Applied Mathematics & Information Sciences
An International Journal

http://dx.doi.org/10.12785/amis/090142

Study on a New Video Scene Segmentation Algorithm

Shaofei Wu[1,2] *and Maozhu Jin*[3,*]

[1] Hubei Province Key Laboratory of Intelligent Robot, Wuhan Institute of Technology, China
[2] School of Computer Science and Engineering, Wuhan Institute of Technology, Wuhan, China
[3] Business School, Sichuan University, Chengdu 610064, China

Received: 11 Apr. 2014, Revised: 12 Jul. 2014, Accepted: 13 Jul. 2014
Published online: 1 Jan. 2015

Abstract: A new video scene segmentation algorithm is proposed based on semantic overlapped shot linked algorithm in this paper. The video shot similarity is calculated by using multi-modality subspace correlation propagation. The experiments show that the video scene can be effectively separated by the method proposed in the paper, and the MAP values, M values reached 50%, 83.4% respectively.

Keywords: temporal associated co-occurrence; support vector machine; scene segmentation

1 Introduction

Video scene, the smallest semantic unit in video structure, is composed of one or more consecutive semantically related lens. Video scene segmentation is the key technology for video data mining and analysis of video content effectively, which has become one of the most challenging research content in the field of video retrieval.

Many different approaches for video scene segmentation has been proposed by the researches at home and abroad, in general it can be divided into three categories, namely consolidation method, decomposition method and model method. Consolidation method is a bottom-up process of constantly merging. Nevertheless, decomposition method is a top-down decomposition process constantly so as to get video scene set of not further to merger or decompose, such as Sidiropoulos [1] used high-level audiovisual to video scene segmentation, which is used of multi-feature fusion ideas to switch scene transition graph (STG) [2] into multiple sub-graph so as to obtain segmented video scene. Not only that, compared with using only low-level features of the video scene segmentation algorithm, this method has a higher accuracy and versatility. However, it cant resolve the non-linear and high-dimensional problems caused by multi-feature fusion. Model method developed in recent years is a new video scene segmentation method by using statistical model for scene modeling. Among them, the most typical is Monte Carlo algorithm based on Markov chain proposed by Zhai [3], which is the use of Monte Carlo random sampling to simulate the scene generation process, and introduces three kinds of update mode (diffusion, merger and decomposition) in order to determine the scene boundaries. This method has a good mathematical model, but it is sensitive to the model parameters used for determine the number of scene. Combined with machine learning method for video scene segmentation, not only does it not need the prior knowledge of sample distribution, but it can effectively solve the problems of non-linear, high-dimensional and parameter sensitivity. For instance, video scene segmentation algorithm based on semantics has been proposed by Cao [4], which is the use of image features of the video key frame to construct and train support vector machine (SVM) [5,6], and the video scene can be well segmented by the semantic difference of the video frame. Compared with the classic SIM [7] algorithm, the precision of this method has improved greatly, but it considers only the image features of the video frame, without considering the interaction and integration of multi-modality features in video which play an important role in reducing the ”semantic gap”, leading to the low rate of semantic concept detection.

In order to solve the problems of non-linear, high dimension and semantic concept detection in the above video scene segmentation algorithm, multi-modality video scene segmentation algorithm with semantic concept has been proposed in this paper. This method, take full account of temporal associated co-occurrence

* Corresponding author e-mail: jinmaozhu@scu.edu.cn

(TAC) [8]character between multimodal media data (image, audio, text), is the use of SimFusion [9] algorithm to calculate the similarity relations between video lenses so that it can effectively solve the non-linear problem caused by multimodal fusion. In addition, it can also effectively reduce the "curse of dimensionality" by locality preserving projections (LPP)[10], which can be a high-dimensional feature vectors map to low-dimensional semantic subspace by reducing the dimensionality of feature vectors (reflecting the similarity relations between video lenses). Then, semantic space coordinates obtained by reducing the dimensionality is as the input into the SVM so as to construct a number of different semantic concept training classifier and predict the semantic concept vectors of video key frames, by means of semantic overlapped shot linked algorithm to get the video scene at the end of this paper. The authors experiments show that the video scene can be effectively separated by the method proposed in the paper, and the MAP values, M values reached 50%, 83.4% respectively.

2 VIDEO LOW-LEVEL FEATURES EXTRACTION

Multimodal features are composed of corresponding low-level features of media data in video, and multimodal choice, integration and collaboration play an important role to eliminate the gap between the high-level semantics and the low-level features of video [11].

A. Image Features Extraction

Image feature is the basic property of the video image, which is a natural feature recognized by human visual. This paper selected color histogram, texture and edge features as the image features.

