21st International Conference on Pattern Recognition (ICPR 2012)
November 11-15, 2012. Tsukuba, Japan

Multi-modality Movie Scene Detection Using Kernel Canonical Correlation Analysis

Guangyu Gao, Huadong Ma
*Beijing Key Laboratory of Intelligent Telecommunications Software and Multimedia,
Beijing University of Posts and Telecommunications, Beijing 100876, China,
ggyalzj@gmail.com, mhd@bupt.edu.cn

Abstract

Scene detection is the fundamental step for efficient accessing and browsing videos. In this paper, we propose to segment movie into scenes which utilizes fused visual and audio features. The movie is first segmented into shots by an accelerating algorithm, and the key frames are extracted later. While feature movies are often filmed in open and dynamic environments using moving cameras and have continuously changing contents, we focus on the association extraction of visual and audio features. Then, based on the Kernel Canonical Correlation Analysis (KCCA), all these features are fused for scene detection. Finally, spatial-temporal coherent shots construct the similarity graph which is partitioned to generate the scene boundaries. We conduct extensive experiments on several movies, and the results show that our approach can efficiently detect the scene boundaries with a satisfactory performance.

1. Introduction

With rapid advances in digital technologies, feature movies take a large portion in the powerful growth videos. In order to feasibly browse and index these movies, movie scene detection is the critical step. Generally, scenes are defined as sequences of related shots chosen according to certain semantic rules. Shots belonging to one scene are often taken with a fixed physical setting as well as that the continuity of ongoing actions performed by the actors are also seemed as a scene.

Over the last decades, many scene detection methods have been proposed. Tavanapong [7] introduced a stricter scene definition for narrative films and visual features from selected local regions of key frames are extracted. Then features were compared by the continuity-editing techniques for film making. Besides, Chasanis et al. [1] clustered shots into groups, and then a sequence alignment algorithm was applied to detect scenes when the pattern of shot labels changed. In [6], a weighted undirected graph called Shot Similarity Graph (SSG) was constructed, then scene detection was transformed into a graph partitioning problem.

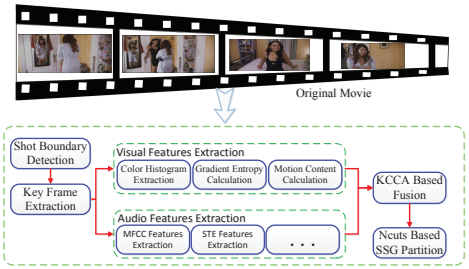

Figure 1. Flowchart of the proposed scene detection scheme.

However, most of the above mentioned methods merely exploit visual information, mostly the color features. There were also several multimodality features based methods. Rasheed et al. [5] have incorporated the shot length and motion contents to analyze scene properties. Zhu et al.[9] analyzed both the auditory and the visual sources to semantically identify video scenes. In [4], an enhanced set of eigen-audio frames was created and visual information was used to align audio scene change indications with neighboring video shot changes. In these mentioned multimodality scene

*The work reported in this paper is supported by the National Science Fund for Distinguished Young Scholars under Grant No. 60925010, the National Natural Science Foundation of China under Grant No. 60833009, the Foundation for Innovative Research Groups of the National Natural Science Foundation of China under Grant No. 61121001, the Program for Changjiang Scholars and Innovative Research Team in University under Grant No. IRT1049, the Co-sponsored Project of Beijing Committee of Education.

detection methods, a non-trivial issue is how the audio information is integrated with the visual information. While KCCA is very useful to improve the performance of multimodality recognition systems which involve modalities with a mixture of correlated and uncorrelated components, we proposed to integrate both features of colors, motion contents, edges and audio features using KCCA fusion mechanism for scene detection, as illustrated in Fig. 1

The rest of the paper is organized as follows. The proposed approach is presented in Section 2, such as details of feature extraction and fusion. The experimental results are illustrated in Section 3. Finally, we conclude our work in Section 4.

2. Proposed Approach

2.1. Shot boundary detection

Firstly, an accelerating shot boundary detection method of our previous work [2] is adopted to segment movie into shots efficiently. Concretely, for each frame, the more the pixels near the center of the frame, the more they are important. Thus the Focus Region (FR) in each frame is defined. Further more, by using a skipping interval of 40 frames, it not only speeds up the detection speed, but also finds more gradual transitions. Besides, the camera and object motions are detected as the candidate shot boundaries. Thus, the corner distribution analysis is adopted to exclude them as false boundaries. Whereas, these camera or object motions are also used in key frames extraction.

