MULTI-MODAL TOPIC UNIT SEGMENTATION IN VIDEOS USING CONDITIONAL RANDOM FIELDS

Su Xu, Bailan Feng and Bo Xu

Interactive Digital Media Technology Research Center,
Institute of Automation Chinese Academy of Sciences, Beijing, China
{su.xu, bailan.feng, xubo}@ia.ac.cn

ABSTRACT

In this paper a novel approach of video segmentation into topic units is presented. This approach is built upon the design in which topic unit segmentation is transformed into label identification problem by defining four types of shots that reveal semantic structure of it. To implement our algorithm, four middle-level features including shot difference signal, scene transition graph, shot theme and audio type are extracted to depict the label properties of each shot, and then CRFs model is employed to identify the labels sequence. CRFs model integrates context information, so it produces accurate results in topic unit segmentation. The proposed approach is verified by two types of data: documentary and news. Experiments on testing data set yield average 86% F-measure, which illustrates that the proposed method can accurately detect most topic units in different genres of programs.

Index Terms— Topic unit segmentation, Conditional random field

1. INTRODUCTION

Topic unit is regarded as a series of shots for which a subject matter is discussed [1]. This term is most of the time used with event-driven video content, such as documentary and news. The task of automatic topic unit segmentation is to divide the video into topically homogeneous segments. Such a techniques is an essential prerequisite for a wide range of video manipulation applications, such as content indexing, and retrieval, non-linear browsing, summarization etc [1].

In the early stage, the methods of topic unit segmentation are early based on a set of production rules of how program should be composed. In [2], Gao et al. assume that news stories begin with the anchorperson shots, so they propose an unsupervised method that groups these shots using minimum spanning tree (MST) clustering to segment the news. In [3], Wang Ce et al. add silence and caption information to design heuristic rule for news story segmentation. However, these methods are not applicable to various genres of TV-news program. Instead of designing heuristic rule, some authors exploit model-based methods for topic unit segmentation. In [4], Chaisorn et al. employ decision tree technique to classify the shots into one of 13 predefined categories and then perform the Hidden Markov Models (HMM) analysis to locate the boundaries of topic units. In [5], Xie et al. propose a method of topic unit segmentation based on the robustness speech recognition technique. In [6], Wang et al. employ a SVM-based method to identify program boundaries using a variety of low-level multi-modal features. The model-based methods are very likely to break one topic into a few segments leading to poor precision. To address this problem, Feng et al. proposed an SVM-detector and dynamic programming (DP) refiner scheme to segment news story [7]. A common deficiency of the reviewed techniques is that they are only against one kind of topic unit—TV-news. The documentary is another kind of program that is organized by topic unit, but these methods are not applied on it. Another deficiency is that they ignore the context information of neighboring shots to segment topic units. Due to semantic connectivity in the units, the features of neighboring shots provide important context information to judge topic unit boundaries, which can effectively decrease miss or falseness of segmentation.

In this paper, we propose a novel approach for topic segmentation. Compared with prior work in the field, one novelty of our approach is that we transform topic unit segmentation into label identification problem by defining four kinds of labels. Such a definition is easily applicable to different styles of topic unit: news and documentary. Another novelty of our method is that the proposed approach employs Conditional Random Fields (CRFs) technique to predict labels. Because this technique adequately utilizes the context information of neighboring shot, it delivers significantly more accurate results than previous methods.

2. TOPIC UNIT SEGMENTATION ALGORITHM

2.1. Analysis of proposed algorithm

According to the observations on a great number of actual data, positional property of shots in the topic units can be classified into four categories: begin shot (BS) is a start point of the new topic unit; end shot (ES) indicates that a uint ends in this shot; middle shot (MS) is the internal shot between BS and ES; single shot (SS) is a kind of independent shot that contains a complete topic. Such structures are given in Fig. 1 (a) in which BS {1, 8}, MS {2, 3, 4, 5, 9, 10}, ES {6, 11} and SS {7} construct three typical topic units in video sequence. In the topic unit sequence, the production rule of news and documentary can help to identify the types of shots. Taking TV-news for instance, some news programs are produced by following rules: 1) a new topic begins with an anchorperson shot or a silence segment; 2) one topic has a headline caption to abstract its content. Therefore, anchorperson or silence feature indicates that a shot probably belong to BS; see shot 1 and shot 8 in Fig. 1 (a). The ES does not have distinguishing features to indentify it, but it can be judged by the features of neighboring shots. For example, if the a shot has anchorperson or silence segment, the previous one may be the ES; see shot 6 in Fig. 1 (a). MS can be identified by headline caption; see shot 5 in Fig. 1 (a). SS commonly contains a complete subject matter in a single shot that can extract both anchorperson and head-

