL FRAMES IN VIDEO SEQUENCES

If the feature extraction for each frame and matching for all adjacent frames, the amount of calculation will be very large as there are a lot of frames in video, which results in large complexity in practical application. The shot also has a certain length and usually contains a number of frames, but the shot change doesn’t occur between every two frames, so it is not necessary to calculate the features of all frames and match between every two frames. Therefore, a certain number of frames can be selected in the way of interval frames for detection processing. The interval frame selection of video sequence is to select the representative key frames every few frames to detect and match the point features and fingerprint features. The smaller the parameter value of $\omega$ is, the more keyframes will be selected, and the larger the calculation amount is. This will improve the accuracy, but the recall will decrease and the effect will be worse for the long shot changes. While the larger the parameter value of $\omega$ is, the fewer keyframes will be selected, and the less the calculation amount is, however, the accuracy will be lower. In this article, the value of $\omega$ is 20 [40].

2) IMAGE SCALING IN SURF EXTRACTION

Although with scale invariant, image size has a certain impact on the complexity of SURF detection. In order to reduce the complexity of interest point detection and feature extraction as well as ensure the accuracy of features, the image of each frame is scaled by sampling to a certain size in feature detection. In this article the height is scaled to 216 pixels and the width will be scaled synchronously [40]. Moreover, before the detection and extraction of interest point SURF, the video frames are also grayed.

3) THE NUMBER OF BLOCKS OF DCT MATRIX

On the one hand, when the fingerprint features are extracted using DCT, the input image $I$ needs to be divided into $N \times N$ matrix blocks in preprocessing, and it is usually with $8 \times 8, 16 \times 16, 32 \times 32, 64 \times 64$ and so on. Most of the energy of image and sound signal is concentrated in the low-frequency region after DCT. So, the high-frequency region with less information content can be removed directly as the human eyes are not sensitive to detailed information, and a small amount of low-frequency information can be used to represent the image features. In this way, it can not only describe the global features of the image, but also reduce the feature dimension. Therefore, it is necessary to ensure that some of the high frequency signals can be eliminated after DCT. On the other hand, DCT will increase the complexity of the algorithm, and in general, the complexity of one-dimensional DCT is $O(n^2)$, and that of two-dimensional DCT is $O(n^4)$. Therefore, the matrix of DCT should not be too large, otherwise, the complexity of algorithm will be greatly increased and the efficiency of feature extraction will be reduced. When considering the two matters, it can be divided into $16 \times 16$ matrix blocks in preprocessing. For these blocks, a part of high-frequency signals can be eliminated, and the low-frequency signal features of the upper left corner $8 \times 8$ region can be obtained. The feature vectors are also the same number of dimensions with the SURF and the complexity of the algorithm will not be too high. Therefore, in this article the partition parameter of matrix blocks can be set as $N \times N = 16 \times 16$.

C. ACCURACY ANALYSIS FOR PUBLIC STANDARD DATASET

The test for public standard datasets and comparison with relevant excellent algorithms can show whether the proposed algorithm is outstanding or not. For comparison algorithms of PSOGSA [4], ORB-SSIM [16], HLFPN-KM [26] and the proposed algorithm of FFCT, the experimental results of recall, precision and $F_1$ for the public standard datasets of TRECVID 2001 and TRECVID 2007 are shown in TABLE 3.

TABLE 3. Experimental Results of Accuracy of FFCT Algorithm and Comparison Algorithms.

| Algorithm       | TRECVID 2001 | TRECVID 2007 |
|-----------------|--------------|--------------|
|                 | Recall (%)   | Precision (%)| $F_1$ (%)  | Recall (%) | Precision (%)| $F_1$ (%)  |
| PSOGSA [4]      | 96.30        | 84.07        | 90.13      | 97.53      | 91.23        | 94.27      |
| ORB-SSIM [16]   | 93.02        | 88.92        | 90.92      | 91.88      | 91.34        | 91.61      |
| HLFPN-KM [26]   | 94.65        | 88.37        | 91.40      | 96.89      | 91.86        | 94.31      |
| FFCT            | 95.88        | 90.26        | 92.99      | 97.62      | 92.05        | 94.75      |

