Synopsis Alignment: Importing External Text Information for Multi-Model Movie Analysis

Li Ming[1,2], Liu Yang[1], Zhang Yong-Dong[1], Lin Shou-Xun[1]

[1]Institute of Computing Technology, CAS, Beijing, 100190, China
[2]Graduate School of the Chinese Academy of Sciences, Beijing, China 100039
{mli, yliu, zhyd, sxlin}@ict.ac.cn

Abstract. Text information, which plays important role in news video concept detection, has been ignored in state-of-the-art movie analysis technology. It is so because movie subtitles are speech of roles which do not directly describe content of movie and contributes little to movie analysis. In this paper, we import collaborative-editing synopsis from professional movie sites for movie analysis, which gives detailed descriptive text of movie. Two aligning methods, subtitle alignment and RoleNet alignment, are proposed complementarily to align synopsis to movie to get scene-level text information of movie. The experiment show that proposed methods can effectively align synopsis to movie scene and the imported text information can give a more user-preferred summarization than merely using audiovisual feature.

Keywords: Movie Analysis, Synopsis, Subtitle Alignment, RoleNet Alignment, Movie Summarization.

1 Introduction

From an incomplete statistic of professional movie site www.imdb.com, movie industry produces more than 20,000 movies every recent year. With the advance of easy dissemination of digital movie, watching movie is becoming one of the most popular entertainments. The explosive amount of movie burdens users in information retrieval. Therefore, effective and efficient data organizing techniques based on movie analysis are in urgent need.

Many studies of movie analysis have been proposed, which can be roughly categorized into following 3 fields based on the content level they study: saliency analysis, emotion calculation and event detection. As a popular entertainment, the main purpose of movie is to attract audiences' attention. Saliency analysis selects frames with high energy or sudden change of audiovisual information which hint high attraction. Different visual and acoustical features as well as so-called movie grammar features are used. Rapantzikos et. al. [1] exploit coupled audiovisual features to get salient key frames for movie summarization. Intensity, color and motion features of vision and amplitude, frequency and instantaneous energy features of audio are used to quantify the importance of movie frames. Video tempo information such as shot change frequency is also used to get salient frames in [2].The movie summarization generated by saliency analysis gives a highlight but less-informative compact show. It is full of highlight frames but we don't know which kinds of frames they are indeed or what happens in these frames. Emotion calculation shows a larruping way in analyzing movie content [3]. Low-level features from cinematographic and psychological considerations estimate emotion category of movie. Comparing to saliency analysis, scene-level affective analysis is a more human-like understanding of movie, which gives human reaction of the audiovisual stimulation. It can be considered as an emotion information complementation to the saliency based movie analysis. Little work has been done about event detection. Ying Li et. al [4][5] detect dialogue event in movie, which is classified to 2-speaker dialog and multi-speaker dialog.

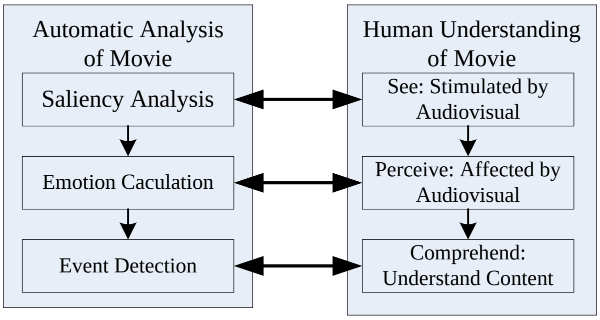
Automatic Analysis of Movie
Saliency Analysis
Emotion Caculation
Event Detection
Human Understanding of Movie
See: Stimulated by Audiovisual
Perceive: Affected by Audiovisual
Comprehend: Understand Content

Figure 1: Comparison between automatic movie analysis and human understanding

Figure 1 shows the 3 research fields of movie analysis comparing to human understanding levels. Although we have elevated automatic movie understanding from human "see" level to "perceive" level, "comprehend" level understanding is still an insurmountable problem. Thus, research on event detection becomes the key point. It is the hardest task among the 3 fields. Broad movie content and various representation methods worsen the semantic-gap problem. Besides, text information in movie, in the form of subtitle, is not as sufficient and detailed as in news, which helps a lot in news video analysis.