The work was supported by National Nature Science Foundation of China(grant No. 61202326)

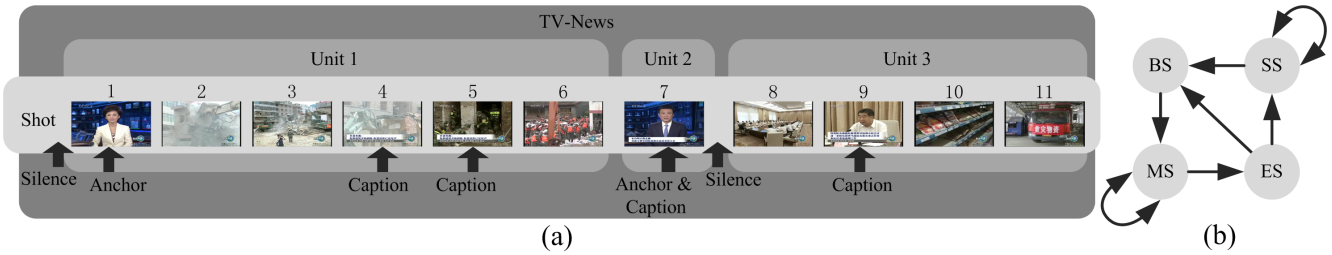

Fig. 1. (a)The structure of shots in topic unit sequence; (b)Sate transition graph of different types of shots in the topic units.

line caption from it; see shot 7 in Fig. 1 (a). Stated thus, every shot in a topic unit has its positional property, so topic unit segmentation can be transformed into label identification problem. If we correctly identify all labels, we can obtain the topic unit boundaries between ES and BS, ES and SS or SS and SS.

To accurately identify the types of shots in a topic unit, the proposed method employs CRFs technique that is an effective algorithm to predict multiple variables depending on each other. Firstly, CRFs model estimates the state transition probabilities, when the states transform between different types of label. State transition relationship is given in Fig. 1 (b), where direction of arrows indicates the way of state transition and these probabilities can be calculated by training data. These transition directions and probabilities simulate the retaliation of different labels in the shot sequence conducting reasonable results of label estimation. Secondly, CRFs technique counts all priori probabilities of the four states, which promote the accuracy of label estimation. For example, if there are not effective features to identify type of a shot, it will count the most likely label from training data as the outputting label. Thirdly, comparing with other models, features of neighboring shots are taken into account when predict the current label. In the sequence of Fig. 1 (a), shot 3 does not have helpful features to identify its type, but due to between anchorperson in shot 1 and headline caption in shot 4 this shot should express the detail of the report in a topic unit. For this reason, it probably belongs to MS. CRFs model integrates the above advantages, so it can accurately predict label of shots for topic unit segmentation.

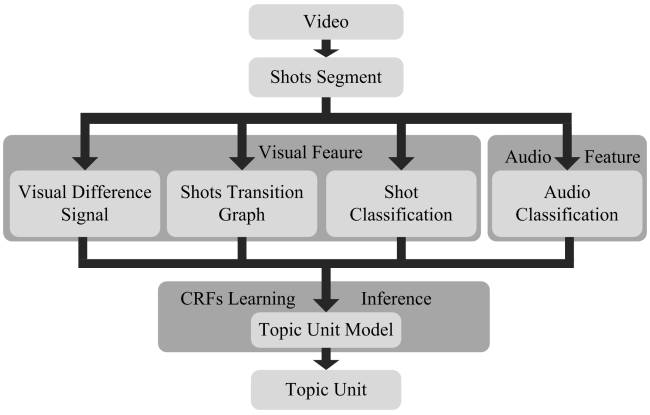

Fig. 2. Flow chart of our algorithm.