From the experimental results, for the dataset of TRECVID 2001, the proposed algorithm of FFCT and the comparison algorithms of PSOGSA, ORB-SSIM, HLFPN-KM may have some advantages in recall or precision. The recall is the
highest for PSOGSA algorithm while the precision is FFCT algorithm. However, for the overall accuracy measure of $F_1$, FFCT algorithm is the highest, which is increased by 2.86%, 2.07% and 1.59% respectively. It shows that compared with PSOGSA algorithm, the precision has been greatly improved with a small loss of recall for FFCT algorithm. At the same time, FFCT algorithm is also superior to ORB-SSIM algorithm and HLFPN-KM algorithm in recall and precision. For the dataset of TRECVID 2007, the recall and precision of each algorithm are high, which shows that they have all achieved high accuracy. Compared with the algorithms of PSOGSA, ORB-SSIM and HLFPN-KM, the FFCT algorithm proposed in this article has improved in recall, precision and $F_1$. The recall is increased by 0.09%, 5.74%, 0.73%, the precision is increased by 0.82%, 0.71%, 0.19%, and the $F_1$ is increased by 0.48%, 3.14%, 0.44%, respectively. In addition, for the datasets of TRECVID 2001 and TRECVID 2007, the $F_1$ of HLFPN-KM algorithm is higher than that of PSOGSA algorithm and ORB-SSIM algorithm. Therefore, the HLFPN-KM algorithm is also used for comparison in actual application dataset testing.

D. EXPERIMENTAL RESULTS AND ANALYSIS FOR ACTUAL APPLICATION DATASET

1) COMPARISON AND ANALYSIS OF OBJECTIVE EFFECTS

For an excellent shot boundary detection algorithm, the recall and precision should be as close as possible to 100%. The higher the recall and precision, the higher the accuracy of the algorithm for shot boundary detection, and the better the effect of shot segmentation. The experimental results of SSA-SURF algorithm, HLFPN-KM algorithm and FFCT algorithm for actual application dataset are shown in TABLE 4, TABLE 5 and TABLE 6 respectively.

| TABLE 4. Experimental Results of SSA-SURF. |
|------------------------------------------|
| Video                    | Correct | Missing | Error | Recall (%) | Precision (%) | $F_1$ (%) |
|---------------------------|--------|---------|-------|------------|---------------|----------|
| Literature repertoire     | 90     | 18      | 10    | 83.33      | 90.00         | 86.54    |
| Drama opera               | 140    | 28      | 15    | 83.33      | 90.32         | 86.68    |
| Heritage sites            | 59     | 19      | 12    | 75.64      | 83.10         | 79.19    |
| Cultural relics           | 48     | 12      | 8     | 80.00      | 85.71         | 82.76    |
| Culture education         | 179    | 18      | 12    | 90.86      | 93.72         | 92.27    |

| TABLE 5. Experimental Results of HLFPN-KM. |
|------------------------------------------|
| Video                    | Correct | Missing | Error | Recall (%) | Precision (%) | $F_1$ (%) |
|---------------------------|--------|---------|-------|------------|---------------|----------|
| Literature repertoire     | 93     | 15      | 7     | 86.11      | 93.00         | 89.42    |
| Drama opera               | 144    | 24      | 12    | 85.71      | 92.31         | 88.89    |
| Heritage sites            | 62     | 16      | 11    | 79.49      | 84.93         | 82.12    |
| Cultural relics           | 49     | 11      | 7     | 81.67      | 87.50         | 84.48    |
| Culture education         | 182    | 15      | 9     | 92.39      | 95.29         | 93.82    |