I) 72HSV color histogram

According to the characteristics of human visual system, HSV is used in this paper, hue space H divided into 8 parts, saturation space S divided into 3 parts, value space divided into 3 parts. In this way, RGB color value can be quantized to 72HSV color space and the one-dimensional feature vector can be calculated accurately, in order to obtain 72 bin color histogram , the expression is as follows:

$$Hist = (h^0, h^1, \cdots, h^{71}), 0 \le h^k \le 1, k = 0, \cdots, 71, \sum h^k = 1 \tag{1}$$

II) Texture feature extraction

Texture feature is the change of image gray level, reflecting the other property of the video image. In this paper, Gabor filter is used to extract texture feature quickly in the frequency domain [12], and the filtered sub-graph energy is selected as a component to structure texture feature vector.

Filtered sub-graph energy $E^{m,n}$ is defined as:

$$\begin{aligned} E^{m,n} &= \frac{1}{M \times N} \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} |P(x,y)|^2 \\ &= \frac{1}{M \times N} (\frac{1}{M \times N} \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} |H(x,y)|^2) \end{aligned} \tag{2}$$

Among them, $P(x,y)$ is the filtered image, $H(u,v)$ is the Fourier transform result, $M \times N$ is the image size, m, n are scale numbers and direction numbers respectively.

Under the condition of without IFFT transformation, $E^{m,n}$ can be obtained directly in the frequency domain, so that it can greatly reduce the calculation and feature extraction time.

Experiments take $m = 4$, $n = 6$, $E^{m,n}$ as the component, the image texture feature vector can be expressed as:

$$T = |E_{00}, E_{01}, \cdots, E_{35}| \tag{3}$$

III) Edge feature extraction

Edge histogram descriptor (EHD) proposed by the literature [13] is used for extracting edge feature in this paper, which makes use of the spatial distribution of the video frame images to describe edge feature. Every frame images are divided into $4 \times 4 = 16$ sub-image, and the boundary distribution of each sub-image is represented by histogram. Divided into 5 boundary types, that is, non-directional boundary, horizontal boundary, vertical boundary, 45^o boundary and 135^o boundary, correspond to 5 template types respectively. Then, calculate for the proportion of different boundary types in each sub-image, and the quantification, draw as a histogram, in this way, 16 sub-images formed the 80bin edge histogram.

In experimenting the video frame image is divided into non-overlapping square block, such as 320×240 pixels video frame image can be divided into 16 80×60 pixels sub-images, and each sub-image is divided into 12 pixels 20×20 image blocks, then each image block is further divided into 4 10×10 pixels sub-block images. By the formula (4) can be calculated the boundary strength value of the $(i,j)th$ image block.

$$\begin{aligned} ver\_edge(i,j) &= |\sum_{k=0}^{3} A_k(i,j) \times ver\_edge\_filter(k)| \\ hor\_edge(i,j) &= |\sum_{k=0}^{3} A_k(i,j) \times hor\_edge\_filter(k)| \\ dia45\_edge(i,j) &= |\sum_{k=0}^{3} A_k(i,j) \times dia45\_edge\_filter(k)| \\ dia135\_edge(i,j) &= |\sum_{k=0}^{3} A_k(i,j) \times dia135\_edge\_filter(k)| \\ mond\_edge(i,j) &= |\sum_{k=0}^{3} A_k(i,j) \times mond\_edge\_filter(k)| \end{aligned} \tag{4}$$

Among them, $A_k(i,j)$ is the average luminance value of the k-th sub-block image of the $(i,j)th$ image block, $edge\_filter(k)$ is the boundary filter coefficient. Find out the maximum of the boundary strength value, if the selected threshold value is less than or equal to the maximum, the boundary type of image block correspond to the boundary type of the maximum; if the selected threshold value is greater than the maximum, the image block is non-directional boundary type. Statistics of the image block number of 5 boundary types of each sub-image, the quantification, then the boundary type of sub-images are arranged by non-directional boundary, horizontal boundary, vertical boundary, 45^o boundary and 135^o boundary, in order to obtain the 5bin edge histogram of each sub-image, in this way, 16 sub-images formed the 80bin edge histogram.

B. Audio Features Extraction

Select from the following three kinds of characteristic quantities constitute the audio feature in this paper, and the corresponding audio sequence of video lens is defined as the audio clip. Assuming the sampling number of the audio frame is $M = L \times S$, wherein, $L(s)$ is the audio frame length, $S(Hz)$ is the sampling rate of audio signal. $I)$ Short time energy (STE)

STE refers to the accumulated average energy of all sampling signal in an audio frame, calculated as follows:

$$E_n = \frac{1}{N}\sum_m [x(n)w(n-m)]^2 \tag{5}$$

Among them, $x(n)$ is the signal value of the n-th sampling point of the m-th short time frame, $w(n)$ is a widow function of length N.