In Fig. 2, we summarize the main steps of our approach. The video is divided into shots using algorithm in [8]. Next, visual and audio features are extracted based on shot units, while those features are discretized to adapt CRFs tools. Finally, a CRFs model is trained and used to predict shot labels.

2.2. Features

We choose four middle-level features: shot difference signal (SDS), scene transition graph (STG), shot theme (ST) and audio type (AT). Visual features are directly extracted from key-frames in shot units. To reduce computational complexity, we employ a common sampling strategy to select key-frames. Assuming a sampling step is n_t, and n_s is the number of frames in one shot. When $n_s > 3n_t$ key-frames are sampled by step n_t, otherwise the first, middle and last frame are selected as the key-frames. Using this strategy, no less than three key-frames would be selected to represent each shot, which is enough to compute features.

The first kind of features in our algorithm is SDS that depicts visual dissimilarity of topic unit boundary, and this signal is calculated by the graph partition model as work [9]. Unlike distance between low-level features, SDS considers visual information of neighboring shots in a temporal interval, so it produces a signal with local invariance. Before calculating SDS, visual distance between each two shots must be defined according to RGB histogram. To be robust to noise, the metric proposed in work [2] is used to compute the distance between two group of key-frames. The SDS must be discretized in order to input CRF++ tools [10] that are open source tools for CRF application. Since the topic boundaries are probably obtained at its local maximum and this value should be larger than median or mean, discrete feature must reflect these characteristic. On one hand, range of the signal is divided into thirteen equal subintervals to map each value in signal sequence. On the other hand, each value is also labeled by three attribute: above or below median, above or below mean and local maximum or not. Therefore, SDS is transformed into 4 dimensional discrete features to input into CRFs model.

In contrast to SDS depicting visual dissimilarity, STG [11] clusters similar shots for purpose of constructing a connecting graph that depicts repeating shot pattern in a topic unit. Such a method divides the video into a lot of segments according to the visual similarity, and the topic unit boundaries are a subset of these segment boundaries. Then, the shots can be classified into two categories: boundary shots and interior shots of segments, which are discrete STG feature. To calculate this feature, we employ the same shot difference measurement function as SDS. In clustering step, a minimum spanning tree (MST) clustering as work [11] is adopted instead of time-constraint hierarchical clustering algorithm, due to that MST can easily add time-constraint in clustering process. Then, STG can be constructed

by backward searching in same shots cluster as work [11].

Fig. 3. Examples of the predefined theme shots: anchorperson (a) and (b), headline caption (c) and (d), highlight (e).

ST is an important feature in topic unit segmentation. In this article, shots are classified into four types of theme: anchorperson, headline caption, highlight and footage, as shown in Fig. 3. Anchorperson and headline are effective features to depict TV news pattern [2, 3], as shown in (a), (b), (c) and (d) of Fig. 2. Highlight shot is an important landmark in some documentaries, when the topic moves to the next subject, as shown in (e) of Fig. 3. If a shot cannot be classified into the above categories, it will be labeled as footage shot. To classify the shot theme, discrete cosine transform (DCT) coefficients feature [12] and three parallel support vector machine (SVM) models are used. Probability of ST is calculated according to following formula: $P_{theme} = \max\{PA_k, PC_k, PH_k\},\ k \in S_i$, where PA_k, PC_k and PH_k are discriminative probability of anchorperson, headline caption and highlight respectively; k is index of key-frame of shot S_i. $P_{theme} \geqslant 0.5$ shot theme is classified to the corresponding type, otherwise shot theme is labeled by footage shot.

A change between topic units commonly accompanies a certain audio type. Therefore, AT is an effective clue that indicates a starting point of new topic unit. For example, there may be a silence or music appearing between different units. A helpful audio classification algorithm in [13] is used to classify sound type. Features are extracted from audio data across two shots during half second in each one. Then, all sounds are classified into silence, speech, music and noise by two cascaded SVMs.

