SEMI-SUPERVISED LEARNING TOWARDS COMPUTERIZED GENERATION OF MOVIE TRAILERS

XingChen Liu
School of Computer Science and Technology
Tianjin University
Tianjin, China
xcliu@tju.edu.cn

JianMing Jiang*
Research Institute for Future Media Computing
Shenzhen University
Shenzhen, China
Jianmin.jiang@szu.edu.cn

Abstract— **Generating movie trailers is recognized as an effective way for promoting films and attracting viewers to increase its publicities. Existing techniques, however, are mainly represented by manual process in generating movie trailers, which is time-consuming, expensive, and unreliable for the subjectivity of movie trailers. In this paper, we make use of multi-modality features, and propose a semi-supervised learning approach towards automated generation of movie trailer shots without involving any human efforts or manual process. By using a number of well-known movies and their manually-made trailers as the ground truth, we carry out extensive experiments to demonstrate that our approach remains competitive and even outperforms the manually supervised method under certain assessment criteria. Since there exists no commonly agreed evaluation mechanisms for movie trailer generation at present, we show and analyze our experimental results in a number of different angles to support the feasibility and reliability of our proposed approach, providing a good potential for the film industry.**

Keywords-Video summarization; movie trailers; video skimming; semi-supervised learning

I. INTRODUCTION

Along with the advancement of multimedia technologies, Internet-based WEB TV, and the higher speed networking facilities, videos are becoming the major examples of big data in multimedia, among which movies, films and soap series remain to be the representative forms. On the other hand, people often find it difficult to make the right choice in watching movies in order to get relaxed and enrich their life experiences. Facing the enormous resources of multimedia big data, how to quickly find the one desirable by viewers becomes a crucial and challenging problem. In the movie-making industry, movie trailers are generated to solve this problem, in which a short video footage with highlight of sensational scenes and actions about the movie content is generated manually to promote the movie. Being impressed by the trailers, viewers can often decide whether this is something worthwhile to watch the film or not according to the level of match between their interests and the trailer content. As such manual process is expensive and time-consuming, automated generation of trailers via computerized analysis of movie content and summarization becomes an interesting and valuable research topic.

In principle, generating a movie trailer is an application of video summarization. There exist two types of video summarization approaches: static video summarization and dynamic video skimming [1] [2]. A static video summarization is a collection of images that can represent the underlying video content. Many approaches of static summarization have been proposed, including those methods based on sampling [3], video temporal segmentation [4], and visual attention etc. [5]. Dynamic video skimming, on the other hand, produces a summarization video sequence with highlight of events, which can be continuously played back [1]. A summary sequence is used to provide viewers with a brief introduction about the entire video content [6], while a highlight only contains the most interesting parts of the original video.

Although movie trailers can be categorized as part of video skimming, there still exist significant differences in comparison with common video summarization techniques. Generating a movie trailer is an artistic process, which is essentially a subjective process, and many other information or video frames about the movie will be artificially made and added into the movie trailer. Examples include video footages of the cast, special effects and clips in different orders etc. in order to attract viewers and enrich their experiences.

According to the literature survey, a good movie trailer should possess five basic properties [7], including:

1) Important objects / people, i.e. the most important objects or actors ought to appear in the trailer;
2) Actions, highlighting the story sequence or events inside the movie;
3) Mood, reserving the mood of the original movie;
4) Dialog, a change of actions or atmosphere usually occurs with the dialog scenes;
5) Disguise the outcome of the story, creating mysterious atmosphere for the movie.

978-1-4799-8697-2/15 $31.00 © 2015 IEEE
DOI 10.1109/SMC.2015.520

For the same movie, different directors would produce different trailers as the design, understanding, approaches and intention are entirely variable. As a result, it is not appropriate to take the official movie trailers as the supervised information. To this end, a semi-supervised context learning algorithm is proposed to exploit other informative clips of the movie. We utilize a small amount of official movie trailers of some movies as the limited known information, and then predict trailers for a new movie by using a semi-supervised learning method via multi-modality features.