$II)$ Zero-crossing rate (ZCR)

ZCR refers to the positive and negative change number of sampling signal values in a short time frame, reflecting the average frequency of audio signal in a short time, calculated as follows:

$$Z_n = \frac{1}{2}\sum_m |sgn[x(n)] - sgn[x(n-1)]|w(n-m) \tag{6}$$

Among them, $sgn[\square]$ represents the sign function.

$III)$ Mel-frequency cepstral coefficients (MFCC)

MFCC refers to the frequency spectrum converted into cepstral domain based on non-linear characteristics of the human auditory system and cepstrum decorrelation principle, calculated as follows:

$$C_r = \sqrt{\frac{2}{M}\sum_{k=1}^{M} log(X_k) \times cos[\tau(k-0.5) \times \frac{\pi}{M}]}, \tau = 1,2,\cdots,12 \tag{7}$$

Among them, M is the triangular filter number, X_k is the k-th filter output, τ is the MFCC dimension. The audio feature can be obtained by the equation (5), (6), (7), expressed as follows:

$$A = (E_n, Z_n, C_1, C_2, \cdots, C_12) \tag{8}$$

C. Text Features Extraction

This paper is directly use of ASR speech recognition result provided by TRECVID so as to get the keyword aggregate documents of each video lens. Each document is expressed as a feature vector of the vector space, expressed as follows:

$$V(d) = (w_1(d), \cdots, w_i(d), \cdots, w_n(d)) \tag{9}$$

Among them, d represents the keyword aggregate documents, n represents the total number of video keyword, $w_i(d)$ represents the feature weight of entry t_i , calculated as follows:

$$w_i(d) = TF(t_i,d) \times IDF(t_i) \tag{10}$$

Among them, TF and IDF are used to describe features of the text content, $TF(t_i,d)$ represents the relative frequency of appearance of entry t_i in the document d, $IDF(t_i)$ can reflect the specificity of entry t_i , calculated as follows:

$$IDF(t_i) = log(\frac{N}{n_i}) \tag{11}$$

Among them, N represents the total number of video lens in this experiment, n_i represents the video lens number of entry t_i of appearance.

3 THE SIMFUSION ALGORITHM AND DIMENSIONALITY REDUCTION

Video is essentially a time-series data, including multiple media (visual, auditory, text, etc.), in many cases, which present TAC character, that is, in a certain period of time, there are multi-modal data (video frame image, audio signal, video transcript) common to express the same semantic content, and the multi-modal data are correlating coupling each other in the duration of the video [14].

By SimFusion algorithm to calculate the similarity between video lenses, it can effectively carry out the relation (TAC character) mining between heterogeneous data, in order to make the results more accurate and reasonable.

A. The SimFusion algorithm

In order to facilitate the inherent character mining between heterogeneous data as a whole, this paper defines unified relation matrix (URM), that is, heterogeneous is treated as the element of the data space; by defining the unified similarity matrix (USM), it can carry out the similarity mining between heterogeneous data as a whole, and it can also use the iterative calculation to improve the accuracy of similarity calculation.

URM can be expressed as:

$$\begin{bmatrix} \lambda_{11}L_{image} & \lambda_{12}L_{i-a} & \lambda_{13}L_{i-t} & \lambda_{14}L_{i-s} \\ \lambda_{21}L_{a-i} & \lambda_{22}L_{audio} & \lambda_{23}L_{a-t} & \lambda_{24}L_{a-s} \\ \lambda_{31}L_{t-i} & \lambda_{32}L_{t-a} & \lambda_{33}L_{text} & \lambda_{34}L_{t-s} \\ \lambda_{41}L_{s-i} & \lambda_{42}L_{s-a} & \lambda_{43}L_{s-t} & \lambda_{44}L_{shat} \end{bmatrix}$$

Among them, the correlation matrix L_{i-a} , L_{i-t} , L_{a-t} from different modal represent respectively the correlation of between image and audio, image and text, audio and text; the similarity matrix L_{image} ,L_{audio} ,L_{text} from the same modal represent respectively the similarity of between image and image, audio and audio, text and text; L_{shot} represents the similarity between video lens; the correlation matrix L_{s-i} , L_{s-a} , L_{s-t} represent respectively the correlation between image, audio, text and lens; parameter λ satisfies $\sum_{\forall j}\lambda_{ij}=1$.

