MULTI-MODAL INFORMATION FUSION FOR NEWS STORY SEGMENTATION IN BROADCAST VIDEO

Bailan Feng[1], Peng Ding[1], Jiansong Chen[1], Jinfeng Bai[1], Su Xu[1], Bo Xu[1.2]

[1]Digital Content Technology Research Center, Institute of Automation
[2]National Lab of Pattern Recognition, Institute of Automation
Chinese Academy of Sciences, Beijing 100190, China

ABSTRACT

With the fast development of high-speed network and digital video recording technologies, broadcast video has been playing a more and more important role in our daily life. In this paper, we propose a novel news story segmentation scheme which can segment broadcast video into story units with multi-modal information fusion (MMIF) strategy. Compared with traditional methods, the proposed scheme extracts a wealth of semantic-level features including anchor person, topic caption, face, silence, acoustic change, audio keywords and textual content. Parallel to this, we make use of a multi-modal information fusion strategy for news story boundary characterization by joining these visual, audio and textual cues. Encouraging experimental results on News Vision dataset demonstrate the effectiveness of the proposed scheme.

Index Terms—News Story Segmentation, Anchor Person Detection, Topic Caption Detection and Track, Audio Detection, Broadcast Video

1. INTRODUCTION

With ever-increasing broadcast channels, we are exposed to overwhelming amounts of news nowadays. As Chua and Chang's statement [1] that "Before these news videos can be conveniently accessed, there is a need to organize the video into meaningful units based on story, which conveys cohesive information about one topic.", the effective story segmentation of the broadcast video is essential to support a variety of user-oriented functions, including the browsing, retrieval and personalization of news video.

From the perspective of the segmentation techniques, the current story segmentation techniques can be basically divided into three broad categories: a) text-based methods, b) heuristic rule-based methods and c) model-based methods.

Text-based methods aim to employ the ASR script of news video and explore the text tiling information to perform story segmentation [2-3]. However, the ASR script is not always reliable, and the intrinsic cohesive of text alone still remains concealed to determine the precise boundary.

Heuristic rule-based methods are generally performed by making use of certain domain knowledge and other cues to explore the regular pattern of news story [4-5]. Nevertheless, all the existing methods above only explore the partial domain knowledge which can only be suitable for limited programs. Otherwise, the accuracy of each middle-level feature restricts its further extension as well.

The work was supported by "The National Key Technology R&D Program", grant No. 2009BAH40B03.

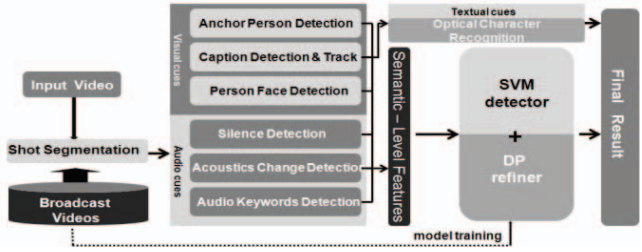

Fig.1 *The framework of MMIF for news story segmentation*

In variance from heuristic rule-based methods, model-based methods are commonly data-driven and attempt to explore the use of fully automated machine learning techniques incorporating multimodality features [6-7]. For example, [6] employs a SVM-based method to identify program boundaries using a variety of low-level multimodal features. Nevertheless, the framework in [6] is program-oriented which is not suitable for story segmentation. Besides, the accuracy of various features extraction is also critical for the final story boundary detection.

Motivated by these analyses, in this paper, we propose a novel news story segmentation scheme under model-based method, named multi-modal information fusion (MMIF) for news story segmentation in broadcast video, to cope with the problems mentioned above. To summarize, the main contributions of this paper are twofold:

1. We extract a wealth of semantic-level features based on visual, audio and textual cues with high quality. All of these can contribute to a significant story segmentation improvement.
2. We propose a novel scheme naming MMIF for news story segmentation in broadcast video, which includes a detector and a refiner. The scheme can integrate with above semantic-level features and enhance the segmenting accuracy and recall gradually.