The rest of this paper is organized as follows: Section 2 briefly describes the related work on movie trailer generation out of our literature survey on the existing research. Section 3 introduces and reports our proposed approach in detail, Section 4 presents our experimental results and analysis, and finally conclusions are drawn in Section 5.

II. Related Work

While there are numbers of representative approaches and categorizations in video summarization [1] [2], a movie trailer has to show some of the most attractive scenes without revealing the story's end [1]. Silvia Pfeiffer et al. proposed a VAbstract system in 1996. They defined a series of standards about movie trailers, and used a number of important cues, such as contrast perception, color perception, context perception and stimuli processing, to extract clips from the original video [7]. Although there are some interesting ideas in this paper, the essential parts of the algorithm are too simple to be effective. Deciding which video segments to be included in a movie trailer is actually a very subjective process, and it is hard to map human cognition into the automated summarization process. The easiest way is to compress the original film with fast-preview mode, for which Omoigui et al. reported their work to enable people to watch the film in a shorter time with fast playback [8]. Similar to Omoigui's work, Amir et al. proposed an approach called Time Scale Modification (TSM) via using audio processing technology. This method was developed for fast video browsing and efficient video-based learning. These audio based approaches, however, only allow a maximum time compression of 1.5-2.5 depending on the speech speed [10], otherwise the speech will become unrecognizable.

Other researchers in this field decided to make use of multiple features in formulating their summarization videos. Smeaton et al. extracted a set of audiovisual features to model the characteristics of shots typically present in trailers, and a SVM (Support Vector Machine) was utilized in order to select the relevant shots [11]. Hermes et al. designed a fully annotated system for producing movie trailers [12]. Different from work in [12], Xu et al. presented a method of selecting candidate movie clips automatically, in which they also trained a supervised SVM with the official movie trailers to classify and determine which clips should be included inside the movie trailer [13]. In addition, there are also many approaches based on scenes and events. While scene-based summarization splits a movie into a set of scenes and hence obtains the index of the candidate scenes [14] [15] [16], event-based movie summarization aims to detect shots in the original film that belong to a certain event type, such as dialogues and violent event detections [17] [18].

Most of the existing approaches extract a few features to detect actions, faces or events, based on which the clips with higher scores are chosen to generate the movie trailer. In [13], the authors used SVM to generate the movie trailer, in which a matching process is utilized by using a so-called SURF (Speeded Up Robust Features) method to find the key-frames similar to those inside the official movie trailer, and labeled it as positive. Since the official movie trailer is made by editors manually and many films have several different movie trailers, it is not appropriate to exclude all the other non-selected key-frames directly. To this end, we resample these key-frames and only label them as negative instances instead of complete exclusion, providing spaces for further learning in the process of movie trailer generation.

Generally speaking, supervised-learning based movie trailer algorithms often perform better than semi-supervised ones. The disadvantage, however, lies in the fact that 1) the supervised learning often requires the training data to include all the key-frames of the film, and hence making it computationally expensive and time-consuming; 2) less flexibility and adaptability as the supervised learning often generate results with too much dependencies upon the trained samples, making it weak in capturing new elements of the varying input. Although supervised-learning algorithms, such as SVM [13], have been used in this field early and gained good performance, it is hard to get all the films and corresponding trailers simultaneously in real life. As a result, we introduce a semi-supervised learning approach in this paper, and demonstrate that the proposed algorithm produces competitive performances in terms of generating movie trailers with high level of consistency when compared with the ground truth.

III. The Proposed Movie Trailer Generation Via Semi-Supervised Learning

Considering the fact that most of the movie trailers will have similar key elements, such as primary actors, dialogues, actions and exciting scenes, we can extract a small sample set out of those official trailers and their corresponding films to learn and hence predict the best possible elements to construct trailers for input movies. This scenario fits in well with the nature of using a semi-supervised learning approach. As a result, our proposed video trailer generation scheme can be characterized with three stages of operations, which include: 1) extraction of key frames and visual features; 2) preparation of labeled and unlabeled data set; 3) classification based on semi-supervised learning and generation of movie trailers. Fig. 1 shows the main framework of our approach.