USM can be expressed as:

$$\begin{bmatrix} S_{11} & S_{12} & \cdots & S_{1T} \\ S_{21} & S_{12} & \cdots & S_{2T} \\ S_{31} & S_{32} & \cdots & S_{3T} \\ \vdots & \vdots & \vdots & \vdots \\ S_{T1} & S_{21} & \cdots & S_{T1} \end{bmatrix}$$

Among them, S represents the data objects (image, audio, text, shot) correlation in the data space; represents the total number of the data objects, namely $T=4\times N$, N represents the component dimension.

By the SimFusion algorithm to calculate the similarity between i-th lens $shot_i$ and j-th $shot_j$, the specific steps are as follows:

Step1 URM initialization

1) The similarity matrix L_{image} , L_{audio} ,L_{text} initialization.

For video lens $shot_i$ and $shot_j$, the color similarity calculation formula is:

$$Sim_C(i,j)=\sum^{7}1_{k=0}min(h_i^k,h_j^k) \tag{12}$$

The texture similarity calculation formula is:

$$Sim_T(i,j)=1-\sum_k\frac{w(k)|T_i(k)-T_j(k)|}{a_k} \tag{13}$$

Among them, $w(k)$ represents the weight of each component, a_k represents normalized parameters.

The edge similarity calculation formula is:

$$Sim_E(i,j)=\frac{\sum_k min[EHD_i(k),EHD_j(k)]}{\sqrt{\sum_k EHD_i(k)\sum_k EHD_j(k)}} \tag{14}$$

In summary, the image feature similarity between the video $shot_i$ and $shot_j$ is expressed as:

$$L_{image}=w_C Sim_C(i,j)+w_T Sim_T(i,j)+w_E Sim_E(i,j) \tag{15}$$

Among them, w_C, w_T,w_E represent respectively the weight of color, texture, edge similarity relations between video lenses.

By Euclidean distance formula to calculate the audio feature similarity between the video $shot_i$ and $shot_j$, expressed as follows:

$$L_{audio}=\|A_i-A_j\|=[\sum_k(a_i(k)-a_j(k))^2]^{\frac{1}{2}} \tag{16}$$

By cosine distance formula to calculate the text feature similarity between the video $shot_i$ and $shot_j$, expressed as follows:

$$L_{text}=\frac{V_i(d)\cdot V_j(d)}{|V_i(d)||V_j(d)|}=\frac{\sum_k w_i(k)\times w_j(k)}{\sqrt{\sum_k w_i(k)^2}\sqrt{\sum_k w_j(k)^2}} \tag{17}$$

2) The correlation matrix L_{s-i} , L_{s-a} , $L_{s-t}$and L_{shot} are initialized to the unit matrix, that is, the correlation between the heterogeneous data and video lens is 1.

3) The correlation matrix L_{i-a} , L_{i-t} , L_{a-t} initialization. By canonical correlation analysis (CCA) proposed by Zhang [15] to calculate the correlation between from different modal.

Step2 To get the similarity relations between video lenses by iterative calculation, S_{usm} represents USM, L_{urm} represents URM, the specific steps are as follows:

1) $S_{usm}^{original}$ is initialized to the unit matrix, by the equation (18) to calculate S_{usm}^{new} , expressed as follows:

$$S_{usm}^{new}=L_{urm}S_{usm}^{original}L_{urm}^T \tag{18}$$

2) By the formula (19) to iterative calculation S_{usm}^{new} , until convergence, expressed as follows:

$$S_{usm}^{n}=L_{urm}S_{usm}^{n-1}L_{urm}^T=L_{urm}^n S_{usm}^0(L_{urm}T)^n \tag{19}$$

3) S_{usm}^{final} is divided into 4×4 sub-matrix, expressed as follows:

$$S_{usm}^{final}=\begin{bmatrix} S_{image} & \S_{i-a} & S_{i-t} & S_{i-s} \\ S_{a-i} & \S_{audio} & S_{a-t} & S_{a-s} \\ S_{t-i} & S_{t-a} & S_{tgxt} & S_{t-s} \\ S_{s-i} & S_{s-a} & S_{s-t} & S_{shat} \end{bmatrix} \tag{20}$$

Among them, S_{shot} represents the similarity relations between different video lenses, to which it need to reduce the dimension.

B. Dimensionality reduction

LPP is an effective combination of linear and non-linear dimensionality reduction method. It can not only solve the problem of surface data, but can be extended to the outside of the training set. Its basic idea is that the original high-dimensional feature vectors map to low-dimensional semantic subspace by transformation matrix. Therefore, this paper adopts the LPP to reduce the dimension on S_{shot} calculated by the equation (20).