2. MMIF FOR NEWS STORY SEGMENTATION

2.1. Overall Framework

As illustrated in Fig.1, our solution combines visual, audio and textual cues to identify the boundaries of news story and catalog the story contents. We firstly estimate the Uniform

LBP to detect shot boundaries, and extract the middle frame of each shot as the key-frame. Secondly, multi-modal features are extracted from each key-frame to represent the story characteristics. These features are all semantic-level ones including anchor person, topic caption, face, silence clip, acoustic change, and audio keywords etc., which can effectively capture the style content of story boundaries. Finally, a SVM-based detector is trained based on these features to identify the candidate boundaries of each story and a dynamic programming (DP) refiner is further carried out to determine the accurate boundaries. Besides, the topic caption of each story is also recognized by OCR module for cataloging the content.

2.2. Multi-Modal Semantic-Level Features

2.2.1. Visual cues—anchor person

News video is usually obtained by linking each anchor shot with all successive news report shots until another anchor shot, or the end of the news video occurs. Hence, anchor person detection can be regarded as an important aspect for news story segmentation. In our framework, we take an unsupervised strategy for anchor person detection as follows:

Step1: Anchor-candidate key-frame extraction and filtration. To keep the low computational cost, we firstly extract the middle frame of each shot as the corresponding key-frame for anchor shot detection. Secondly, we remove the frames without any person face from the candidate collection by face detection [8], due to that whose key-frame does not contain any faces should not be considered as an anchor person frame either.

Step2: Anchor key-frame clustering. A graph-theoretical cluster (GTC) analysis [4] is used for the anchor key-frame clustering. The anchor-candidate key-frames are considered as vertices in the GTC, and then certain similarity metric is necessary to associate a weight to each edge connecting two vertices. From repeated observation, we find that the anchor dress style and the studio background are always appearing consistency in an anchor collection. To capture these local and global traits, on one hand, we take the 4*4 block histograms of [9] as the global strategy to represent the studio background cue, which is formulated as follows:

$$Dis_{i,j}^{global} = \sum_{k=1}^{4*4} \min(b\_his_k(i,j))_1^8 \quad (1)$$

The global distance $Dis_{i,j}^{global}$ between i and j key-frames is defined as the sum of the histogram differences of the largest 8 sub-regions color differences. On the other hand, we extract the trunk region of each anchor candidate, and take the SIFT descriptors of each trunk region as the local strategy to represent the dress style cue. Trunk region is extracted by firstly calculating the face coordinate and then estimating the position of trunk with the following 1.5 times area below. Subsequently, the local distance between two key-frames is calculated as below:

$$Dis_{i,j}^{local} = \max\cos_{bow}(i,j) \quad (2)$$

where $maxcos_{bow}(i,j)$ means the maximum cosine distance of the bag of words of the corresponding trunk region between key-frames i and j. With the representations above, a linear fusion is used to combine the global and local cues, and the weight of each edge is calculated as below, where α is a parameter in [0,1), leveraging the relative contributions to the final distance from the global part to the local part, and in our method, we empirically set α to 0.7.

$$Dis_{i,j}^{anchor} = \alpha Dis_{i,j}^{global} + (1-\alpha) Dis_{i,j}^{local} \quad (3)$$

Step3: Anchor category selection. After the clustering process, a three-level pruning strategy of anchor category selection is used to pick up the real anchor categories. Firstly, the size of the category smaller than 4 will be excluded, since that one person appears less than 4 cannot be recognized as an anchor person empirically. Secondly, we notice an interesting phenomenon that the anchor shots often repeatedly appear along the whole video, and therefore the categories whose life-time exceeds certain value will be excluded as well. Besides, an anchor key-frame usually includes one or two person at most, and the number of person face goes beyond that range will be eliminated finally. Then the rest of the categories are belonging to the anchor ones, which will be used for the news story segmentation.

