A Novel Role-Based Movie Scene Segmentation Method

Chao Liang, Yifan Zhang, Jian Cheng, Changsheng Xu, and Hanqing Lu

National Lab of Pattern Recognition, Institute of Automation,
Chinese Academy of Sciences, Beijing, China 100190
{cliang,yfzhang,jcheng,csxu,luhq}@nlpr.ia.ac.cn

Abstract. Semantic scene segmentation is a crucial step in movie video analysis and extensive research efforts have been devoted to this area. However, previous methods are heavily relying on video content itself, which are lack of objective evaluation criterion and necessary semantic link due to the semantic gap. In this paper, we propose a novel role-based approach for movie scene segmentation using script. Script is a text description of movie content that contains the scene structure information and related character names, which can be regarded as an objective evaluation criterion and useful external reference. The main novelty of our approach is that we convert the movie scene segmentation into a movie-script alignment problem and propose a HMM alignment algorithm to map the script scene structure to the movie content. The promising results obtained from three Hollywood movies demonstrate the effectiveness of our proposed approach.

Keywords: Film script, scene segmentation, Hidden Markov Model.

1 Introduction

As digital video data continues to grow, efficient video accessing becomes increasingly important. Structured analysis is an efficient way to make a long data more accessible. Similar to text parsing performed on the granularity of word, sentence and paragraph, video data can also be analyzed in the level of frame, shot and scene. In cinematography, scene is the basic story unit that consists of one or more consecutive shots which are semantically correlated. For semantic-based movie analysis, accurate scene segmentation is an important and indispensable module.

Extensive research efforts have been devoted to movie scene segmentation in recent years. Rasheed *et al.* [1] transformed the scene segmentation into a graph partitioning problem, where each node represents a shot and each edge represents the temporal-visual coherence between two shots. Through recursively applying the normalized cuts algorithm, their method can generate number-prescribed scene segmentation. However, due to the semantic gap, such content-based method is difficult to describe the high-level semantic meaning, which is an intrinsic thread in scene segmentation. To address this difficulty, Weng *et al.* [2] investigated the usage of social relationship in segmenting movie video. By building the roles' social network and analyzing the context variance, their method reported promising result in story-level segmentation. Different from the above the content-based methods, Cour *et al.*

P. Muneesawang et al. (Eds.): PCM 2009, LNCS 5879, pp. 917–922, 2009.

[3] utilized external text information to segment movie into scenes. By aligning the common dialogues (in both script and close caption) and timestamps (in both close caption and movie), their approach can generate script-specified movie scene segmentation. However, because of the wide discrepancies between script and close caption, the text alignment rate is quite limited, which directly affects the final segmentation accuracy.

Motivated by the social network analysis and the usage of film script, we propose a novel role-based approach for movie scene segmentation. The main idea of our method is to map the text scene structure to the video content based on the role network analysis in both movie and script. Specifically, we first build the semantic link between movie and script through face-name matching, then we adopt the EMD distance to measure the semantic (role component) similarity between role histograms in both movie shots and script scenes and finally search for a global optimal alignment between movie and script under the HMM framework. Compared with previous work, the main contributions of this paper are: 1) We convert the movie scene segmentation into a movie-script alignment problem which is more objective and accurate in terms of high-level semantic meaning; 2) We present a bag-of-roles representation to depict the high-level semantics and their similarities between movie shots and script scene description; 3) We propose a HMM-based alignment algorithm to generate the global optimal alignment between movie and script.

2 Role-Based Movie Scene Segmentation

This section presents in detail the role-based movie scene segmentation approach. It includes face-name matching, bag-of-roles representation and HMM-based Movie/ Script Alignment.

2.1 Face-Name Matching

Face-name matching is an important link bridging the movie and script in our method. For space reason, we focus our following discussion on algorithm's basic idea and final derived result (illustrated in Fig. 1). For readers who are interested in implement details, please reference relevant work in [4] and [5].