Importing external knowledge, especially collaborative editing information on professional movie sites, gives us a new orientation to overcome semantic-gap in movie. Collaborative editing shows great power on current www. Wikipedia [9] is a typical successful case. The great amount of contributors and their sharing spirit makes every problem have its answer and responsible readers monitoring contributions maintains a high accuracy of answers as the knowledge rapidly growing. Professional movie sites attract lots of cinephiles as well as their contributions which are movie information of all kinds including synopsis. Movie synopsis is text paragraph introducing movie story descriptively. On www.imdb.com, it is collaborative editing and the number of movie with synopsis is growing fast while the synopsis maintains high accuracy. We can imagine that synopsis will contribute a lot to event detection in movie analysis as how text acts in news video. As the first step of using synopsis, it should be aligned to movie scene to get a scene-level description of movie. How to align synopsis is what we discuss in this paper.

The main contribution of this paper is as follows: We import synopsis to enhance movie analysis. Two complemental aligning methods, subtitle alignment (section 2.1) and RoleNet alignment (section 2.2), are proposed to get scene-level interpretation of movie. The experiments show that the proposed aligning methods perform well (section 3.1). As illustration, we also show the effect of imported synopsis in movie summarization (section 3.2).

2 Synopsis Alignment

Synopsis alignment aligns synopsis sentences to movie scene. Two aligning methods are proposed complementally to get scene-level text information of movie.

2.1 Subtitle Alignment

Subtitles are mostly the speech texts cinemactors/cinemactresses say in movie. They tell the same story of movie in a different manner from synopsis. Although the information subtitle carries may not be comprehensive, it still has similarity to synopsis for what roles say also leads the story. The alignment between synopsis and subtitle based on the text similarity can import time information to synopsis which will get us scene-level description of movie. Subtitle alignment is described as:

$$C_{M*N} = [c_{ij}]_{M*N} = F(S_{M*N}) = [f(s_{ij})]_{M*N} \quad \textbf{(1)}$$

$$c_{ij} = f(s_{ij}) = \begin{cases} 1 & if \quad s_{ij} > T \\ 0 & otherwise \end{cases} \quad \textbf{(2)}$$

C_{M*N} is subtitle alignment matrix. *M* is scene amount of movie. *N* is sentence amount of synopsis. $c_{ij} = 1$ denotes the *j*th sentence of synopsis describes the *i*th scene of movie. S_{M*N} is subtitle-synopsis similarity matrix. s_{ij} is cosine similarity of *i*th scene of subtitles and *j*th sentence of synopsis. It is defined as:

$$s_{ij} = \cos(\overline{Y}_i, \overline{U}_j) = \frac{\sum_{k=1}^{K} y_{ik} u_{jk}}{\sqrt{\sum_{k=1}^{K} y_{ik}^2} \sqrt{\sum_{k=1}^{K} u_{jk}^2}} \quad \textbf{(3)}$$

y_{ik} and u_{ik} are classic tf*idf feature of subtitle and synopsis. For subtitle, one document is all subtitles of one scene. While for synopsis, one document is one sentence of synopsis. y_{ik} / u_{ik} is defined as:

$$y_{ik} / u_{ik} = \begin{cases} (1+\log(tf_{ik}))\log\dfrac{N}{df_i}, & if \quad tf_{ik} \geq 1 \\ 0, & otherwise \end{cases} \tag{4}$$

N is the number of documents. T is pre-defined threshold.

C_{M*N} gives a primary subtitle alignment result. To further improve alignment accuracy, domain knowledge can be used to give two constraints on subtitle alignment which is shown in figure 2.

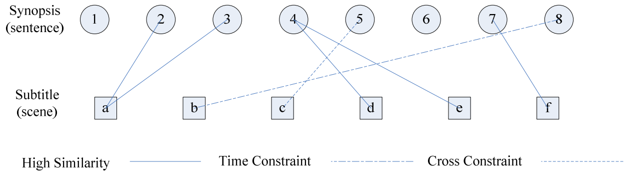

Figure 2. Constraints on Subtitle Alignment

The first is Time Constraint. Subtitle and synopsis are both temporal sequences. Although the relative positions of connected scene of subtitle and sentence of synopsis may not be exactly matched, the distance between them will not be long. Time constraint is described as:

$$\text{For } c_{ij} = 1 \text{, if } \left|\frac{i}{N_{sub}} - \frac{j}{N_{syn}}\right| > \alpha \text{, then } c_{ij} = 0 \tag{5}$$

N_{sub} is the number of subtitle documents. N_{syn} is the number of synopsis documents. α is pre-defined threshold.