2.2. Key frame extraction

In order to choose an appropriate number of key frames, widely used middle frames may represent a static shot with little actor or camera motion well. However, dynamic shots with higher actor or camera motion may not be represented adequately. Thus a variant of key frame selection [6] is used. To initialize the key frames set K, besides middle frames, frames with step of 5 frames in those camera or object motions are also contained. Actually, not only the color histogram, but also mutual information is used to define the frame similarity. The frame similarity which takes account the mutual relation of correspond pixels in two frames is defined as follows,

$$D_{i,i+s} = \frac{min(I_{i,i+1}, I_{i,i+s})}{max(I_{i,i+1}, I_{i,i+s})} \tag{1}$$

where $I_{i,i+1}$ means the mutual information of frames F_i and F_{i+1}, and the same meaning of $I_{i,i+s}$[2].

2.3. Audio-Visual features extraction

Descriptions in terms of hue/lightness/chroma or hue/lightness/saturation are often more relevant, so a 16 bin HSV normalized color histogram is computed for each frame with 8 bins for Hue and 4 bins each for Saturation and Value. However, two different frames may have the same color histogram, so another visual feature named gradient entropy is introduced. First, the gradient and gradient magnitude ∇f are computed. Then, we create the orientation histogram with 180 bins, and each bińs value is calculated by a weighted voting of the gradient magnitude. At last, we get the gradient entropy GE in terms of the orientation histogram $\mathcal{O}$ as:

$$GE = -\sum_{i=1}^{180} \mathcal{O}(i) \log \mathcal{O}(i). \tag{2}$$

Actually, we get three gradient entropies: $\mathcal{O}_r$, $\mathcal{O}_g$ and $\mathcal{O}_b$, in R, G and B channel respectively.

In addition, a global affine motion model is estimated and the velocities of blocks are reprojected [5]. Shot motion content is defined as the magnitude of the difference between the actual and the reprojected velocities.

Generally, while an action scene consists of a successive shots refer to different physical setting, the audio features will be very important. However, there are many audio frames in a shot, which are not so strictly corresponding to video frames. In order to extract audio features which align to key frames, we select ten audio frames, five in forward and five in backward at the time point of each key frame. For each audio frame, we consider a 43-dimensional audio features comprising Mel-frequency cepstral coefficients (MFCCs) and its delta values and acceleration values (36 features), mean and variance of short time energy log measure (STE) (2 features), mean and variance of short-time zero-crossing rate (ZCR) (2 features), short-time fundamental frequency (or Pitch) (1 feature), mean of the spectrum flux (SF) (1 feature), and harmonic degree (HD) (1 feature) [8]. Finally, the mean features of the ten audio frames were calculated for the corresponding key frame.

2.4. KCCA based feature fusion

CCA can be seen as solving the problem of finding basis vectors for two sets of variables such that the correlation between the projections of the variables onto these basis vectors is mutually maximized[3]. Given two centered random multivariables $x\epsilon\ \mathcal{R}^{n_x}$ and $y\epsilon\ \mathcal{R}^{n_y}$, a new coordinate for x is defined by choosing a direction w_x and projecting x onto that direction, $x \rightarrow \langle w_x, x\rangle$ and the same for y, $y \rightarrow \langle w_y, y\rangle$. Let

$$\mathcal{S}_{x,w_x} = (\langle w_x, x_1\rangle, \ldots, \langle w_x, x_n\rangle), \tag{3}$$

$$\mathcal{S}_{y,w_y} = (\langle w_y, y_1\rangle, \ldots, \langle w_y, y_n\rangle), \quad (4)$$

The first stage of canonical correlation is to choose w_x and w_y to maximize the correlation between the two variables. Namely, the function result to be maximized is

$$\rho = \max_{w_x, w_y} \frac{w_x^{'} C_{xy} w_y}{\sqrt{w_x^{'} C_{xx} w_x w_x^{'} C_{yy} w_y}} \quad (5)$$

where,

$$C(x,y) = E[\begin{pmatrix} x \\ y \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}^{'}] = \begin{bmatrix} C_{xx} & C_{xy} \\ C_{yx} & C_{yy} \end{bmatrix} = C. \quad (6)$$

However, CCA may not extract useful descriptors of the data because of its linearity. KCCA who is the powerful nonlinear extension of CCA to correlate the relation between two multidimensional variables, offers an alternative solution by first projecting the data into a higher-dimensional feature space.

$$K(x,y) = \langle \phi(x) \cdot \phi(y) \rangle, \quad (7)$$

The maximum canonical correlation is the maximum of ρ with respect to w_x and w_y. We get the two transformations corresponding to these canonical basis vectors, $w_x = (\alpha_1, \alpha_1, \ldots, \alpha_m), w_y = (\beta_1, \beta_1, \ldots, \beta_m)$[3].