From the experimental process and results, the shot boundary detection effects as well as the recall and precision are affected by the content and characteristics of the video sequences. For the test sequences, the recall and precision are not high for SSA-SURF algorithm without feature clustering and mapping, while HLFPN-KM algorithm and FFCT algorithm both have higher recall and precision, especially for the abrupt shot boundaries. Compared with SSA-SURF algorithm, FFCT algorithm improves recall and precision greatly, which shows that the clustering and mapping for fusion features are effective for shot boundary detection. Moreover, compared with HLFPN-KM algorithm with higher detection accuracy, FFCT algorithm proposed in this article has the best detection effects and the fewest misses and errors. For each kinds of test sequences, the recall is increased by 5.56%, 4.77%, 7.69, 8.33% and 3.55% respectively, the precision is increased by 5.02%, 3.29%, 4.54%, 8.93% and 3.66% respectively, and the $F_1$ is increased by 5.32%, 4.08%, 6.19%, 8.62% and 3.60% respectively. The average recall, average precision and average $F_1$ of all kinds of test sequences are increased by 5.98%, 5.09% and 5.56% respectively, which shows that the FFCT algorithm proposed in this article has the higher recall and precision and can be more accurate to realize the shot boundary detection.

Although many existing algorithms can detect abrupt shot boundaries accurately, they are always complex and difficult for gradual shot boundaries. Many algorithms have low accuracy in practice, and have obvious problems of missing and error in detection. The FFCT algorithm proposed in this article can not only achieve the accurate detection of abrupt shot boundary, but can also achieve the detection of gradual shot boundary more accurately. Especially, for the video sequences with lots of gradual shot boundaries, the detection effects including recall and precision of FFCT algorithm are obviously better than SSA-SURF algorithm and HLFPN-KM algorithm. The comparison of gradual shot boundary detection results is shown in TABLE 7.

From the experimental results of gradual shot boundary detection of SSA-SURF algorithm, HLFPN-KM algorithm and FFCT algorithm, the SSA-SURF algorithm without feature clustering and mapping has low detection accuracy, although some sequences have the same recall and precision as the HLFPN-KM algorithm, while the FFCT algorithm proposed in this article has obvious advantages. For each kinds of test sequences, compared with HLFPN-KM algorithm with higher detection accuracy, the FFCT algorithm improves the recall, precision and $F_1$, and the recall is increased by...
TABLE 7. The Comparison of Gradual Shot Boundary Detection Results.

| Video        | SSA-SURF [40] | HLFPN-KM [26] | FFCT      |
|--------------|---------------|---------------|-----------|
|              | Missing Error | Recall (%)    | Precision (%) | $F_1$ (%)|
| Literature repertoire | 9 5 | 64.00 76.19 | 69.57 | 7 3 | 72.00 85.71 | 78.26 | 4 1 | 84.00 95.45 |
| Drama opera   | 16 10 | 65.22 75.00 | 69.77 | 13 8 | 71.74 80.49 | 75.86 | 6 3 | 86.96 93.02 |
| Heritage sites | 10 8 | 83.61 86.44 | 85.00 | 9 7 | 85.25 88.13 | 86.67 | 5 4 | 91.80 93.33 |
| Cultural relics | 5 3 | 66.67 76.92 | 71.43 | 5 3 | 66.67 76.92 | 71.43 | 2 1 | 86.67 92.86 |
| Culture education | 5 3 | 58.33 70.00 | 63.63 | 4 3 | 66.67 72.73 | 69.57 | 3 2 | 75.00 81.82 |

TABLE 8. The Comparison of Long Shot Detection Results.

| Video        | SSA-SURF [40] | HLFPN-KM [26] | FFCT      |
|--------------|---------------|---------------|-----------|
|              | Missing Error | Recall (%)    | Precision (%) | $F_1$ (%)|
| Literature repertoire | 6 5 | 62.50 76.92 | 68.96 | 5 2 | 68.75 84.62 | 75.86 | 3 1 | 81.25 92.86 |
| Drama opera   | 9 5 | 59.09 72.22 | 65.00 | 8 4 | 63.64 77.78 | 70.00 | 4 2 | 81.82 90.00 |
| Heritage sites | 6 4 | 66.67 75.00 | 70.59 | 6 4 | 66.67 75.00 | 70.59 | 2 2 | 88.89 88.89 |
| Cultural relics | 5 3 | 66.67 76.92 | 71.43 | 4 3 | 73.33 78.57 | 75.86 | 2 1 | 86.67 92.86 |
| Culture education | 5 4 | 64.29 69.23 | 66.67 | 4 3 | 71.43 76.92 | 74.07 | 1 0 | 92.86 100.00 |