2.2.2. Visual cues—topic caption

Topic caption is one kind of artificial text in news video, which generally appears in the bottom of the frames. Different from other common texts, topic caption is an important symbol in news story. That is to say, a news story can only correspond with one unique topic caption both in content and form. Therefore, if we can exploit the location and the duration of each topic caption in the corresponding video, the candidate interval of each news story will basically arise, which is critical to the final story segmentation. In our framework, we propose one topic caption detection and track strategy, which consists of the following three steps:

Step1: Text spatial location. As a kind of artificial caption which is added by post-editing, topic caption in news videos generally appears in certain specific area (commonly in the 1/4 bottom region). Hence, a two-level text location strategy is performed based on above prior knowledge. Firstly, gray-scale difference statistics is used for the row location of text region. Secondly, the column location of text region can be performed by vertical projection in above row sub-region, and the intersection region is regarded as the textual region.

Step2: Topic caption region recognition. With the location process above, we obtain the textual regions of the corresponding video, which include not only topic captions but also various common texts. To separate the topic captions from noisy ones, we choose an adaptive k-means clustering algorithm [10] for the fast topic caption clustering, and one program-based topic caption example is configured beforehand for topic caption region recognition by example-to-category matching.

Step3: Topic caption temporal track. In order to meet the requirement of exploring the duration of each topic caption, we calculate each two temporally neighboring frames of the topic caption collection based on local region descriptor matching as follows:

$$Sim(A,B)=\min_{\|V_{AB}\|\le d_{th}}\left\{\frac{\sum_i \min_{p_j^B\in H_A} d_{\text{M-dist}}(p_i^A+V_{AB},p_j^B)+\sum_j \min_{p_i^A\in H_B} d_{\text{M-dist}}(p_i^A+V_{AB},p_j^B)}{n+m}\right\} \quad (4)$$

where *Sim(A,B)* is the gray similarity of topic caption sub-regions *A* and *B*. $H_A=[p_1^A, p_2^A,..., p_n^A]$ and $H_B=[p_1^B, p_2^B,..., p_m^B]$ are the Harris corner location of *A* and *B* respectively, and p_i is the coordinates point. V_{AB} is a translation vector from O^A directing to O^B, and O^A and O^B denote the origin of *A* and *B*. After this module, we can gain the location and the duration of each topic caption in the corresponding video which will be used for the news story segmentation.

2.2.3. Visual cues—person face

The various properties of person face can contribute to the news story characteristics as well. And in our framework, we also add commonly used face-related properties in visual aspect, which include face number, face position, face size and face coordinate based on the existing algorithms [8].

2.2.4. Audio cues—silence

In common sense, silence segments are mostly present in regular news videos, especially at the story transitions. To capture this audio cue, in our framework, shot time peak-valley energy [11] is used for the detection of audio clips endpoint, and we propose a dynamic threshold method to determine the audio silence. We firstly calculate the average energy $EN_{Average}$, relative energy *ENR(i)* and average relative energy $ENR_{Average}(i)$ of the whole audio clips respectively. Then if $EN(i)>EN_{Average}$, or $ENR(i)<0.8*ENR_{Average}(i)$, the frame *i* is regarded as one audio frame, else others are all belonging to silence ones. After a silence frame smooth, certain length audio can be segmented as one silence clip finally, and in our framework, we empirically consider the minimum silence segment lasting for at least 0.5s.

2.2.5. Audio cues—acoustic change

The purpose of acoustic change detection is to discover the change point of different speaker acoustic clips, which consists of the candidate news story boundaries in high probability. According to [12], we take a two-level unsupervised speaker-based segmentation for acoustic change detection. Briefly, the acoustic change detection process is divided into two levels: region level and boundary level. A modified generalized likehood ratio (MGLR) metric is firstly employed to search for the potential speaker change regions in continuous local windows at the region level. A bayesian information criterion (BIC) algorithm is secondly performed to refine the acoustic change boundaries within the potential windows. Due to the limited space, more details can refer to our previous work [12], and in this framework, the length of windows above is set to be 5s experimentally.

2.2.6. Audio cues—audio keywords

Some keywords in audio are also important cues and can help to indicate the story transitions. For example, "*This is CCTV Radio*", "*BTV News with…*" are all the common beginning keywords mostly appearing in news stories. Therefore, we resort to the Mandarin keyword spotting (KWS) system in our previous work [13] for these audio keywords detection. According to the needs of system, we have used two-week Chinese broadcast news videos as the training data, which includes 18 kinds of programs. Besides, we provide a keywords list for each program in advance as an initial configuration, and then apply the KWS to explore the audio keyword points of the corresponding program.