All videos share the common nature that dynamic effects can be revealed by playing back quickly, and the content of adjacent frames is often similar to each other. In order to process the video more efficiently, we extract key-frames of official movie trailers and films by using the technique reported in [19]. Following the key frame extraction, we then extract two visual features, GIST and Color Moment, out of these key

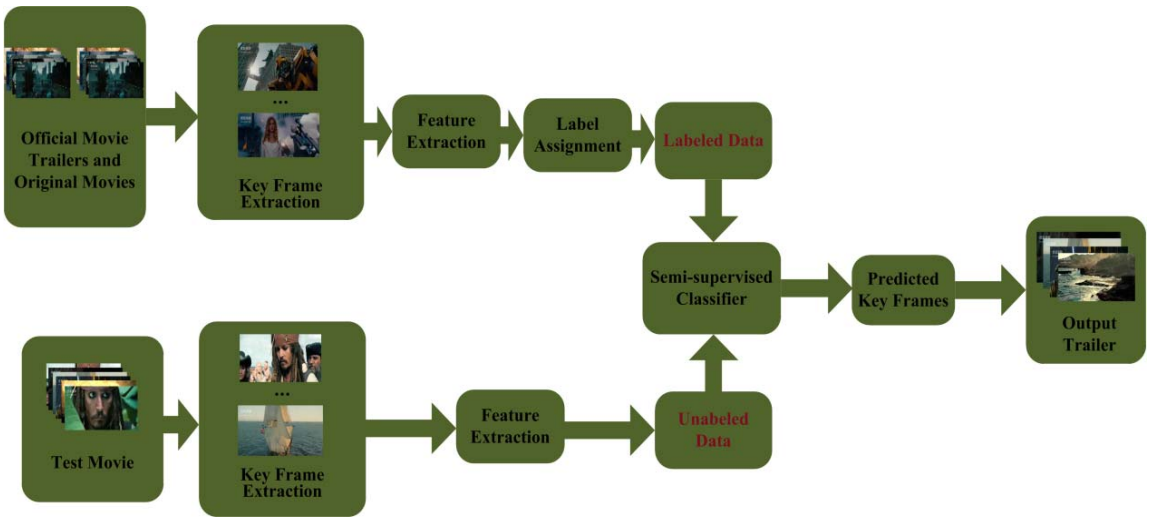

Figure 1. Main framework of the proposed movie trailer generating approach.

frames to construct the corresponding feature matrices in order to pave the way for classification of key frames.

To prepare the training set and drive our proposed semi-supervised learning algorithm, we prepare sequences of positive and negative instances as the labeled data, in which the key frames extracted from the official movie trailers (training set) are used as the positive samples, and the remaining key frames with the parts similar to trailers being removed are used as the negative samples. Key-frames out of the films without trailers and the remainder are taken as the unlabeled data.

Given the set of feature GIST, $G \in R^{(l+u)\times 512}$. There is a set of labeled data $\{g_i, y_i\}_{i=1}^{l}$, and a set of unlabeled data $\{\hat{g}_j\}_{j=1}^{u}$, where $g_i, g_j^{'} \in G$, $y_i \in \{\pm 1\}$, l and u are the number of labeled data and unlabeled instances, respectively. We have to find a function which can predict the labels of unlabeled data $\hat{y}$, the objective function of the single large-margin low-density is formulized as follows:

$$h(f,\hat{y}) = \min \frac{\|f\|_{\mathrm{H}}}{2} + C_1 \sum_{i=1}^{l} \ell(y_i, f(g_i)) + C_2 \sum_{j=1}^{u} \ell(\hat{y}_j, f(\hat{g}_j)) \quad (1)$$

$$f \in \mathrm{H}, y' \in \mathrm{B}$$

where H is the Reducing Kernel Hilbert Space (RKHS), B is a set of label assignments obtained from domain knowledge, and $\ell(y, f(g)) = \max\{0, yf(g) - 1\}$ is the hinge loss. C1 and C2 are regularization parameters.