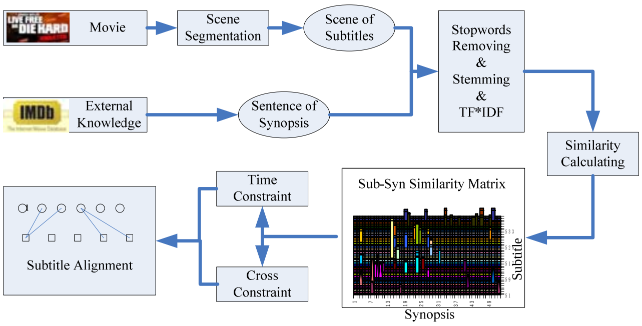

Figure 3. Subtitle Alignment

The second is Cross Constraint. Two conditions are managed separately. If crossed connections are more than 2, remove the connection which crosses the most other connections until no such cross left. If the crossed connections are only 2, the connection with lower similarity will be removed. The whole subtitle alignment process is shown in figure 3.

2.2 RoleNet Alignment

Subtitle alignment gives a good result when the subtitles of scene take enough information, especially when it is a dialogue scene. But some movie scenes have little subtitles. Sometimes there may even be no subtitle at all lasting for couple of minutes. For these kinds of scene, subtitle alignment will not work. We need to consider other ways to handle this condition.

One main factor of movie is role. The whole movie is a story of roles while each scene tells part of the story with certain roles. Based on analyzing roles in movie and their relations, RoleNet [6] has been proposed to present a brave new way to analyze movie content. We construct RoleNet for the movie and its synopsis. According to graph comparing, RoleNet alignment is proposed based on the result of subtitle alignment.

The RoleNet construction for synopsis includes two steps. The first step is synopsis segmentation. Based on subtitle alignment, the alignment between scene and synopsis has a primary result. The synopsis sentences between connected points are segmented into story level based on RoleNet information. The synopsis sentences and the roles in mentioned in them are viewed as a bipartite graph which is shown in figure 4 (a).

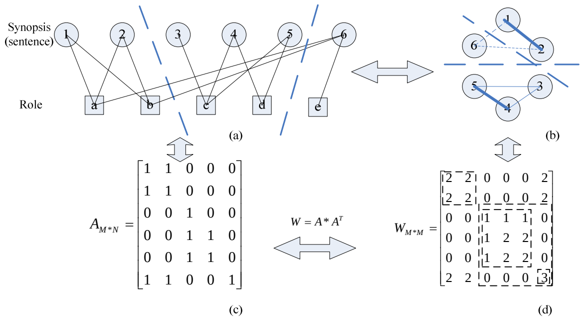
Synopsis
(sentence)
Role
1
2
3
4
5
6
a
b
c
d
e
(a)
(b)
$A_{M*N} = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 1 & 1 & 0 & 0 & 1 \end{bmatrix}$
$W = A * A^T$
$W_{M*M} = \begin{bmatrix} 2 & 2 & 0 & 0 & 0 & 2 \\ 2 & 2 & 0 & 0 & 0 & 2 \\ 0 & 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 1 & 2 & 2 & 0 \\ 0 & 0 & 1 & 2 & 2 & 0 \\ 2 & 2 & 0 & 0 & 0 & 3 \end{bmatrix}$
(c)
(d)

Figure 4. Synopsis Segmentation

The circular nodes denote synopsis sentences, and the square nodes denote roles. The edge between the *i*th circular node and the *j*th square node represents that the *j*th role appears in the *i*th synopsis sentence. For synopsis sentences between neighbored connected points during subtitle alignment which consist of *M* sentences and *N* roles, we can express the status of occurrence by a matrix $A = [a_{ij}]_{M*N}$ (figure 4 (c)), where the element

$$a_{ij} = \begin{cases} 1 & \text{if the } j\text{th role appears in the } i\text{th sentence} \\ 0 & \text{otherwise} \end{cases} \quad \textbf{(6)}$$

Based on this occurrence matrix, we can identify role's occurrence amount both in *i*th sentence and *j*th sentence by

$$W = A * A^T = [w_{ij}]_{M*M} \qquad w_{ij} = \sum_{k=1}^{N} a_{ik} a_{jk} \tag{7}$$

W (figure 4 (d)) is an approximate block diagonal matrix. Each block on the diagonal stands for one story segmentation of synopsis (figure 4 (b)).