2.2.7. Textual cues—textual content

Commercial OCR is directly used to extract the textual contents from above topic caption collection, and the textual scripts are simply organized for the catalog of stories. But the performance of OCR in video is not satisfactory till now, which restricts its further content analysis.

2.3. News Story Segmentation

As Chua's prediction that "It is advantageous to employ rigorous machine learning techniques, along with judicious use of full multi-modality features, to achieve good segmentation performance." [1], we introduce a supervised news story segmentation scheme by combining our multi-modal semantic-level features with machine learning techniques.

Firstly, to guarantee a high-recall performance, a boundary detector for each news program is trained using above totally 27 dimensional semantic-level features (for the shots *n*, *n*-1 and *n*+1, their features are 9 dimensional respectively, the difference between each two neighboring shots is cumulative to form a 27-dimension in total). As a solid approach of a variety of supervised learning ones, SVM is chosen as our model. The RBF kernel is used for model training, and the optimal parameters are obtained from grid-based search within predetermined ranges. With this detector, we can receive a high-recall candidate segment point collection by ranking those shot boundary points who does not appear in any topic caption duration, and selecting out the appropriate top ones based on the empirical criterion of 1.2 times standard story amount of each program.

Secondly, to further obtain a high-accuracy performance, a dynamic programming (DP) based refiner is employed to improve an optimal segmentation result. We transform the candidate segment point verification to HMM decoding problem, and use the Viterbi algorithm to smooth the detection results. All the anchor point, the first topic caption point of each duration, silence point, acoustic change point and audio keyword point detected in the previous sections are considered as the states in the HMM framework. Apart from these, a null state is added too, which means it does not exist a story boundary there. The observation probability of each state comes from the probability output of each module,

and the null state is set as 0.03 experimentally. The transfer probability is simply set as 1/6, which implies that every state has the same possibility to transfer to other states. After the Viterbi process, those points falling on null state are considered as noisy ones and be removed from the candidate collection.

3. EXPERIMENTS

We conduct the MMIF scheme on News Vision dataset, which is a Chinese broadcast television collection updating each day. We have annotated two-month data from Jul. 2011 to Sep. 2011, containing both 9 China Central TV programs and 9 China Local TV ones, totally about 450 hours. The first two-week data is used for model training, and the rest is for testing.

We firstly evaluate the detecting performance of two important cues, anchor person and topic caption, and the evaluation criterion is F_1*-measure*, i.e., the harmonic mean of precision and recall, calculated as follows:

$$F_1 = 2 * Precision * Recall / (Precision + Recall) \quad (5)$$

Table 1 reports the average performances of both anchor person and topic caption, with F_1 of 0.959 and 0.997 respectively, and from the table we can conclude that the semantic-level features embedded in our scheme are of high quality.

Tab.1 *the average detecting performance of anchor person and topic caption in MMIF scheme*

Anchor Person Detection			Topic Caption Detection		
Recall	*Precision*	F_1	*Recall*	*Precision*	F_1
0.967	0.986	0.959	0.995	1	0.997

Secondly, we further evaluate the news story segmentation performance with above F_1*-measure* criterion, and a detected story boundary is considered correct if it lies within a 10-second tolerant window on each side which is consistent with TRECVID criterion. We measure the performance of two segmentation strategies:

MMIF with SVM detector: *segmenting news stories based on our MMIF using SVM detector only.*
MMIF with SVM detector + DP refiner: *refining the above result based on the DP refiner.*

The average performance over 18 programs is shown in Figure 2. From the figure, we can see that by combining our high-quality multi-modal semantic features, MMIF with SVM detector can receive an average F_1 performance of 0.744, which shows the important role of semantic features.