To obtain multiple separators $\{f_t\}_{t=1}^{T}$ and the corresponding label assignments, the following formulation is proposed:

$$\min_{\{f_t, \hat{y}_t \in B\}_{t=1}^{T}} \sum_{t=1}^{T} h(f_t, \hat{y}_t) + M\Omega(\{\hat{y}_t\}_{t=1}^{T}) \quad (2)$$

where T is the number of separators, M is a large constant enforcing large diversity and Ω is a number of penalty about the diversity of separators.

To make the improved performances, we learn a classifier y over an inductive SVM [20] and maximize its roles by projecting into an optimization problem:

$$\arg\max_{y \in \{\pm 1\}^u} \min_{\hat{y} \in \{\hat{y}_t\}_{t=1}^{T}} earn(y, \hat{y}, y^{svm}) - \lambda lose(y, \hat{y}, y^{svm}) \quad (3)$$

where $earn(y, \hat{y}, y^{svm})$ and $lose(y, \hat{y}, y^{svm})$ are the increased and decreased accuracies compared to the inductive SVM, respectively. λ is a parameter for trading-off how much risk you want to undertake, y^{svm} denotes the predicted labels of inductive SVM, and $\hat{y}$ is the worst-case separator in $\{\hat{y}_t\}_{t=1}^{T}$, we assume it can realize the ground truth boundary.

To increase the reliability of our automated generation of movie trailers, we propose three measures as described below.

A. Higher Confidence Coefficient

Introducing a confidence coefficient to increase its reliability in classifying positive samples, and the confidence coefficient is defined as given below:

$$\delta_j = \frac{1}{T} \sum_{i=1}^{T} \exp(-1/|d_i(g_j)|), 1 \le j \le u \quad (4)$$

where $d_i(g_j)$ is the distance between g_j with i-th separator. As a result, if the positive instances predicted in (3) have higher confidence coefficient than a threshold we set, we label them as positive and the others are negative.

Figure 2. Some scenes in official trailer and our generated trailer of "*Edge of Tomorrow*". Picture (a) ~ (e) are scenes in the official movie trailer, and picture (f) ~ (j) are scenes in our generated movie trailer.

B. Fusion Process

Extracting two features from test movies to drive the S4VMs, and the final result is derived from a simple fusion process, in which only under the condition that both the two S4VMs predicting one instance as positive, we label it as positive sample, i.e.:

$$\hat{y}_j = \begin{cases} 1 & \hat{y}_j^G = \hat{y}_j^C \\ 0 & \hat{y}_j^G \neq \hat{y}_j^C \end{cases}, 1 \le j \le u \qquad (5)$$

where $\hat{y}_j^G$ and $\hat{y}_j^C$ are labels predicted by GIST-based and Color-Moment-based classifications respectively.

C. Audio Information collabration

For a video, there is not only visual content, audio information occupies a big part of the video. To further improve the performance and reliability of the proposed algorithm, we also take the audio information into consideration, due to the fact that audio information remains to be an important factor for drawing attentions, such as screams or explosive sounds etc. Since different films often have their own audio features, such audio analysis is carried out as a post-processing of the S4VM-based learning process. Specifically, the volume and tone are extracted to refine the candidate key-frames, in which a threshold is used to choose the final key frames when their associated audio volume and tone are higher than the threshold.

Generally speaking, shot is often used as the fundamental element in making movies. However, since movie trailers are short summarization of movies, only part of a shot is often used in constructing movie trailers. It is revealed that human visual system takes about 400 to 700ms to process one visual stimulus [7], and human has to take at least 3.25s to analyze a scene. Meanwhile, our empirical studies discover that the time length a scene occurs in a trailer is no more than 3s in general. As a result, the movie trailer is generated with clips constructed around each key-frame, where neighboring frames are reserved to make the clip about 3s long.