12.00%, 15.22%, 6.55%, 20.00% and 8.33% respectively, the precision is increased by 9.74%, 12.53%, 5.20%, 15.94% and 9.09% respectively, while the $F_1$ is increased by 11.10%, 14.03%, 5.89%, 18.23%, and 8.69% respectively. Accordingly, the average recall, average precision and average $F_1$ of all kinds of test sequences are increased by 12.42%, 10.50% and 11.59% respectively, which realizes the detection of gradual shot boundary more accurately and effectively.

The existence of long shot in video is an important reason for the low accuracy of shot boundary detection. Long shot is mainly produced by the camera’s push, pull, shake, move and other actions, and has a smooth transition of a long sequence frames. Usually, there is a large difference in visual content among the previous sequence frames and the rear sequence frames. So, long shot may be considered to have several shots and contain several gradual shot boundaries. In practice, many algorithms which can well implement in gradual shot boundary detection, but lead to a high error rate in long shot detection. Another obvious advantage of the FFCT algorithm proposed in this article is that it can detect long shot more accurately and improve the accuracy, and the effects are also better than the other algorithms. The comparison of long shot detection results is shown in TABLE 8.

From the experimental results of long shot detection of SSA-SURF algorithm, HLFPN-KM algorithm and FFCT algorithm, the detection accuracy of SSA-SURF algorithm is low, while the FFCT algorithm proposed in this article has obvious advantages, which shows that the algorithm is also effective. Furthermore, for each kinds of test sequences, compared with HLFPN-KM algorithm with higher detection accuracy, FFCT algorithm improves the recall by 12.50%, 18.18%, 22.22%, 13.34% and 21.43% respectively, the precision by 8.24%, 12.22%, 13.89%, 14.29% and 23.08% respectively and the $F_1$ by 10.81%, 15.72%, 18.30%, 13.80 and 22.23% respectively. Accordingly, the average recall, average precision and average $F_1$ are increased by 17.53%, 14.34% and 16.17% respectively, which greatly improves the accuracy of long shot boundary detection.

However, it should be noted that on the one hand, due to some of long shots in various test sequences and gradual shot boundaries in test sequences, when there are a small number of misses or errors, the recall and precision will be greatly reduced. So, for large-scale video shot boundary detection, especially with large number of gradual shot boundaries and long shots, the recall and precision of FFCT algorithm will be better. Of course, on the other hand, long shot contains gradual transition shots and gradual transition shot may also be long shot, so there will be errors in the detection.

2) COMPARISON AND ANALYSIS OF SUBJECTIVE EFFECTS

The accuracy of abrupt shot boundary detection is usually high for many algorithms, while the problem of missing detection is that the gradual shot boundary fails to be detected accurately, and the front and rear sequences are considered to belong to the same shot clip. For the errors in detection, one reason is that the recognition of gradual shot boundary is wrong although it can be detected. Due to the change of visual features, the accuracy of non-compressed algorithm is usually not high in gradual shot boundary detection [41]. The other is that the long shot is mistakenly detected as several shots and considered to contain several boundaries.

For the FFCT algorithm proposed in this article, it can not only detect the abrupt shot boundary accurately, but can also detect the gradual shot boundary as well as long shot more accurately. An important reason for gradual shot boundary detection more accurately is that the mapping based on LDA is introduced for clustering, which enhances the differences between shots in correlation calculation. At the same time, the main reason for long shot detection more accurately is that, although it may be detected wrongly as
non-shot boundaries in correlation calculation, the non-shot boundaries can be merged due to the hash and aggregation of the mapping, so the false boundaries can be eliminated.