2.3. Topic unit model based on CRFs

There are many articles to introduce principle of CRFs, such as [14, 15], and due to the limited space, we will not go into these details of CRFs in this paper. To implement our algorithm, we use CRF++ tools to train the model and predict labels. We use 7 dimensional features in which components of SDS, STG, AT and ST are 4, 1, 1 and 1 respectively. To train a CRFs model, each feature vector must map a tag to indicate the type of shot. In predicting process, we only input a feature vector list to obtain boundaries of topic unit between ES and BS, ES and SS or SS and SS.

Let $S = \{s_i,\ i \in n\}$ represents n labels of shot sequence, and $X = \{X_i,\ i \in n\}$ is corresponding feature vector sequence. Each X_i in X represents a group of audio and visual features that are extracted from shot i. The goal of topic unit segmentation is to maximize the number of labels s_i that are correctly classified, which need to learn an independent per-position classier that maps $X = \{X_i\} \rightarrow S = \{s_i\}$ for each shot i. The solution of CRFs to this problem is to model the conditional distribution $p(S|X)$. The probability assigned to a label sequence for a particular sequence of shots by a linear-chain CRFs is given by the equation below:

$$p(S|X) = \frac{1}{Z(X)} \exp\left(\sum_{i=1}^{n}\sum_{k=1}^{m} \lambda_k f_k\left(s_{i-1}, s_i, x_i\right)\right) \quad (1)$$

where $Z(X)$ is a normalization function:

$$Z(X) = \sum_{s_i \in S} \exp\left(\sum_{i=1}^{n}\sum_{k=1}^{m} \lambda_k f_k\left(s_{i-1}, s_i, x_i\right)\right) \quad (2)$$

function $f_k(\cdot) \in \{0,\ 1\}$ represents empirical function that depends on input variable. In theory, current label s_i can depends on the feature vector of all shots, but the feature vectors of neighboring shots are only considered in practice. In formula, k donates range of neighboring feature vectors to predict s_i. CRF++ tools use a template to control the value of k, and the detail can refer to [10]. Using $\lambda = \{\lambda_k,\ k \in m\}$ that is estimated in learning process, the maximum probability of the label sequence $S = \{s_i,\ i \in n\}$ in the condition $X = \{X_i,\ i \in n\}$ can be calculated, which $S = \{s_i,\ i \in n\}$ is desired of label sequence.

3. EXPERIMENT RESULTS

Table 1. The information of data set.

	genre	segment	time(min)	shots	Topic
Topic unit	news	6	167	2234	138
	documentary	6	160	1839	63
	total	12	327	4073	201

We choose about five and half hours data and cross-validation strategy to evaluate the performance of our method. In cross-validation, the half data set that is used to train model are not overlaps with testing data. The data set contains two kinds of program: documentary and news. Table 1 summarizes the information of data set. For each video, ground-truths of the topic unit boundaries are obtained by a human observer in accordance with definition in work [1]. Recall, precision, and F-measure are selected following the work [1] to evaluate the performance. In addition to use CRF++ tools, we also implement our method and comparative methods by C++ language and Opencv tools.

3.1. Impact of template ranges on performance

As mentioned above, the template is employed to control the range of neighboring features used to predict current label in CRF++ tools. In the first experiment, we compare the results using different templates on topic unit segmentation. In Fig. 4 (a), the performances of our algorithm are respectively presented by varying the template from 1 to 4. It can be observed that the algorithm yields better results with template range increasing. The probable reason for this phenomenon is that large template provides more context information on predicting labels than small one. Taking shot 3 in Fig. 1 (a) for example, when the template equal to 1, only the features from shot 2 and shot 4 can be used to predict this label. However, if the template increase to 2, the features of shot 1 and shot 5 also are employed to predict its label. Therefore, the large template is more possible to contain important clues to identify shot types.