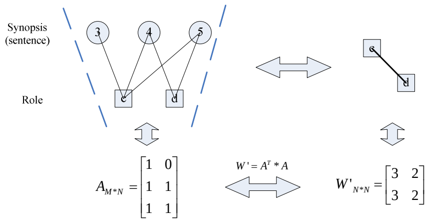
Synopsis
(sentence)
Role
3
4
5
c
d
c
d
$A_{M*N} = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 1 \end{bmatrix}$
$W' = A^T * A$
$W'_{N*N} = \begin{bmatrix} 3 & 2 \\ 3 & 2 \end{bmatrix}$

Figure 5. RoleNet Construction for Synopsis Segmentation

The second step is RoleNet construction for each segmentation which is shown in figure 5. It is similar to figure 4 but the *W'* is completely different from *W*.

$$W' = A^T * A = [w'_{ij}]_{N*N} \qquad w'_{ij} = \sum_{k=1}^{M} a_{ki} a_{kj} \tag{8}$$

w_{ij} stands for the relation between roles. The larger w_{ij} is (thicker edge), the closer the two roles are.

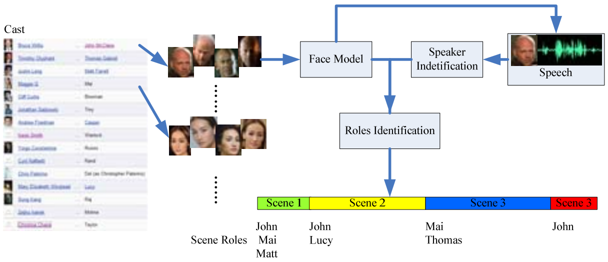

Figure 6. Scene-level Movie Role Detection

The RoleNet for movie video is in scene-level. The roles in each scene are detected based on face recognition and speaker identification. On IMDB website, cast of movie is provided. The cast list can lead to introduction of every cinemactor/cinemactress and the face model can be trained by their photos. According to face detection and speaker clustering result, speaker identification is proposed to enhance the face detection result. The scene-level role detection process is shown in figure 6. The RoleNet construction for scene is similar to that for synopsis segmentation.

RoleNet graph similarity [10] between synopsis and scene is calculated for RoleNet alignment. Similarly to subtitle alignment, the RoleNet similarity matrix is formed and time constraint and cross constraint is also processed. The RoleNet Alignment process is shown in Figure 7.

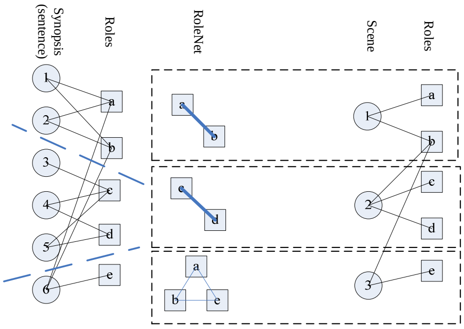

Figure 7. RoleNet Alignment

After subtitle and synopsis alignment, the synopsis sentences are assigned to each scene to give a scene-level description of movie. The imported synopsis can be used in many aspects of movie analysis, including movie summarization, content based scene indexing and retrieval. In following experiment, we will show the effectiveness of the proposed aligning methods and illustrate the application of imported text information in movie summarization.

3 Experiments And Discussion

3.1 Synopsis Alignment

With various techniques introduced above, it is important to understand how they can individually affect alignment. We screen through 5 top popular movies consisting of 10 hours video data and manually give scene segmentation of these movies. The synopsis sentences of these movies on IMDB website is manually aligned to the movie scene to give the ground truth of this experiment.

The experiment is designed to test the effectiveness of: (1) sub-syn similarity based subtitle alignment; (2) time constraint and cross constraint; (3) RoleNet alignment. The experiment is carried out on following 5 systems: (1) SA: subtitle alignment using sub-syn similarity matrix. (2) SAC: SA with time and cross constraints. (3) RA: RoleNet alignment. (4) RAC: RA with time and cross constraints. (5) SRAC: SAC + RAC

Table 1. Performance of synopsis alignment

	SA	SAC	RA	RAC	SRAC
Precision	0.42	0.88	0.25	0.37	0.80
Recall	0.23	0.21	0.82	0.79	0.83

From table 1, we can draw the following observations. First, the high precision and low recall results of subtitle alignment shows that the story clues subtitles take is effective in alignment but not sufficient. The high recall and low precision results of RoleNet alignment shows the discriminabiltity of RoleNet is limited. Second, the two constraints performed well in improve alignment precision with little miss-removing of accurate alignment. Third, the SRAC run yield best performance, demonstrating sequentially processing subtitle alignment and RoleNet alignment complementally is effective.