For example, in the gradual transition of test sequence of national costume ‘cheongsam’, the effect displayed in frames from 9980 to 10010 is the fade in and fade out of two shots, which is shown in Fig. 3 and only 30 frames are selected and listed. As the visual content of the two shots before and after is quite different, although the fade in and fade out effect achieves a smooth transition, the HLFPN-KM algorithm and the FFCT algorithm proposed in this article both can better detect the shot boundary, and for HLFPN-KM algorithm the shot boundary is after the frame of 9990, while for FFCT algorithm the shot boundary is after the frame of 9993. In addition, although SSA-SURF algorithm detects the existence of the shot boundary, the position is not accurate, which also shows the effectiveness of FFCT algorithm.

Also, in the test sequence of national costume ‘cheongsam’, the frames from 10340 to 10401 is still the transition effect of fade in and fade out of two shots, which is shown in Fig. 4 and only 30 frames are selected and listed at average intervals. In the transition, although the before and after frames belong to different shots, there is little difference in content between the shots. In the experiment of shot boundary detection, the transition effect as well as the front and rear video clips are considered as one shot for both SSA-SURF algorithm and HLFPN-KM algorithm because of the great similarity in video content, while the boundary is detected after the frame of 10359 for the FFCT algorithm proposed in this article, which better realizes the detection of the gradual shot boundary.

For the video of long shot, for example the frames from 739 to 1159 in the test sequence of cultural site ‘Memorial Archway’, it is a shot that formed by the shake action of the camera, which is shown in Fig. 5 and only 30 frames are selected and listed at average intervals. The start and end boundaries of the shot are abrupt, but the duration of the shot is long and the content of the front and rear frames is quite different, and the similarity is also very small in visual content. The SSA-SURF algorithm, HLFPN-KM algorithm and FFCT algorithm all realize the shot detection accurately at the boundary before and after. However, in the detection, it is recognized as two shots for both HLFPN-KM algorithm and SSA-SURF algorithm, and the frames from 883 to 1005 and 874 to 1019 are considered as gradual transitions by HLFPN-KM algorithm and SSA-SURF algorithm respectively. While for the FFCT algorithm proposed in this article, it can recognize and detect the boundary as one shot accurately, which improves the accuracy.

The dissimilarity, that is \( \text{sim}(X_t, X_{t-1}) \), between adjacent frames is calculated by Euclidean distance and normalized.
into the range of [0,1]. As an example, for the frames from 9900 to 10500 of national costume 'cheongsam', the plot of dissimilarity of the algorithms SSA-SURF, HLFPN, FFCT and the ground truth of transition boundaries are shown in Fig. 6. There are four shot boundaries in the clip, the first and the third are gradual transitions, while the second and the fourth are abrupt transitions. As can be seen from Fig. 6, the plot of the algorithm FFCT is more obvious in shot boundaries, especially in gradual transitions (the third shot boundary is not detected for the other two algorithms), which can realize detection of shot boundaries accurately.

3) COMPARISON AND ANALYSIS OF TIME EFFICIENCY
In order to evaluate the efficiency of the algorithms of SSA-SURF, HLFPN-KM and FFCT, the running time of shot boundary detection is tested. In the experiment, the three algorithms are implemented to detect the shot boundary of the test sequences respectively, and the detection time and number of detected boundaries are recorded and compared, which are shown in TABLE 9.

### FIGURE 6. The plot of dissimilarity between adjacent frames from 9900 to 10500 of national costume ‘cheongsam’.

| Video            | SSA-SURF [40] | HLFPN-KM [26] | FFCT   |
|------------------|---------------|---------------|--------|
| Number (frames)  | Time (s)      | Number (frames) | Time (s) | Number (frames) | Time (s) |
| Literature repertoire | 100           | 119.26        | 101    | 96.76          |
| Drama opera | 155            | 194.67        | 159    | 149.66         |
| Heritage sites | 71             | 202.14        | 76     | 154.54         |
| Cultural relics | 56             | 114.65        | 56     | 89.98          |
| Culture education | 191           | 106.88        | 191    | 84.12          |