Assume that X represents the original high-dimensional space matrix, namely $X = [x_1, x_2, \cdots, x_n]$, among them, x_i represents $i-th$ lens of video sequence; Y represents the low-dimensional subspace matrix, namely $Y = [y_1, y_2, \cdots, y_n]$, in this way, y_i representative of the corresponding high-dimensional vector x_i, expressed as $y_i = A^T x_i$, among them A represents the transformation matrix.

The main steps of LPP are as follows:

1) To establish adjacent graph for X, the vertices of adjacent graph represent video lens, S_{shot} represents the edge-weight.

2) Eigenvectors and eigenvalues are calculated as follows:

$$XLX^T a = \lambda XDX^T a \tag{21}$$

Among them,$L - D - S_{shot}$, D represents the diagonal matrix, the diagonal element values of D represent the sum of the corresponding column element values of S_{shot} , that is, $D_{ij} = \sum_j S_{shot}$.

According to the corresponding eigenvalues $\lambda_0 \leq \cdots \leq \lambda_{l-1}$, it can be sorted the vector $(a_0, a_1, \cdots, a_{l-1})$ calculated by the equation (21), the result is as follows:

$$x_i \to y_i = A_T x_i,\ A = (a_0, a_1, \cdots, a_{l-1}) \tag{22}$$

Among them, y_i represents a one-dimensional column vector, A represents a $n \times lCdimensional$ transformation matrix.

In this way, the original high-dimensional feature vectors can be mapped to low-dimensional semantic subspace, in order to explore for the intrinsic characteristics of the original data.

4 VIDEO SCENE SEGMENTATION ALGORITHM BASED ON SEMANTIC CONCEPT

Low-dimensional semantic space coordinates y_i obtained by LPP is as the input into the SVM so as to construct a number of different semantic concept training classifier and predict the semantic concept vectors of video key frames, by means of semantic overlapped shot linked algorithm for video scene segmentation at the end of this paper.

A. Construction and training of SVM

SVM implementation is based on LIBSVM 2.88 experimental tool, using radial basis function (RBF), expressed as $K(x_i, x_j) = exp(-g|x_i - x_j|^2), g > 0$.

The major objects of video lens key frames are artificially divided into k classes, corresponding to k-types semantic concept, namely $C_1 - C_2$. If $k > 2$, it can carry out the dual classification; if , it can do multi-classification, that is, it takes one-to-one to construct $k(k-1)/2$ training classifiers. A semantic concept

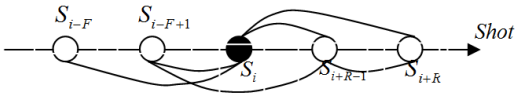

Fig. 1: Semantic overlapped shot linked construction diagram

corresponds to a training classifier so as to construct the different training data sets, and each training set uses the same feature data to train the classifier. If the semantic concept of training data sets appears in the video key frame, the video key frame number is set to 1, otherwise 0. By the cross-contrast training to complete the construction of SVM, it can get the optimal error penalty parameter C and kernel parameter g. Semantic concept "airplane", for example, the main steps of training semantic concept extraction model are as follows:

1) Save the extracted video key frames feature.

2) Select the RBF.

3) Optimal parameter selection. In experiment, this paper selects the gridregression.py tool to conduct k-fold cross-validation in order to obtain the optimal error penalty parameter C and kernel parameter g, in this way, it can effectively avoid over-fitting caused by machine learning, and reduce the model dependence of specific samples.

4) Use the training tool svm-train.exe to train the training data sets in order to obtain the training model file airplane.model.

5) Use the predictive tool svm-predict.exe to predict the test data sets.

B. Semantic concept classification

This paper uses the trained $k(k-1)/2$ training classifiers to classify the video key frames in order to get the semantic concept of each video key frame. Using the 1 and 0 represent respectively if j-th video key frame contains i-th semantic concept, among them, 1 represents that the video key frame contains certain semantic concept, 0 means do not contain the semantic concept. Assuming that the classified results of the training classifiers as shown in the equation (23), the semantic concept vector of 1-th video key frame is $V_{s1} = [0,1,0,0,1,0,1]$

$$\begin{cases} 0\ 1\ 0\ 0\ 0\ 1\ 0\ 0\ 0 \\ 1\ 0\ 0\ 0\ 1\ 1\ 1\ 0\ 0 \\ 0\ 1\ 1\ 0\ 0\ 0\ 0\ 0\ 0 \\ 0\ 0\ 1\ 0\ 0\ 1\ 0\ 0\ 1 \\ 1\ 1\ 0\ 0\ 0\ 0\ 1\ 0\ 1 \\ 0\ 0\ 1\ 0\ 0\ 0\ 0\ 0\ 1 \\ 1\ 0\ 1\ 0\ 1\ 0\ 1\ 0\ 1 \end{cases} \tag{23}$$