IV. Experimental Settings And Evaluations

To evaluate the proposed algorithm, we set up a database of 14 movies, and their descriptions are given in Table 1. Out of these 14 movies, since the semi-supervised method we use, 4 movies and their corresponding trailers are selected as the labeled data and the others are used as unlabeled data. Table 2 provides additional information about all the four official trailers produced by the movie makers. Visual observation of these movie trailers indicate that many artificial special effects have been added to movie trailers. To ensure an accurate learning, we removed these special effects before the features are extracted to drive the S4VM-based learning unit.

To trade off the time cost and the accuracy of classification, we choose the linear Kernel to construct the S4VMs [20], and its related parameters are set up as: C1 = 100, C2 = 0.01, and sample-time = 10 from validation of our empirical studies.

Fig. 2 shows some sample scenes from both the official trailer and our generated trailer for the film "Edge of Tomorrow", it can be seen that our predicted movie trailer has maintained a high level of similarities as compared with the official trailer in many cases. Our trailer contains the basic properties we mentioned in Section 1, i.e., main actors, dialogues, intense explosion and actions in the film, as shown in picture (f) ~ (i), providing a good match with picture (a) ~ (d) inside the official trailer. In addition, movie trailers do not always cover fast moving or violent scenes only. Some peaceful scenes, such as the one shown in picture (e), will also be included in many cases. From the picture (j), it is demonstrated that our approach is capable of extracting these quiet and natural scenes which can also advertise movies at the same time.

Fig. 3 illustrates the timeline of some scenes occurring in the official trailer of "Men in Black 3", and the bottom of the timeline shows the scenes included in our automatically

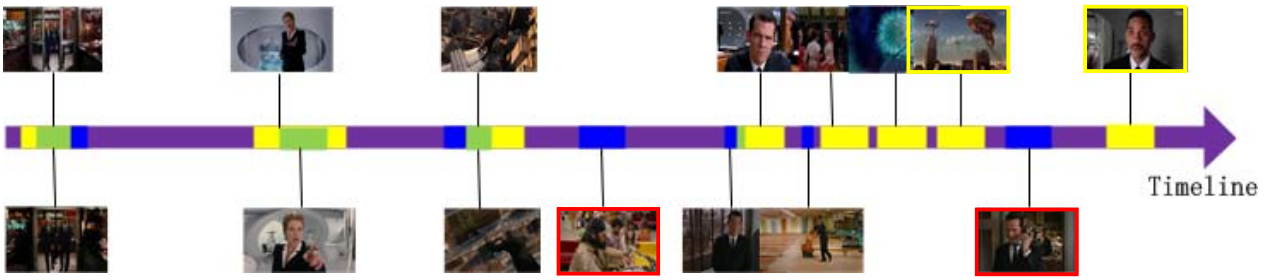

Figure 3. Some matched scenes between our generated movie trailer and the official trailer of "*Men in Black 3*" in a continuum movie clips. This is timeline of the film. The yellow parts on the timeline are scenes belong to official trailer, and blue parts are components in our generated trailer. The green parts denote that our trailer overlap the official trailer in the original film.

generated trailer. We can see that our generated trailer overlaps with the official trailer on several scenes, which demonstrates that our approach achieves competitive performances in producing trailers. It can also be noted that a few scenes in our produced trailer (highlighted with red rectangles) are not included in the official trailer, which indicates that our proposed algorithm provides sufficient space and flexibilities for producing attractive scenes. Further examination reveals that all these selected scenes are very salient on many aspects,

TABLE I. Movie information in our experiments.