From the experimental results in TABLE 9, for each algorithm the time length and the number of frames of test sequences are the same, and the number of shot boundaries detected is almost the same. However, the time complexity of FFCT algorithm proposed in this article is greatly reduced compared with the other two algorithms. Moreover, compared with HLFPN-KM algorithm with higher efficiency, the time consumption is reduced by 16.39%, 20.37%, 21.50%, 18.39% and 19.01% respectively, and the average time of shot boundary detection for all sequences is reduced by 19.13%, which shows that the time complexity of FFCT algorithm is lower than that of HLFPN-KM algorithm. In view of the time complexity of the three algorithms, it can be found that in SSA-SURF algorithm, although the interval selection reduces the number of frames, there are many feature points detected in each frame and no main feature selection is executed, so the matching calculation is large and the time efficiency is low. Similarly, in HLFPN-KM algorithm, the implementation of fuzzy Petri net model and the unselected and unoptimized features lead to high complexity and long time-consuming. While for FFCT algorithm, the similarity of cluster center points of adjacent frames is only calculated when judging the correlation after clustering and mapping. Moreover, only the similarity of main features is calculated according to the feature selection and density when judging the similarity of features. So, the time complexity is optimized and the time consumption is reduced by reducing the calculation amount and complexity.

### V. CONCLUSION
The FFCT algorithm proposed in this article considers the global and local factors of the interval frames of video. It not only extracts the compressed domain features, but also extracts the non-compressed domain features, which realizes the comprehensive and accurate extraction and fusion of the features. Through clustering and mapping for the extracted features of frames, as well as the calculation of correlation and similarity, it realizes the detection of shot boundary accurately and efficiently, especially for the gradual shot boundaries and long shots. The advantages of this method are that it can achieve high cohesion and low coupling within and among classes when using feature clustering and mapping. So, it can improve the closeness of frame feature similarity within shot and expand the dissimilarity among shots, which can better achieve the accurate detection of gradual boundaries and long shots. In addition, because only the center point feature correlation and the main feature similarity are calculated, the complexity is reduced and the time efficiency of shot boundary detection is improved. It overcomes the problems of inaccuracy for gradual shot boundary and long shot detection as well as the high complexity in detection, especially for the large-scale video data. In the future research, we will focus on large-scale video content analysis and homogeneous or heterogeneous unstructured data matching and retrieval.

### REFERENCES
[1] A. Amiri and M. Fathy, “Video shot boundary detection using QR-decomposition and Gaussian transition detection,” EURASIP J. Adv. Signal Process., vol. 2009, no. 1, pp. 1–12, Feb. 2010.
[2] D. Thounaojam, T. Khelchandra, T. Jayshree, S. Roy, and K. Singh, “Colour histogram and modified multi-layer perceptron neural network based video shot boundary detection,” Int. Arab J. Inf. Technol., vol. 16, no. 4, pp. 686–693, Jul. 2019.
[3] A. SenGupta, D. Meitei Thounaojam, K. Manglem Singh, and S. Roy, “Video shot boundary detection: A review,” in Proc. IEEE Int. Conf. Electr. Comput. Commun. Technol. (ICECCT), Mar. 2015, pp. 1–6.
S. Chakraborty and D. M. Thounaojam, “A novel shot boundary detection system using hybrid optimization technique,” Int. J. Speech Technol., vol. 49, no. 9, pp. 3207–3220, Sep. 2019.

S. H. Abdulhussein, A. R. Ramli, M. I. Saripan, B. M. Mahmmod, S. A. R. Al-Haddad, and W. A. Jassim, “Methods and challenges in stationary shot boundary detection: A review,” Entropy, vol. 20, no. 4, pp. 1–42, Apr. 2018.

A. Kethsy Prabavathy and J. Devi Shree, “Histogram difference with fuzzy rule base modeling for gradual shot boundary detection in video cloud applications,” Cluster Comput., vol. 22, no. S1, pp. 1211–1218, Jan. 2019.

A. Singh, D. M. Thounaojam, and S. Chakraborty, “A novel automatic shot boundary detection algorithm: Robust to illumination and motion effect,” Signal Image Video Process., vol. 14, no. 4, pp. 645–653, Jun. 2020.