C. Semantic overlapped shot linked algorithm

This paper uses semantic overlapped shot linked algorithm to get the similarity between video lens. Its basic idea is that if between video lens exist the same semantic concept, they are similar in the semantics, i.e. $sim(S_i, S_j) = 1$, the two video lens is considered to be in

the same video scene, otherwise dissimilar, i.e. $sim(S_i,S_j)=0$, it means that the video content has been changes. Fig.1 describes the clustering process of video lens, and the specific steps of video scene segmentation algorithm are as follows:

Step1 Select the key frames of video lens. This paper see the first frame and the last frame of each video lens as the video key frame, then all video lenses and key frames are numbered, i.e. n_i , n_j,(n_i ,n_j are positive integers).

Step2 Calculate the semantic concept vectors of video lens. This paper uses RBF and SVM to classify the video key frames so as to obtain the semantic concept vectors of each video lens, namely V_s.

Step3 Define the variable. Assuming that N represents the total lens number of video sequence, S_i represents the current video lens ($1 \le i \le N$, i initialized to 1), F and R represent respectively the video lens number of forward search and backward search (F, R are positive integers,$F \ge R$), f and r represent the actual video lens number of forward search and backward search, represents the video lens number from the current video lens to the start video lens, $Futshot$ represents the video lens number from the current video lens to the last video lens, $sim(S_i,S_j)$ represents the semantic similarity between video lens.

Input: N,V_s,F,R

Output: video scene boundary n_i (video lens number)

Step4 Forward comparison. Suppose $f = min(F,Futshot)$, if $f \neq 0$, $sim(S_i,S_{i+m}) = 0, m = 1,2,\cdots,f$, that is, to judge whether f the previous video lens of the current video lens and S_i have the same semantic concept, if it find the same semantic concept between them, the video lens and S_i belong to the same video scene,$S_i = S_{t+m}$, update $Preshot$ and $Futshot$, return Setp4. Step5 Backward comparison. Suppose $r = min(R,Preshot)$, if $r \neq 0$, $sim(S_i,S_{i-r+m}) = 0, r, i-r+m \neq i, m = 1,2,\cdots,$, that is, to judge whether the following video lens of the current video lens and S_i have the same semantic concept, if it find the same semantic concept between them, the video lens and belong to the same video scene, $S_i = S_{i-t+m}$, update $Preshot$ and $Futshot$, skip Step4.

Step6 Output the video scene boundary. Record the segmented video scene boundary (start and end numbers), update $Preshot$ and $Futshot$. If $Futshot \neq 0$, skip Step4, then to detect the next video scene, otherwise the end of video scene segmentation.

5 ANALYSIS OF EXPERIMENTAL RESULTS

To validate the proposed algorithm for video semantic concept detection performance, this paper selected part of the video data provided by TRECVID, a total of 100 video sequences, 16875 video lenses, 32134 video key frames, which contain mainly news video and a small amount of advertisements, sports events, TV shows. Fig.2 shows part of the video key frame images.

Fig. 2: Video key frame images

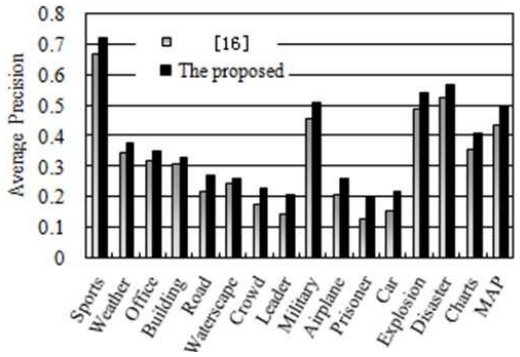

Fig. 3: The detection results of different semantic concepts

In the experiment, this paper selected the 10000 images in front of video key frames as the training set, the rest as the test set, and selected the 15 common semantic concepts in news video, that is, Sports, Weather, Office, Building, Road, Waterscape, Crowd, Government-Leader, Military, Airplane, Prisoner, Car, Explosion, Disaster, Charts. Fig.3 shows the detection results of different semantic concepts, among them, the ordinate represents average precision (AP), MAP is the average value of AP, expressed as follows:

$$AP = \frac{1}{R}\sum_{k=1}^{S}\frac{R_k}{k} \times I_k \tag{24}$$

Among them, S represents the total video lens number of the test set, R represents the associated video lens number of certain semantic concept, R_k represents the associated video lens number of the first k test lenses. If k-th video lens and the test lens are relevant, $I_k = I$, otherwise $I_k = 0$.