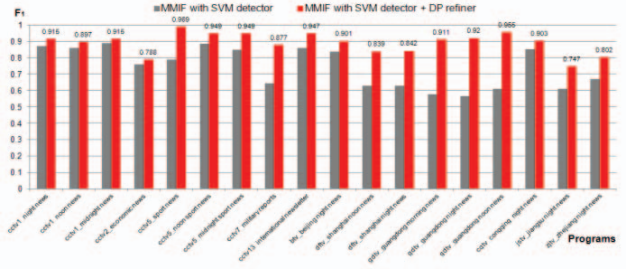

Fig.2 *Segmentation results of MMIF on 18 programs*

By further appending DP refiner to the framework, the MMIF with SVM detector + DP refiner strategy gets a higher F_1 performance of 0.891 on average, which is able to be applied to the practical engineering application*. Both observations above demonstrate the effectiveness of our MMIF scheme for news story segmentation.

4. CONCLUSIONS

In this paper, we have presented a novel multi-modal information fusion scheme for news story segmentation. Unlike conventional methods, it extracted a wealth of semantic-level features with high quality, and then made use of a multi-modal information fusion strategy by joining these visual, audio and textual cues for news story segmentation and cataloging. Our idea has been successfully tested on the News Vision dataset and experimental results demonstrate its practical engineering effectiveness.

5. REFERENCES

[1] T.S.Chua, S.F.Chang, L.Chaisom, W.Hsu, "Story Boundary Detection in Large Broadcast News Video Archives – Techniques, Experience and Trends," *ACM MultiMedia*, pp.656-659, 2004.

[2] L.Xie, J.Zeng, W.Feng, "Multi-Scale TextTiling for Automatic Story Segmentation in Chinese Broadcast News," *In: Proceedings of LNCS*, pp.345-355, 2008.

[3] L.Xie, Y.Yang, "Subword Lexical Chaining for Automatic Story Segmentation in Chinese Broadcast News," *In: Proceedings of PCM,* pp248-258, 2008.

[4] X.Gao, X.Tang, "Unsupervised Video-Shot Segmentation and Model-Free Anchorperson Detection for News Video Story Parsing," *IEEE Trans. CSVT*, 12(9), pp.765-776, 2002.

[5] H.Lee, J.Yu, Y.Im, J.M.Gil, D.Park, "A Unified Scheme of Shot Boundary Detection and Anchor Shot Detection in News Video Story Parsing," *Multimedia Tools and Applications*, 51(3), pp.1127-1145, 2011.

[6] J.Q.Wang, L.Y.Duan, Q.S.Liu, H.Q.Lu, J.S.Jin, "A Multimodal Scheme for Program Segmentation and Representation in Broadcast Video Streams," *IEEE Trans. Multimedia*, 10(3), pp.393-408, 2008.

[7] G.J.Poulisse, M.F.Moens, "Multimodal News Story Segmentation," *In: Proceedings of IHCI*, pp.95-101, 2009.

[8] M.H.Yang, D.Kriegman, N.Ahuja, "Detecting Faces in Images: A Survey," *IEEE Trans. PAMI*, 24(1), pp.34-58, 2002.

[9] M.D.Santo, P.Foggia, G.Percannella, C.Sansone, M.Vento, "An Unsupervised Algorithm for Anchor Shot Detection," *IEEE ICPR*, pp.1238-1241, 2006.

[10] R.Maitra, A.D.Peterson, A.P.Ghosh, "A Systematic Evaluation of Different Methods for Initializing the K-means Clustering Algorithm," *IEEE Trans. KDE*, 2010.

[11] S.N.He, J.B.Yu, "A Novel Chinese Continuous Speech Endpoint Detection Method Based on Time Domain Features of the Word Structure," *IEEE ICCCAS*, pp. 992-996, 2002.

[12] S.L.Zhang, S.W.Zhang, B.Xu, "A Two-level Method for Unsupervised Speaker-based Audio Segmentation," *IEEE ICPR*, pp.298-301, 2006.

[13] J.E.Liang, M.Meng, X.R.Wang, P.Ding, B.Xu, "An Improved Mandarin Keyword Spotting System Using MCE Training and Context-Enhanced Verfication," *IEEE ICASSP*, pp.1145-1148, 2006.

*This work has been applied to the News Vision System under Chinese SARFT program, and has gained an excellent performance and user experience.