B. Chakraborty, S. Bhattacharyya, and S. Chakraborty, “Generative model based video shot boundary detection for automated surveillance,” Int. J. Ambient Comput. Intell., vol. 9, no. 4, pp. 69–95, Oct. 2018.

Z.-M. Lu and Y. Shi, “Fast video shot boundary detection based on SVD and pattern matching,” IEEE Trans. Image Process., vol. 22, no. 12, pp. 5136–5145, Dec. 2013.

C. Bi, Y. Yuan, J. Zhang, Y. Shi, Y. Xiang, Y. Wang, and R. Zhang, “Dynamic mode decomposition based video shot detection,” IEEE Access, vol. 6, pp. 21397–21407, Apr. 2018.

L. Wu, S. Zhang, M. Jian, Z. Lu, and D. Wang, “Two stage shot boundary detection via feature fusion and spatial-temporal convolutional neural networks,” IEEE Access, vol. 7, pp. 77268–77276, Jun. 2019.

S. Dhiman, R. Chawla, and S. Gupta, “A novel video shot boundary detection framework employing DCT and pattern matching,” Multimedia Tools Appl., vol. 78, no. 24, pp. 34707–34723, Sep. 2019.

C.-R. Huang, H.-P. Lee, and C.-S. Chen, “Shot change detection via local keypoint matching,” IEEE Trans. Multimedia, vol. 10, no. 6, pp. 1097–1108, Oct. 2008.

W. Tong, L. Song, X. Yang, H. Qu, and R. Xie, “CNN-based shot boundary detection and video annotation,” in Proc. IEEE Int. Symp. Broadband Multimedia Syst. Broadcast, Jun. 2015, pp. 1–5.

M. G. De Klerk, P. W. C. Venter, and P. A. J. Hoffman, “Parameter analysis of the Jensen–Shannon divergence for shot boundary detection in streaming media applications,” SAIEE Afr. Res. J., vol. 109, no. 3, pp. 171–181, Sep. 2018.

H. Liu, T.-H. Tan, and T.-Y. Kuo, “A novel shot detection approach based on ORB fused with structural similarity,” IEEE Access, vol. 8, pp. 2472–2481, Jan. 2020.

P. P. Mohanta, S. Prabha, and B. Chanda, “A model-based shot boundary detection technique using frame transition parameters,” IEEE Trans. Multimedia, vol. 14, no. 1, pp. 223–233, Feb. 2012.

A. Barjata, “Block matching algorithms for motion estimation,” IEEE Trans. Evol. Comput., vol. 8, no. 3, pp. 225–239, Apr. 2004.

P. G. G. Lakshmi and S. Domnic, “Walsh-Hadamard transform kernel-based feature vector for shot boundary detection,” IEEE Trans. Image Process., vol. 23, no. 12, pp. 5187–5197, Dec. 2014.

T. Kar and P. Kanungo, “Video shot boundary detection based on Hilbert and wavelet transform,” in Proc. 2nd Int. Conf. Man Mach. Interfacing (MAMI), Dec. 2017, pp. 1–6.

S. H. Abdulhussein, A. R. Ramli, B. M. Mahmmod, M. I. Saripan, S. A. R. Al-Haddad, and W. A. Jassim, “Shot boundary detection based on orthogonal polynomial,” Multimedia Tools Appl., vol. 78, no. 14, pp. 20361–20382, Jul. 2019.

S. H. Abdulhussein, S. A. R. Al-Haddad, M. I. Saripan, B. M. Mahmmod, and A. Hussien, “Fast temporal video segmentation based on Krawtchouk–Tchebichef moments,” IEEE Access, vol. 8, pp. 72347–72359, Apr. 2020.

J. Ren, J. Jiang, and J. Chen, “Shot boundary detection in MPEG videos using local and global indicators,” IEEE Trans. Circuits Syst. Video Technol., vol. 19, no. 8, pp. 1234–1238, Aug. 2009.

B. Han, Y. Hu, G. Wang, W. Wu, and T. Yoshigahara, “Enhanced sports video shot boundary detection based on middle level features and a unified model,” IEEE Trans. Consum. Electron., vol. 53, no. 3, pp. 1168–1176, Aug. 2007.