As can be seen from Fig.3, the proposed algorithm can better detect different semantic concepts in video lens. In addition, the detection results of different semantic concepts are different, which is mainly depend on the complexity and mutual influence of between

different semantic concepts, such as the AP value of the more complex semantic concept ”Prisoner” is 20%, which is far less than the simple semantic concepts ”Sports” 72%, ”Military” 51%, ”Explosion” 54%, ”Disaster” 57%, not only that, ”Prisoner” and ”Crowd”, ”Government-Leader” are easily confused. Besides, the semantic concept number in video key frames affects directly the detection results, such as compare No.9, No.10, No.504 video key frames (including a semantic concept) with No.11, No.15, No.1013 video key frames (including a variety of semantic concepts), the former semantic concept detection rate is higher. Compared with the literature [16], the video semantic concept detection algorithm proposed by this paper is more effective, MAP value higher, about 50%, mainly due to the proposed algorithm selected multiple key frames in the same video lens, but the literature [16] selected only a key frame. Yet, a key frame is not enough to express the entire contents of the video lens, easily leading to the semantic concepts missed. Furthermore, video of shooting, editing and post-processing have different effects in the semantic concept detection results. It can be seen, how to establish a standard video ontology library and select the appropriate video key frames is need further research in the field of video retrieval.

In the experiment, the video scene is divided into indoor scene, outdoor scene, nature scene and man-made scene. Among them, the outdoor scene contains some typical scenarios, such as city, building, forest, desert, sea etc; the indoor scene contains some typical scenarios, such as office, bedroom, kitchen, living room etc. And, the typical scenarios are used as the basis for dividing the video scene. This paper selected three types of video sequences, respectively, BBC News, Hollywood movies, Yahoo! Sports, total 2:30’50”, 2329 video key frames, 1219 video lenses, 65 video scenes. To validate the proposed algorithm for video scene segmentation performance, this paper adopted recall, precision and M widely used in the field of information retrieval, defined as follows:

$$Recall = \frac{n_c}{n_c + n_m} \times 100\% \tag{25}$$

$$Precision = \frac{n_c}{n_c + n_f} \times 100\% \tag{26}$$

$$M = \frac{2 \times Preision \times Recall}{\times Preision + Recall} \times 100\% \tag{27}$$

Among them, n_c represents the correct detected scene number, n_m represents the missed scene number, n_f represents the false detected scene number. Tab.1 shows the different results to compare with the classical algorithm STG and the proposed algorithm in order to illustrate the performance of different algorithms.

As can be seen from Tab.1, compare with the STG algorithm proposed by the literature [2], the video scene segmentation effect is better, and recall, precision, M value are higher in this paper, mainly due to the STG algorithm does not consider TAC character between multimodal media data. However, the proposed algorithm, taking full account of the TAC character, adopted the SimFusion algorithm for multimodal subspace correlation transitive to get the similarity relation between video lens, in order to better reduce the ”semantic gap” between the low-level feature and high-level semantic in video. In addition, F, R values have a direct impact on the video scene segmentation, such as F, R take 3, the recall, precision, M value of the proposed algorithm are respectively 72.3%, 85.5%, 78.3%, then the STG algorithm are respectively 63.0%, 82.0%, 71.2%; F, R take 4, the recall, precision, M value of the proposed algorithm are respectively 81.5%, 85.5%, 83.4%, and M reaches its maximum value, then the STG algorithm are respectively 64.6%, 79.2%, 71.1%. The main reason is that the video sequence contains multiple lenses, if the selected search range is too small, it can lead to segment falsely the video lenses (expressing the same video content), otherwise it can lead to cluster falsely the video lenses (expressing the different video content). It can be seen, it will be the next stage of research to seek appropriate search range.

6 CONCLUTIONS

Multi-modality video scene segmentation algorithm with semantic concept proposed in this paper, taking full account of TAC character between multimodal media data, is the use of SimFusion algorithm to calculate the similarity relations between video lenses so that it can effectively solve the non-linear problem caused by multimodal fusion. In addition, it can also effectively reduce the curse of dimensionality by LPP, which can be a high-dimensional feature vectors map to low-dimensional semantic subspace by reducing the dimensionality of feature vectors (reflecting the similarity relations between video lenses). Then, semantic space coordinates obtained by reducing the dimensionality is as the input into the SVM so as to construct a number of different semantic concept training classifier and predict the semantic concept vectors of video key frames, by means of semantic overlapped shot linked algorithm to get the video scene at the end of this paper. The authors experiments show that this paper presents an effective method for video scene segmentation.

ACKNOWLEDGEMENT

The work is supported by Doctor Fund of Wuhan Institute of Technology (201210304007).