In our approach, while x and y refer to the visual and audio features respectively, the resulting combined audio-visual feature vector is thus given by

$$\mathcal{Z} = [X \quad Y] = \begin{pmatrix} w_x \\ w_y \end{pmatrix}^T \begin{pmatrix} x \\ y \end{pmatrix} \quad (8)$$

Hence, using the KCCA fusion algorithm, the 63-dimension features (20 for visual features and 43 for audio features) in subsection 2.3 are transferred to the new fused feature vectors for scene detection.

2.5. Scene segmentation based on graph cuts

Graph partitioning techniques are widely used for scene detection with the construction of a graph in which each node represents a shot, and the edges are weighted by their similarity or affinity. In our approach, the similarity between shots s_i and s_j is defined as,

$$ShotSim(s_i, s_j) = \max_{p \epsilon K_i, q \epsilon K_j} (AvSim(p,q)) \quad (9)$$

where p and q are the key frames in key frames set K_i and K_j respectively. Meanwhile, $AvSim(p,q)$ is the Euclidean distance of the fused audio-visual feature vectors in two key frames.

Table 1. Summary of our test data set

Movies	Duration(s)	#Frames	#Shot	#Scenes
L.F.H.	3324	83100	1308	35
C.A.	2353	58825	942	28
G.L.S.	2835	70875	1126	30
P.F.	3715	92875	1611	41

A weighted undirected graph $G = (V, E)$ is constructed with all the shots, where the node set V denote the shots and the weight of edge in set E denotes the node distance. Therefor, scene detection means to seek the optimal partition $V_1, V_2, \ldots, V_M$ of V, which maximizes the similarity among the nodes of each sub-graph (V_i, E_i) and minimizes the across similarity between any two sub-graphs. In this paper, the Normalized graph Cuts (NCuts) is employed to partition the shot similarity graph. The optimal disassociation is the minimum cut cost as a fraction of the total edge connections to all the nodes in the graph called NCuts,

$$Ncut(V_1, V_2) = \frac{cut(V_1, V_2)}{assoc(V_1, V)} + \frac{cut(V_1, V_2)}{assoc(V_2, V)} \quad (10)$$

And

$$cut(V_1, V_2) = \sum_{\nu_1 \epsilon V_1, \nu_2 \epsilon V_2} w(\nu_1, \nu_2) \quad (11)$$

$$assoc(V_1, V) = \sum_{\nu_1 \epsilon V_1, \nu \epsilon V} w(\nu_1, \nu) \quad (12)$$

where w is the node distance between ν_1 and ν_2,

$$w(\nu_i, \nu_j) = exp(-\frac{(m_i - m_j)^2}{N^{\frac{1}{2}} \sigma^2}) \times ShotSim(\nu_i, \nu_j) \quad (13)$$

where σ is the standard deviation of shot duration in the entire video; m_i and m_j are the middle frame number of shots ν_i and ν_j; N is the shots number. The detail of the NCuts based graph partition can be found in [6][10].

3 Experimental Results

To evaluate the performance of our approach, we have experimented with four movies including three Hollywood movies: *Love at First Hiccup (L.F.H.), City of Angels (C.A.), The Goods: Live Hard, Sell Hard (G.L.S.)* and a Chinese movie: *The Piano in a Factory (P.F.)*, as seen in Table 1.