Movie	Time (hh: mm: ss)	fps
Les Misérables	02: 38: 07	23
The Dark Knight Rises	02: 44: 32	23
Transformers 3	02: 34: 21	23
Inception	02: 28: 07	23
Iron Man 3	02: 10: 32	23
Men in Black 3	01: 45: 50	23
Pirates of the Caribbean 4	02: 16: 24	23
The Twilight Saga	02: 01: 50	23
X-Men 2	02: 13: 47	23
Thor 2	01: 52: 03	23
Edge of Tomorrow	01: 53: 27	23
Harry Potter and the Goblet of Fire	02: 37: 05	23
Prometheus	01: 58: 42	25
Resident Evil: Retribution	01: 35: 37	23

TABLE II. Official trailer information of the labeled movies.

Movie	Trailer Time (hh: mm: ss)	fps
Les Misérables	00: 01: 30	25
The Dark Knight Rises	00: 00: 29	25
Transformers 3	00: 00: 30	25
Inception	00: 00: 52	15

including audio, color and actions, among all other frames. It can also be noticed that some scenes, as highlighted in yellow rectangle, inside the official trailer are not included or missed in our result, which indicates that these scenes are not very meaningful to S4VMs, and hence S4VMs classify them into negative classes.

As there is no criterion to judge whether a movie trailer is good or not, human often apply subjective observations and

TABLE III. Match rate (%) of different methods on official movie trailers.

Movie	Random Selection	Supervised SVM	Ours
Iron Man 3	3.77	23.08	**29.12**
Men in Black 3	3.33	16.67	**21.67**
Pirates of the Caribbean 4	3.05	17.24	**26.09**
The Twilight Saga	15	20.69	**24.81**
X-Men 2	5.88	17.48	**22.78**
Thor 2	3.92	13.10	**47.06**
Edge of Tomorrow	15.13	18.88	**26.32**
Prometheus	12.5	15.21	**28.33**
Resident Evil: Retribution	6.45	12.24	**25.81**
Harry Potter and the Goblet of Fire	10.17	18.25	**21.85**
Mean match rate	7.92	17.28	**27.38**

comparisons to make the initial assessment of the proposed algorithm. We propose a further comparative evaluation which can be carried out by including objective elements and consideration of those official trailers as the ground-truth. For movie trailers, trailer published by the official movie makers is the ground-truth. Calculate the similarities between each frame in an official trailer and all frames in our trailer by their correlation. Specifically, a threshold of 0.9 is set up to screen all the frames inside our generated trailers in comparison with the ground truth. In the case of the maximum similarity being bigger than the threshold, we identify the two matched frames matched, and then delete the two matched frames from the official trailer as well as our generated trailer. Finally, the matching rates between our movie trailers and the official trailers are produced.

To provide a benchmark, we randomly extract several clips from the original films to compose movie trailers and see how such a scheme compares with the proposed algorithm. Table 3 shows the results about randomly extracting clips from original films, recognizing trailer clips with supervised SVM and our algorithm. The Best results are bold in the table. We can see that our approach outperforms the other two methods on all test movies and mean value, which provides additional benchmarking for the proposed scheme.

V. Conclusion And Future Work

In this paper, we proposed a semi-supervised learning approach with the consideration of multi-modality features to generate movie trailers. We extract visual and audio features

from key frames of official movie trailers and the original films to train the SVM-based machine learning unit, and activate it as semi-supervised learning to classify the input videos and generate movie trailers. Experimental results show that our produced trailers are close to those official trailers, and outperform the randomly selected clips, which provide good potentials for human editors to select candidate scenes for artificial production of movie trailers.

ACKNOWLEDGMENT

The authors wish to acknowledge the financial support from the Chinese Natural Science Foundation under the grant No 61373103.

REFERENCES

[1] Y. Li, T. Zhang, D. Tretter, "An Overview of Video Abstraction Techniques,"Technical report, HP Laboratory Technical Report, HPL-2001-191, Jul. 2001.

[2] S. Lu, "Content Analysis and Summarization for Video Documents,"PhD thesis, Research Associate, VIEW lab, the Chinese University of Hong Kong, Department of Computer Science and Engineering, Dec. 2004.