S. Tippaya, S. Sitjongsataporn, T. Tan, M. M. Khan, and K. Channongthai, “Multi-modal visual features-based video shot boundary detection,” IEEE Access, vol. 5, pp. 12563–12573, Jun. 2017.

R.-K. Shen, Y.-N. Lin, T. T.-Y. Juang, V. R. L. Shen, and S. Y. Lim, “Automatic detection of video shot boundary in social media using a hybrid approach of HLP FN and keypoint matching,” IEEE Trans. Comput. Soc. Comput., vol. 5, no. 1, pp. 210–219, Mar. 2018.

J. Yuan, H. Wang, L. Xiao, W. Zheng, J. Li, F. Lin, and B. Zhang, “A formal study of shot boundary detection,” IEEE Trans. Circuits Syst. Video Technol., vol. 17, no. 2, pp. 168–186, Feb. 2007.

Y.-N. Li, Z.-M. Lu, and X.-M. Niu, “Fast video shot boundary detection framework employing pre-processing techniques,” IET Image Process., vol. 3, no. 3, pp. 121–134, Jun. 2009.

Z. Yang, L. Tian, and C. Li, “A fast video shot boundary detection employing OTSU’s method and dual pauta criterion,” in Proc. IEEE Int. Symp. Multimedia (ISM), Dec. 2017, pp. 583–586.

R. Liang, Q. Zhu, H. Wei, and S. Liao, “A video shot boundary detection approach based on CNN feature,” in Proc. IEEE Int. Symp. Multimedia (ISM), Dec. 2017, pp. 489–494.

D. Jung, J.-W. Son, and S.-J. Kim, “Shot category detection based on object detection using convolutional neural networks,” in Proc. 20th Int. Conf. Adv. Commun. Technol. (ICACT), Feb. 2018, pp. 36–39.

M. Gygli, “Ridiculously fast shot boundary detection with fully convolutional neural networks,” in Proc. Int. Conf. Content-Based Multimedia Indexing (CBMI), Sep. 2018, pp. 1–4.

H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, “SURF: Speeded up robust features,” Comput. Vis. Image Underst., vol. 110, no. 3, pp. 346–359, 2008.

C. Zauner, “Implementation and benchmarking of perceptual image hash functions,” M.S. thesis, Upper Austria Univ. Appl. Sci., Upper Austria, Austria, 2010.

P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, “Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 19, no. 7, pp. 711–720, Jul. 1997.

R. O. Duda, P. E. Hart, and D. G. Stork, Pattern Classification, 2nd ed. Hoboken, NJ, USA: Wiley, 2001.

TRECVID. Accessed: May 21, 2020. [Online]. Available: http://trecvid.nist.gov

National Digital Culture Network. Audio-Visual Space. Accessed: Oct. 20, 2020. [Online]. Available: http://shating.wimas.com

National Public Culture Cloud. Audio-Visual Space. Accessed: Oct. 20, 2020. [Online]. Available: https://www.culturedc.cn/vod.html

National Public Culture Cloud. Audio-Visual Space. Accessed: Oct. 20, 2020. [Online]. Available: http://shating.wimas.com

FENG-FENG DUAN received the B.S. degree in computer science from Anhui Normal University, in 2007, the M.S. degree in computer science from Huazhong Normal University, in 2010, and the Ph.D. degree in computer science and technology from the Communication University of China, in 2016. He is currently an Associate Professor with the School of Journalism and Communication, Hunan Normal University, Changsha, China. He is also a Researcher with the Hunan Social Public Opinion Monitoring and Network Public Opinion Research Center, Changsha. His research interests include computer vision, image processing, and content-based retrieval.

FEI MENG received the B.S. degree in radio and television science from Jinan University, in 2019. She is currently pursuing the M.S. degree with the School of Journalism and Communication, Hunan Normal University, Changsha, China. She is also a Researcher with the Hunan Social Public Opinion Monitoring and Network Public Opinion Research Center, Changsha. Her research interests include multimedia communication and video content analysis.