In order to evaluate different approaches with ours, we used $Recall$ and $Precision$ measure as well as their combination named $F - Measure$:

$$FM = \frac{2 \times Recall \times Precision}{Recall + Precision}. \quad (14)$$

Table 2. Comparison of our approach with Rasheed's[6] and Marios'[4]

Movie Name	*L.F.H.*	*C.A.*	*G.L.S.*	*P.F.*	*Total*
Ground Truth	35	28	30	41	134
Proposed Approach					
Detected	38	30	31	44	143
Recall (%)	83.0	85.7	83.3	87.8	**85.1**
Precision (%)	76.3	80.0	80.6	82.0	**80.0**
FM (%)	79.5	82.8	81.9	84.8	**82.5**
Method of Rasheed [6]					
Detected	45	39	33	53	170
Recall (%)	68.6	75.0	66.7	80.5	73.1
Precision (%)	53.3	54.0	60.6	62.3	57.6
FM (%)	60.0	62.8	63.5	70.2	64.4
Method of Marios [4]					
Detected	39	33	27	49	148
Recall (%)	80.0	82.1	66.7	80.5	77.6
Precision (%)	71.8	69.7	74.1	67.3	70.3
FM (%)	75.7	75.4	70.2	73.3	73.8

In our experiment, the four test movies are quickly and accurately segmented into shots and the key frames are also extracted firstly. Then, the visual features of color histogram, gradient entropy and the motion contents, as well as the 43-dimension audio features are extracted for feature fusion. Actually, in order to fuse all these features using the KCCA algorithm, the Gaussian Kernel ($K(z, z_i) = exp(-||z - z_i||^2/2\sigma^2)$ with $\sigma = 1e6$ is used. The two transformation parameters (w_x and w_y) are pre-trained. We have chose 80 scenes which contain 4126 shots from the Hollywood movies. All these scenes are manually segmented for training. Finally, the NCut based shot similarity graph partition is applied to detect all the scenes in test movies.

In order to evaluate the result of our KCCA based multimodality scene detection approach, Table 2 lists the comparison results of our method with that of Zeeshan Rasheed et al. [6] and Vasileios T. Chasanis et al. [4]. Method of [6] merely exploited unimodality of visual information, so the detection performance is not so satisfactory compared with the multimodality based methods. Method of [4] aligned audio scene change indications with neighboring video shot changes, while our approach takes account of visual features of color and gradient, motion contents as well as several audio features for KCCA based fusion. Although [4] can deal with news video well, for movie contents, the results of our approach can be seen to be more robust and promising, as shown in Table 2.

4 Conclusions

It is still a challenging problem to robustly detect diverse movie scene changes. In this paper we address this problem using KCCA to fuse multimodality features in feature movies. Firstly, we extracted the multimodality features including visual color histogram, gradient entropy, motion contents and several audio features respectively. Then, we take use of KCCA to fuse all the extracted audio-visual features. Finally, the popularly used SSG partition method based on Ncuts was introduced to detect the scene boundary, and the experimental results show that our approach have archived a satisfactory accuracy.

References

[1] V. T. Chasanis, A. C. Likas, and N. P. Galatsanos. Scene detection in videos using shot clustering and sequence alignment. *IEEE Trans. on Multimedia*, 11(1):89–100, 2009.

[2] G. Gao and H. Ma. Accelerating shot boundary detection by reducing spatial and temporal redundant information. In *Proc. of IEEE Inter. Conf. on Multimedia and Expo.*, pages 1–6, 2011.

[3] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural Computation*, 16(12):2639–2664, 2004.

[4] M. Kyperountas, C. Kotropoulos, and I. Pitas. Enhanced eigen-audioframes for audiovisual scene change detection. *IEEE Trans. on Multimedia*, 9(4):785–797, 2007.

[5] Z. Rasheed and M. Shah. Scene detection in hollywood movies and tv shows. In *Proc. of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, pages 343–348, 2003.

[6] Z. Rasheed and M. Shah. Detection and representation of scenes in videos. *IEEE Trans. on Multimedia*, 7(6):1097–1105, 2005.

[7] W. Tavanapong and J. Zhou. Shot clustering techniques for story browsing. *IEEE Trans. on Multimedia*, 6(4):517–527, 2004.

[8] J. Wang, L. Duan, H. Lu, J. S. Jin, and C. Xu. A mid-level scene change representation via audiovisual alignment. In *Proc. of IEEE Inter. Conf. on Acoustics, Speech and Signal Processing*, volume 2, pages 409–412, 2006.

[9] D. Z. Yingying Zhu. Scene change detection based on audio and video content analysis. In *Proc. of Inter. Conf. on Computational Intelligence and Multimedia Applications*, pages 229–234, 2003.

[10] Y. Zhao, T. Wang, P. Wang, W. Hu, Y. Du, Y. Zhang, and G. Xu. Scene segmentation and categorization using ncuts. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, pages 1–7, 2007.