[3] Y. Mills, A. Akutsu, Y. Tonomura, and H. Hamada, "An intuitive and efficient access interface to real-time incoming video based on automatic indexing," in ACM International Conference on Multimedia, pp. 25-33, Nov. 1995.

[4] W. Sabbar, A. Chergui, and A. Bekkhoucha, "Video Summarization Using Shot SegmentaionAnd Local Motion Estimation," in International Conference on Innovative Computing Technology, pp. 190-193, Sep. 2012.

[5] P. Jiang and X. L. Qin, "Keyframe-Based Video Summary Using Visual Attention Clues,"IEEE Multimedia, vol. 17, no. 2, pp. 64-73, Apr-Jun. 2010.

[6] K. Dale, E. Shechtman, S. Avidan, and H. Pfister, "Multi-Video Browsing and Summarization,"Signal Processing, vol. 89, no. 12, pp. 2354-2366, Dec. 2009.

[7] S. Pfeiffer, R. Lienhart, S. Fischer, and W. Effelsberg, "Abstracting Digital Moives Automatically,"Visual Communication and Image Representation, vol. 7, no. 4, pp. 345-353, Dec. 1996.

[8] N. Omoigui, L. He, A. Gupta, J. Grudin, and E. Sanocki, "Time-Compression: Systems Concerns, Usage, and Benefits," in ACM Conference on Computer Human Factors in Computing Systems, pp. 136-143, May. 1999.

[9] A. Amir, D. Ponceleon, B. Blanchard, D. Petkovic, S. Srinivasan, and G. Cohen, "Using Audio Time Scale Modification for Video Browsing," in International Conference on System Science, vol. 1, Jan. 2000.

[10] G. W. Heiman, R. J. LEO, and G. Leighbody, "Word intelligibility decrements and the comprehension of time-compressed speech,"Perception and Psychophysics, vol. 40, no. 6, pp. 407-411, 1986.

[11] A. F. Smeaton, B. Lehane, N. E. O'Connor, C. Brady, and G. Craig, "Automatically Selecting Shots for Action Movie Trailers,"in ACM International Workshop on Multimedia Information Retrieval, pp. 231-238, Oct. 2006.

[12] T. Hermes and C. Schultz, "Automatic generation of hollywood-like movie trailers,"in cat1.netzspannung.org, Dec. 2006.

[13] Z. Xu and Y. Zhang, "Automatic Generated Recommendation for Movie Trailers," in IEEE International Symposium on Broadband Multimedia Systems and Broadcasting, Jun. 2013.

[14] J. R. Kender and B. -L. Yeo, "Video scene segmentation via continuous video coherence," in IEEE Conference on Computer Vision and Pattern Recognition, pp.367-373, Jun. 1998.

[15] H. Sundaram and S. -F. Chang, "Determining computable scenes in films and their structures using audio-visual memory models," in ACM International Conference on Multimedia, pp. 95-104, 2000.

[16] Y. Cao, W. Tavanapong, K. Kim, and J. Oh, "Audio-assisted scene segmentation for story browsing," in International Conference on Image and Video Retrieval, pp. 446-455, Jul. 2003.

[17] L. Chen, S. J. Rizvi, and M. Otzu, "Incorporating audio cues into dialog and action scene detection," in SPIE Conference on Storage and Retrieval for Media Database, pp. 252-264, Jan. 2003.

[18] J. Nam, M. Alghoniemy, and A. H. Tewfik, "Audio-visual content-based violent scene characterization," in International Conference on Image Processing, vol. 1, pp. 351-357, Oct. 1998.

[19] A. -A. Wael, "Online, Simultaneous Shot Boundary Detectionand Key Frame Extractionfor Sports Videos Using Rank Tracing,"in International Conference on Image Processing, pp. 3200-3203, Oct. 2008.

[20] Y. -F. Li, Z. -H. Zhou, "Towards Making Unlabeled Data Never Hurt,"Pattern Analysis and Machine Intelligence, pp. 1-14, Jan. 2014.