References

[1] Sidiropoulos P., Mezaris V., Kompatsiaris I., Meinedo H., Bugalho M., Trancoso I.. Temporal video segmentation to scenes using high-level audiovisual features [J]. IEEE Trans. on Circuits and Systems for Video Technology, **21**,1163-1177 (2011).

[2] Yeung M, Yeo B. Segmentation of video by clustering and graph analysis [J]. Computer Vision and Image Understanding, **71**, 94-109 (1998).

Table 1: The performance comparison results of different algorithms

Value table	Algorithm	Index Level	Actual scene	Segmented scene	Correct scene	Recall%	Precision%	M%
F=3	The proposed	2329	65	55	47	72.3	85.5	78.3
R=3	STG	2329	65	50	41	63.0	82.0	71.2
F=4	The proposed	2329	65	62	53	81.5	85.5	83.4
R=4	STG	2329	65	53	42	64.6	79.2	71.1
F=3	The proposed	2329	65	56	48	73.8	85.7	79.3
R=4	STG	2329	65	55	43	66.2	78.2	71.7
F=4	The proposed	2329	65	59	50	76.9	84.7	80.6
R=3	STG	2329	65	58	44	67.7	75.9	71.5

[3] Yun Zhai, Shah M. Video scene segmentation using Markov chain Monte Carlo [J]. IEEE Trans. on Multimedia, ,**8**, 686-697(2006).
[4] Jianrong Cao. An algorithm of video scene segmentation based on semantics [J]. Journal of Image and Graphics, **11**, 1657-1660, (2006).(in Chinese)
[5] Chih-Chung Chang, Chih-Jen Lin. LIBSVM: A library for support vector machines [J]. ACM Trans. on Intelligent Systems and Technology, **2**, 322-324(2011).
[6] Shifei Ding, Bingjuan Qi, Hongyan Tan. An overview on theory and algorithm of support vector machines [J]. Journal of University of Electronic Science and Technology of China, **40**, 2-10 (2011).(in Chinese).
[7] T.H. Hsu and S.H. Nian, Journal of the Chinese Institute of Transportation, **10**, 79-96. (1997).
[8] Panagiotis Sidiropoulos, Vasileios Mezaris, etc. Temporal video segmentation to scenes using high-level audiovisual features [J]. IEEE Trans. on Circuits and Systems for Video Technology, **21**, 1163-1177 (2011).
[9] Xi, W., Fox E.A.,et al. SimFusion: Measuring similarity using unified relationship matrix [C]. Proc. of the 28th Annual Int'l ACM SIGIR Conf. on Research and Development in Information Reteieval. New York,USA:ACM Press, 130-137 (2005).
[10] Xiaofei He, Niyogi P. Locality preserving projections [C]. Advances in Neural Information Processing Systems Conference, Cambridge: MIT Press, 153-160 (2003).
[11] Babaguchi N, Kawai Y, Kitahashi T. Event based indexing of broadcast sports video by intermodal collaboration [J]. IEEE Trans. on Multimedia, **4**, 68-75 (2004).
[12] Li Liu, Gangyao Kuang. Overview of image textural feature extraction methods [J]. Journal of Image and Graphics, **14**, 622-635(2009). (in Chinese)
[13] Won C.S., Park D.K., Park S.J. Efficient use of mpeg-7 edge histogram descriptor [J]. ETRI J., **24**, 23-30 (2002).
[14] Yanlan Liu, Fei Wu. Video semantic concepet detection using multi-modality subspace correlation propagation [C]. Proc. of the 13th Int'l Multimedia Modeling Conference, Berlin, Germany:Springer, 527-534 (2007).
[15] Hong Zhang, Fei Wu, Yueting Zhuang, Jianxun Chen. Cross-Media Retrieval method based on content correlations [J]. Chinese Journal of Computers, **31**, 820-826 (2008).(in Chinese).
[16] Fei Wu, Yanan Liu, Yueting Zhuang. Tensor-based transdutive learning for multi-modality video semantic concepet detection. IEEE Trans. on Multimedia, **11**, 868-878 (2009).

Shaofei Wu is the teacher of School of Computer Science and Engineering, Wuhan Institute of Technology, Wuhan, Hubei, P.R. China. He received the PhD degree in Management Science and Engineering at Huazhong University of Science and Technology. His research interests is in the area of Management Information System.

Maozhu Jin is an instructor of Business School,the tutor of MBA operations management and innovation and entrepreneurship management in Sichuan University. He has been engaged in the teaching of core curriculums such as operations management and management consulting. His current research interests include the areas of operations management, organizational process reengineering, strategic management, service operations management, platform-based mass customization and risk management. He has published two books and over ten research papers in authoritative journals of high quality both at home and abroad, and ten of them are retrieved by SCI and EI.