3.2 Movie Summarization

This series of test investigate the effectiveness of importing text information for movie summarization. Besides traditional audiovisual saliency, textual saliency is proposed in scene-level using scene-level description. Three kinds of text feature are provided to reflect the information, fondness and classic aspects of the description, which are Information Feature (IF), Fondness Feature (FF) and Classic Feature (CF). IF is formed by top N tf*idf value of each scene description. The scene with high IF value means it contains more information than others. FF is sum of tf*idf value of "interested" words. The interested words' list is generated by WordNet [7] with seed words provided by users as their preference. The contributors of synopsis always refer to classic dialogue in the synopsis to reveal their subject highlight judgment. The Boolean-valued CF denotes this information. 1 means having synopsis highlight in the scene while 0 means not. The text saliency (TS) function can be formed as below.

$$TS(s) = \alpha IF(s) + \beta FF(s) + \gamma CF(s) \qquad \textbf{(8)}$$

α , β , γ are weights and they can be set according to which aspects users prefer.

Table 2. Performance of multi-model saliency based summarization

	VA	VAIF	VAFF	VATS
Evaluation	3.1	3.0	3.8	4.2

The experiment is designed to test the effectiveness of imported text information in video summarization. Ten volunteers of cinephiles are selected to evaluate the 5 minutes summarizations generated by the following 3 systems. (1) VA: traditional audiovisual saliency based summarization. (2) VAIF: VA + IF only. (3) VAFF: VA + FF only. (4) VATS: VA + TS. Volunteers are asked to rank the systems in terms of summarization preferences from a scale from 5 to 1 corresponding to “best” to “worst”.

From table 2 we can draw the following observations:First, VAIF shows a little lower performance than VA which means IF is useless. That mainly because movie summarization is a compact showing of movie content, highlight but not story clue is what users want to see. It’s also hard to show users the story clue in a so short summarization. Second, VAFF shows better performance than VA which means FF performs well. FF gives high score of audiovisual saliency when there is corresponding words of users’ seed appear in description text of the scene. Third, CF contributes a lot for VATS performing best. Although there are only one or two classic scenes in one movie, the appearance of these well-known scenes apparently gives users good impressing and gets VATS the highest score.

4 Conclusion

We have proposed two alignment methods to import synopsis to movie as text information to enhance movie analysis. Experiments showed that the two alignment methods complementally performed well. As illustration, the experiment on movie summarization showed importance of text information.

5 Acknowledgments

This work was supported in part by the National Nature Science Foundation of China (60873165), the National Basic Research Program of China (973 Program, 2007CB311100), the National High Technology and Research Development Program of China (863 Program, 2007AA01Z416).

References

1. K., Rapantzikos, G., Evangelopoulos, P., Maragos, and Y., Avrithis. 2007. An audio-visual saliency model for movie summarization. In IEEE 9th Workshop on Multimedia Signal Processing, 2007, MMSP 2007, pages 320-323, October 2007.
2. Liu, Anan, Li, Jintao, Zhang, Yongdong, Tang, Sheng, Song, Yan, and Yang Zhaoxuan. 2008. An innovative model of tempo and its application in action scene detection for movie analysis. In IEEE 2008 Workshop on Application of Computer Vision, WACV 2008.
3. Wang, Hee Lin, and Cheong, Loong-Fah. 2006. Affective understanding in film. In IEEE Transactions on Circuits and Systems for Video Technology, pages 689-704, June 2006.
4. Li, Ying, Lee, Shih-Hung, Yeh, Chia-Hung, and Kuo, J. C.-C. 2006. Techniques for movie content analysis and skimming. In IEEE Signal Processing Magazine, vol. 23, no. 2, pages 79-89, March 2006.

5. Li, Ying, S., Narayanan, and C.C.J., Kuo. 2004. Content-Based Movie Analysis and Indexing Based on AudioVisual. In IEEE Transactions on Circuits and Systems for Video Technology, pages 1073-1085, August 2004.
6. Weng, Chung-Yi, Chu, Wei-Ta, and Wu, Ja-Ling. 2007. RoleNet: Treat a Movie as a Small Society. In Multimedia Information Retrieval, MIR 2007, pages 51-60.
7. Fellbaum, Christiane, etc. 1998. WordNet: An Electronic Lexical Database. MIT press, May 1998.
8. Wikipedia: www.wikipedia.org
9. Blondel D., Vincent, Gajardo, Anahi, Heymans, Maureen, Senellart, Pierre, and Van Dooren, Paul. 2004. A measure of similarity between graph vertices: applications to synonym extraction and web searching. In society for industrial and applied mathmatics, vol. 46, no. 4, pages 647-666. October 2004.