3.2. Impact of features on performance

Having examined the performance of different templates, we then compare the impact of different features on topic unit segmentation. Performance of different features in topic unit segmentation is given in Fig. 4 (b): 1- STG, 2- SDS, 3-AT, 4-STG, SDS and AT, 5-ST, 6-STG, SDS and ST, 7-AT and ST, 8-STG, SDS, AT and ST. From

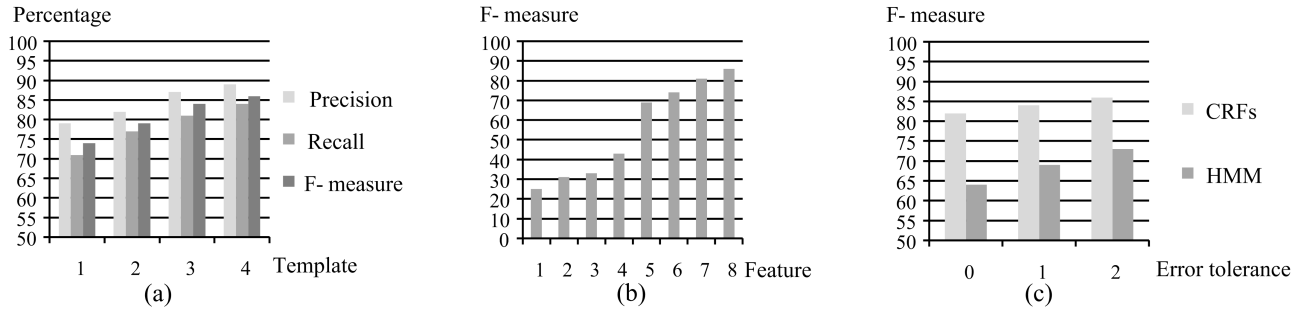

Fig. 4. (a) Result of topic unit segmentation using different template in CRFs model; (b) Result of topic unit segmentation using different features; (c) Comparative results (using F-measure) with different error tolerance on topic unit segmentation.

it, we can learn that STG and SDS are less effective than AT, and furthermore the most effective feature is ST. Reasons of the above phenomenon are following: STG and SDS only depict visual distance of neighboring shot. However, a new topic unit usually begins with silence and music, and thus AT is more useful than STG and SDS. Since ST indicates the key point in the structure of topic unit, it is the most effective feature in the segmentation. It is concluded that performance of CRFs model depends on the effectiveness of features. Fortunately, CRFs model has ability to accept a large number of input features for prediction.

3.3. Comparison with HMM model on topic unit segmentation

Table 2. Comparative result with HMM model on topic unit segmentation using precision, recall and F-measure.

	CRF			HMM		
	Precision	Recall	F-measure	Precision	Recall	F-measure
news01	100%	89%	**94%**	88%	79%	83%
news02	100%	**100%**	**100%**	82%	75%	78%
news03	**96%**	**93%**	**95%**	85%	79%	82%
news04	**100%**	**94%**	**97%**	77%	72%	74%
news05	**100%**	**96%**	**98%**	88%	85%	86%
news06	**100%**	**100%**	**100%**	83%	77%	79%
documentary01	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**
documentary02	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**
documentary03	67%	60%	63%	55%	47%	50%
documentary04	71%	57%	63%	43%	50%	46%
documentary05	69%	55%	61%	45%	49%	47%
documentary06	66%	58%	62%	46%	46%	46%
average	89%	84%	86%	74%	72%	73%

In this section, we compare the segmentation results with HMM model. Because there is not a publicly available dataset in topic unit segmentation and features in our method are quite different with other methods, we only compare results with HMM model based on same features. As shown in Table 2, performance of CRFs model on news has an obviously improvement. This is because CRFs model predicting labels uses more context information than HMM model that only use last state and the current features to predict the label according to Markov assumption.

Testing set of documentaries are two programs that contain six video segment. The first two video come from one program, while the last four come from another. In the first program, there is highlight shot in ST features between different topic units, as shown in (e) of Fig. 3. Such a program has excellent results on both CRFs model and HMM model, as shown in Table 2. In contrast, highlight shot does not exist in the second program. Although the performance of CRFs model is also superior to HMM model, it has obviously decline in comparison with the first program. This phenomenon further prove CRFs model depending feature efficiency on label production.