Motivated by the social network analysis, a similar RoleNet [2] was built on the basis of the co-occurrence status among roles. An intuitive understanding of RoleNet is a weighted graph $\mathbf{G} = \{\mathbf{V}, \mathbf{E}\}$ where $\mathbf{V}$ represents the set of roles and $\mathbf{E}$ represents the set of social link among roles. The more scenes where two roles appear together, the closer the two roles are. In our approach, this RoleNet is realized as a face-net in the movie and a name-net in the script respectively. After that, an inexact graph matching algorithm [5] is applied to find the face-name correspondence by clustering face and name nodes in the dimension-reduced space. The reason of adopting the inexact graph matching is twofold. Firstly, it allows the matching between graphs with unequal node numbers, which loosen the number restriction in face clustering. Secondly, it generates a soft matching where a face cluster can be assigned to various names with different probabilities, which limits the negative effects of error matching.

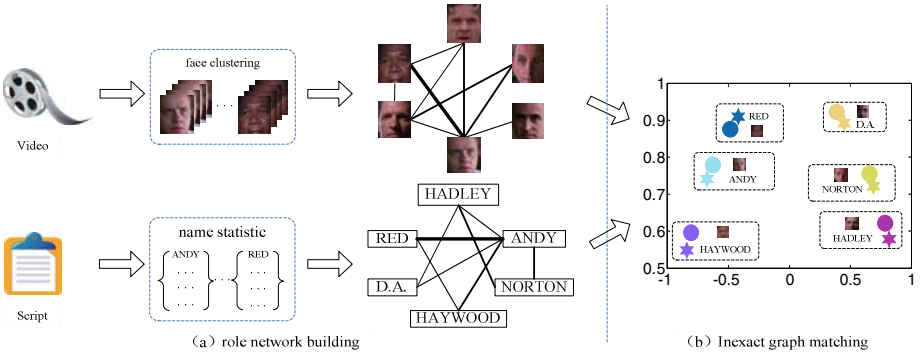

Fig. 1. Face-name matching with two main steps: (a) role network building and (b) inexact graph matching between face-net and name-net

The final result of the face-name matching is a line-normalized matrix **M** with each element $\mathbf{M}_{i,j}$ represent the matching probability between the i^{th} name and j^{th} face:

$$\mathbf{M}_{i,j} = exp\{-\frac{dist^2(name_i - face_j)}{2\sigma^2}\}, \quad with \ \sum_i \mathbf{M}_{i,j} = 1 \tag{1}$$

where $dist(name_i\text{-}face_j)$ represents the euclidean distance between the i^{th} name and j^{th} face in the dimension-reduced space and σ is the tuning parameter for accommodating the deformation between face-net and name-net in dimension reduction process.

2.2 Bag-of-Roles Representation

With the idea of bag-of-word representation in natural language processing, we propose a bag-of-roles representation to denote the semantics of a movie shot or script scene using a set of characters appearing in that segment. An intuitive understanding of such expression is a role histogram where each bin represents a character and its related bin value reflects the character's occurrence frequency in that movie segment.

In our approach, this role histogram corresponds to the face histogram in the movie and name histogram in the script respectively (illustrated in Fig. 2). Based on the generated face-name matching result, the semantic similarity between a movie shot and a script scene description can be measured by the earth mover distance (EMD) between their related face and name histograms.

2.3 HMM-Based Movie/Script Alignment

After the representation stage, the movie video is converted into a shot sequence $V = \{v_1, v_2, \ldots, v_m\}$ where m is the shot number in the sequence and each shot v_i is related with a face histogram. Similarly, the film script is converted into a scene descriptions sequence $D = \{d_1, d_2, \ldots, d_n\}$ where each scene description d_j is represented by a name histogram. The goal of our movie scene segmentation is to assign each movie shot v_i to a specified script scene description d_j so that the script scene structure can be semantic-invariably mapped into movie video.

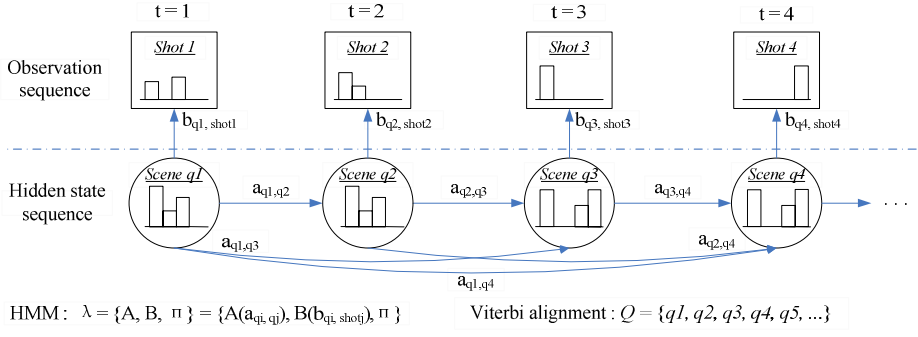

Fig. 2. Role-histogram-based HMM model

This corresponds to finding the optimal assignment sequence $S^* = \{s_1, s_2, \ldots, s_m\}$ which maximizes the *a-posteriori* probability:

$$S^* = \arg\max_{S \in \mathcal{S}} p(V \mid S)\, p(S) \tag{2}$$

where $\mathcal{S}$ is the set of all possible assignment S sequences. If we regard the movie shot sequence V as an observation sequence and the assignment sequence S as a 'hidden' state sequence, then equation 2 can be considered as an observation explanation problem and solved by the Viterbi algorithm under the HMM framework.

The HMM model $\lambda = \{A, B, \Pi\}$ adopted in our approach is illustrated in Fig. 2. The observation sequence is a list of movie shots represented with face histograms and the hidden state sequence is a collection of undetermined script scenes depicted by name histograms. Given the observation sequence, our target is to find the optimal hidden state sequence that can best explain the observation. In this work, we adopt the viterbi algorithm to solve the above problem with following parameter definitions:

- The element of state transition probability distribution $\boldsymbol{A} = \{a_{i,j}\}$ is defined as:

$$\begin{aligned} a_{i,j} &= P\{q_{t+1} = d_j \mid q_t = d_i\} \\ &= \begin{cases} 0, & i > j \\ \dfrac{1}{K}\exp\{-\dfrac{|j-i|^2}{2\sigma^2}\}, & 0 < i \le j \le m, \quad K = \sum_{j=i}^{m}\exp\{-\dfrac{|j-i|^2}{2\sigma^2}\} \end{cases} \end{aligned} \tag{3}$$

where q_t is the hidden state variable at time t, d_i and d_j represent two successive script scenes, K is a normalization parameter to guarantee $\boldsymbol{A}$ is a probability distribution, and σ is the tuning parameter to control the jump distance's effect on transition probability.

- The element of observation probability distribution $\boldsymbol{B} = \{b_{i,j}\}$ is defined as:

$$\begin{aligned} b_{i,j} &= P\{v_j \mid q_t = d_i\} \\ &= \frac{1}{K}\exp\{-\frac{\mathrm{EMD}^2(v_j, d_i)}{2\sigma^2}\}, \quad K = \sum\nolimits_j b_{i,j} \end{aligned} \tag{4}$$

where function $\mathrm{EMD}(v_j, d_i)$ calculate the EMD distance between face histogram in shot v_j and the name histogram in the script scene d_i, K is a normalization parameter to guarantee $\boldsymbol{B}$ is a probability distribution and σ is the tuning parameter to control the effect of EMD distance on the observation probability.

- The initial state distribution $\Pi = \{\pi_i\}$ is designate as 0-1 distribution with the first shot duly belongs to the first script scene.

3 Experiments

In order to verify the proposed method, we conduct the experiments over a corpus of 3 Hollywood movies, which are 'The Shawshank Redemption' (SR), 'You've Got Mail' (YGM) and 'Sleepless in Seattle' (SS), with total video length approximates 5 hours. We first evaluate the face-name matching result by a new search quality measure, then we compare the scene segmentation result between Rasheed method [1] with ours in terms of the commonly adopted *purity* index.

3.1 Face-Name Matching

Motivated by the mean reciprocal rank index in information retrieval, we propose a weighted reciprocal rank (*WRR*) index to depict the matching quality of our soft face-name matching result (given in Table 1). Specifically, the *WRR* is defined as follows:

$$WRR = \sum_{i=1}^{n} \omega_i \frac{1}{rank_i} = \sum_{i=1}^{n} \frac{\#name_i}{\#total\ names} \cdot \frac{1}{rank_i} \tag{5}$$

where n represents the number of face clusters in the matching process, and $rank_i$ corresponds to the rank of the matching probability of ground truth name in the i^{th} column of face-name matching matrix and ω_i denotes the occurrence proportion of the i^{th} ground truth name in the script.