3.4. Accuracy of topic unit boundaries

In the last experiment, we compare accuracy with different methods. In topic unit segmentation, a reasonable boundary error tolerance is no more than two shots. Average length of the shots near two seconds, so average of boundary error is no more than five seconds, which is acceptable according to audience experience. However, if a method can provide more accurate results, the feeling of audience in browsing video will be better. In the experiment, we find that CRFs model tends to miss the boundaries, but accuracy of right labels is better than other methods, as shown in Fig. 4 (c). When error tolerance becomes more rigorous, there is a slight F-measure decline using CRFs model. The reason of this phenomenon is that CRFs model chooses the global optimum solution during inference.

4. CONCLUSION

In this paper a novel topic unit segmentation method, making use of CRFs technique, is presented. As one contribution of this method, the algorithm is developed for topic unit segmentation by transforming it into label identifying problem that reveals the semantic structure of topic unit. Another contribution of this method is that CRFs modal is exploited to identify label sequence. Since the context information of neighboring shots is taken into account, CRFs model delivers accurate results. In experiment, our method is successfully validated on documentary and news, while the encouraging experimental results demonstrate its effectiveness to topic unit segmentation.

5. REFERENCES

[1] Christian Petersohn, "Logical unit and scene detection: a comparative survey," in *Multimedia Content Access: Algorithms and Systems II*, 2008, vol. 6820, pp. 02–17.

[2] Gao Xinbo and Tang Xiaoou, "Unsupervised video-shot segmentation and model-free anchorperson detection for news

video story parsing," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 12, no. 9, pp. 765–776, 2002.

[3] Wang Ce, Wang Yun, Liu Hua-Yong, and He Yan-Xiang, "Automatic story segmentation of news video based on audio-visual features and text information," in *Machine Learning and Cybernetics, 2003 International Conference on*, 2003, vol. 5, pp. 3008–3011 Vol.5.

[4] Lekha Chaisorn, Tat-Seng Chua, and Chin-Hui Lee, "A multi-modal approach to story segmentation for news video," *World Wide Web*, vol. 6, no. 2, pp. 187–208, 2003.

[5] Lei Xie, Jia Zeng, and Wei Feng, "Multi-scale texttiling for automatic story segmentation in chinese broadcast news information retrieval technology," vol. 4993 of *Lecture Notes in Computer Science*, pp. 345–355. 2008.

[6] Jinqiao Wang, Lingyu Duan, Qingshan Liu, Hanqing Lu, and J.S. Jin, "A multimodal scheme for program segmentation and representation in broadcast video streams," *Multimedia, IEEE Transactions on*, vol. 10, no. 3, pp. 393 –408, 2008.

[7] Bailan Feng, Peng Ding, Jiansong Chen, Jinfeng Bai, Su Xu, and Bo Xu, "Multi-modal information fusion for news story segmentation in broadcast video," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, 2012, pp. 1417 –1420.

[8] Yuan Jinhui, Wang Huiyi, Xiao Lan, Zheng Wujie, Li Jianmin, Lin Fuzong, and Zhang Bo, "A formal study of shot boundary detection," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 17, no. 2, pp. 168–186, 2007.

[9] Ufuk Sakarya and Ziya Telatar, "Video scene detection using graph-based representations," *Signal Processing: Image Communication*, vol. 25, no. 10, pp. 774–783, 2010.

[10] Taku Kudo, "Crf++: Yet another crf toolkit," 2005.

[11] Minerva Yeung, Boon-Lock Yeo, and Bede Liu, "Segmentation of video by clustering and graph analysis," *Computer Vision and Image Understanding*, vol. 71, no. 1, pp. 94–109, 1998.

[12] Zhong Yu, Zhang Hongjiang, and A. K. Jain, "Automatic caption localization in compressed video," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 22, no. 4, pp. 385–392, 2000.

[13] Ying Li and Chitra Dorai, "Svm-based audio classification for instructional video analysis," in *Acoustics, Speech, and Signal Processing, 2004. Proceedings. (ICASSP '04). IEEE International Conference on*, 2004, vol. 5, pp. 897–900.

[14] C. Sutton and A. McCallum, "An Introduction to Conditional Random Fields," *ArXiv e-prints*, 2010.

[15] Roman Klinger, Katrin Tomanek, and Roman Klinger, "Classical probabilistic models and conditional random fields," 2007.