Table 1. Weighted Average Precision on three evaluation movie

Evaluation Movies	SR	YGM	SS
Number of Roles	16	14	24
Number of Face Cluster	14	14	21
Weighted Average Precision (WAP)	0.89	0.90	0.85

3.2 Movie Scene Segmentation

The movie scene segmentation is evaluated based on the 'purity' criteria used in [6]. Given a sequential data, a ground truth segmentation $S = \{(s_1, \Delta t_1), \ldots, (s_g, \Delta t_g)\}$, and an automatic segmentation $S^* = \{(s^*_1, \Delta t^*_1), \ldots, (s^*_a, \Delta t^*_a)\}$, the purity π is defined as

$$\pi = \left(\sum_{i=1}^{g} \frac{\tau(s_i)}{T} \sum_{j=1}^{a} \frac{\tau^2(s_i, s^*_j)}{\tau^2(s_i)}\right) \cdot \left(\sum_{j=1}^{a} \frac{\tau(s^*_j)}{T} \sum_{j=1}^{g} \frac{\tau^2(s_i, s^*_j)}{\tau^2(s^*_j)}\right) \tag{6}$$

Where $\tau(s_i, s^*_j)$ is the length of overlap between the scene segment s_i and s^*_j, $\tau(s_i)$ is the length of the scene s_i, and T is total length of all scene. In each parenthesis, the first term is the fraction of recording a segment accounts for, and the second term is a measure of how much a given segment is split into small fragments. The purity value ranges from 0 to 1, with larger value means that the result is closer to the ground truth.

Table 2 compares our movie scene segmentation approach with Rasheed method [1] in two cases, which are only shots with faces and all shots. In the process of computing the *purity* index, the ground truth scene segmentation is obtained by manual labeling with reference to the film script. As shown in Tab. 2, the average purity of our approach is obviously higher than that of the Rasheed (graphcut-based) method,

Table 2. Comparative movie scene segmentation result

Data Type	Method	Average	SR	YGM	SS
Only shots with faces	Our	83%	85%	87%	78%
	Rasheed	72%	75%	73%	67%
All shots with & without faces	Our	78%	81%	82%	72%
	Rasheed	73%	74%	75%	69%

which can be attributed to the semantic guidance provided by the film script and the goal optimal alignment inferred by the HMM. In addition, since our approach is built on the basis of face-name matching, shots without faces may cause some negative effects to the movie-script alignment (about 5% purity decrease in our experiments). But we would argue that the number of shots without faces is quite limited in the movie (usually less then 5%), hence the above influence is not obvious and hence totally acceptable.

4 Conclusion and Future Work

In this work we have presented a role-based method for movie scene segmentation. Our key idea is to map the script scene structure to the target movie data through movie-script alignment. Based on the correspondence between faces and names, we model the alignment problem as hidden state inference under HMM framework and search the global optimal alignment with the help of Viterbi algorithm. Comparative experiments with state-of-the-art method validate the effectiveness of our proposed approach. In the future, we will research more advanced matching algorithm to improve the scene segmentation precision in our current approach.

Acknowledgement

This work is supported by National Natural Science Foundation of China No. 60833006, Natural Science Foundation of Beijing No. 4072025, and 973 Program Project No. 2010CB327900.

References

1. Rasheed, Z., Shah, M.: Detection and Representation of Scenes in Videos. IEEE Transactions on Multimedia 7, 1097–1105 (2005)
2. Weng, C.Y., Chu, W.T., Wu, J.L.: RoleNet: Movie Analysis from the Perspective of Social Networks. IEEE Transaction on Multimedia 11(2), 256–271 (2009)
3. Cour, T., Jordan, C., Miltsakaki, E., Taskar, B.: Movie/Script: Alignment and Parsing of Video and Text Transcription. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part IV. LNCS, vol. 5305, pp. 158–171. Springer, Heidelberg (2008)
4. Zhang, Y.F., Xu, C.S., Lu, H.Q., Huang, Y.M.: Character Identification in Feature-length Films Using Global Face-Name Matching. In: IEEE-T-MM (to appear)
5. Caelli, T., Kosinov, S.: An Eigenspace Projection Clustering Method for Inexact Graph Matching. IEEE Transaction on Pattern Analysis and Machine Intelligence 26(4), 515–519 (2004)
6. Vinciarelli, A., Favre, S.: Broadcast News Story Segmentation Using Social Network Analysis and Hidden Markov Models. In: Proc. ACM Multimedia, pp. 261–264 (2007)