Automatic Parsing of News Video Using Multimodal Analysis*

WANG Wei-qiang[1], GAO Wen[1,2]

[1](*Institute of Computing Technology, The Chinese Academy of Sciences, Beijing* 100080, *China*);

[2](*Department of Computer Science and Engineering, Harbin Institute of Technology, Harbin* 150001, *China*)

E-mail: wqwang@ict.ac.cn

http://www.ict.ac.cn

Received October 24, 2000; accepted February 22, 2001

Abstract: The paper presents an approach, which exploits multimodal information (video, audio and text) to automatically parse news video. In the paper, audio features extraction, as well as multimodal information integration scheme, are addressed in detail. Integration of multiple information sources can overcome the weakness of the approach only exploiting the image analysis techniques. That makes our approach have wider adaptation to variable existence situations of news items. On test data with 184 100 frames, when the system detects bound aries between news items, the recall 95.1% and the accuracy 93.3% are obtained. The experiment results show the approach is valid and robust.

Key words: MPEG-2 video; automatic segmentation of news items; audio and visual information analysis; anchor shot; caption detection

As more and more video becomes available, it is significant to efficiently manage the content of the video, so as to provide the support for other applications. To characterize its content, video structure parsing is required for indexing. Many literatures have addressed the shot boundary detection techniques, such as [1~3]. Some researchers have explored scene extraction algorithms. For instance, Refs. [4,5] present the similar approaches. In their approach, key frames are chosen for each shot first. Then the key frames are clustered and each cluster is given a label. At last, scenes are identified through analyzing the repetition pattern of the labels on the temporal axis.

Due to the state-of-the-art of machine vision and signal analysis, automatic extraction of high level semantic structure, such as scene, story, is difficult to implement for general video programs. But since the temporal syntax of a news video is commonly very straight, some priori knowledge can help to identify high level semantic structures accurately. Zhang *et al.*[6] proposed an approach of parsing news video based on image analysis. The kernel of their system is the algorithm that locates and identifies anchorperson shots. Due to their assumption that each news item starts with an anchor shot followed by a sequence of news shots, their system can not identify the new items that are only read by anchorpersons without news shots, as well as start without an anchorperson shot. The limitation can not be overcome by analyzing only visual signals.

* Supported by the National Natural Science Foundation of China under Grant No. 69789301 (国家自然科学基金); the National High Technology Development Program of China under Grant No. 863-306-ZT03-01-2 (国家863高科技发展计划)

WANG Wei-qiang was born 1972. He is a Ph.D. student at the Institute of Computing Technology, the Chinese Academy of Sciences. His current research interests are multimedia techniques and artificial intelligence. **GAO Wen** was born in 1956. He is a professor and doctoral supervisor of the the Institute of Computing Technology, CAS, and Department of Computer Science, Harbin Institute of Technology. His current research areas include multimedia data compression, image processing, computer vision, multimodal interface, artificial intelligence, virtual reality.

Recently, more literatures are seen to apply audio analysis techniques in characterizing video content. Reference [7] exploited multiple audio features and a neural net classifier to differentiate five classes of TV programs, including advertisement, basketball, football, news, and weather. Reference [8] proposed a heuristic rule-based approach for the segmentation and annotation of generic audio data. Audio recordings are segmented and classified into basic audio types such as silence, speech, music, environmental sound, etc.[5,10] combined audio and visual features to detect shot boundaries. Based on audio and visual features, Ref. [11] applied the HMM classifier to video scene segmentation and classification. In this paper, we present an approach, which integrates visual, audio and text information to automatically parse news video. It effectively overcomes the limitation of the approach in Ref. [6]. Therefore the automatic segmentation of news items will be discussed in detail by the paper. The resulting system was tested on CCTV news, so we will focus the discussion of the techniques on them.

The rest of the paper is organized as follows. Section 1 first overviews the whole system. That is followed by the discussion of shot segmentation and two important event detection, i.e., anchor shots and caption text. Then audio analysis is presented in detailed. At the end of the section, the approach of news item extraction through information fusion is described. In Section 2, experiment results and analysis are given. Section 3 concludes the paper.

1 Parsing of News Video

The parsing system consists of five function modules, as shown in Fig. 1. The shot segmentation module segments a video stream into a sequence of shots. The anchor shot detection module identifies the appearance and disappearance event of anchorperson frames, so as to locate the clips composed of anchor shots. The caption text detection module identifies the clips whose frames are overlapped by captions. In news video, captions are often overlapped on original natural video, which forms concise annotation of relevant video clips. Accurate identification of them not only helps fast browsing of the news content, but also provides an important cue for parsing the news programs. Audio, as another time-dependent media in a video document, can supplement visual information, and supply a unique cue for video content analysis. For instance, visual content is almost unchanged for anchor shots, but it is possible that multiple news items are being reported by an anchorperson. The audio analysis module categorizes audio into the classes, including music, speech and silence. Then silence segments are used to segment news items as one of the important cues. The kernel of the system is the integration module. Based on priori knowledge and temporal structure models, the module combines the information from multiple channels to analyze high-level structures and output them.

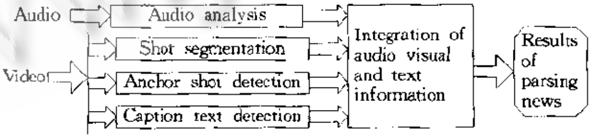


Fig. 1 The block diagram of our system

1.1 Shot segmentation and anchor shot detection

Generally, TV news involves two types of shots, i.e., anchor shots and news shots. The former only uses audio to present news content while the latter uses the synchronized visual and audio streams to do that. Therefore, if an anchor shot involves multiple news items, vision-based analysis can not provide enough cues for news items segmentation. For a clip made up of multiple news shots, if a news item in the clip is transferring to another one, the boundary between the two news items must also be a shot boundary. In this case, shot segmentation is significant, since it will provide potential candidates of new item boundaries. The shot segmentation modules can output useful information of shot boundaries.

Anchor shot detection is another significant aspect of our system. The detection results are used to determine which channels are exploited to extract news items. Zhang *et al.*[5] proposed a scheme to detect anchor shots. Their method is influenced by the accuracy of shot segmentation, and it works on the uncompressed data streams, involving high complexity computation. Compared with the scheme, our system exploits a fast anchor shot detection algorithm based on background choronminance and skin tone models. It does not involve shot segmentation, and operates in compression domain, and the computation complexity is very low. In evaluation experiments on a big test set, the accuracy 98.9% and the recall 100% have been obtained. Owing to the primary purpose of the paper, the details of the algorithm will be described in another paper. But we can assume accurate detection of anchor shots has been done and corresponding information has been gained.

1.2 Caption text detection

Caption present in video frames plays an important role in understanding video content. Some literatures have addressed the caption detection techniques. In Ref. [12], Smith and Kanade considered a typical text region is characterized with clustered sharp edges. Their algorithm required no priori knowledge, and can be applied to general video. But since it operated on original images, intensive decode computation resulted in its relatively low detection speed. Yeo and Liu[13] proposed a fast algorithm in MPEG compression domain. Based on the assumption that caption appearance and disappearance often occur in the middle of a video shot, a similar technique as Ref. [1] was applied to compute inter-frame content difference in caption regions. At the same time, the algorithm identified and ignored the large content difference caused by shot transitions, to locate caption appearance and disappearance events.

But our observation of CCTV news shows existence of caption text commonly covers multiple shots. So the algorithm in Ref. [13] is no longer applicable in the context. Based on statistics features of caption texts' chrominance components, we also propose a fast algorithm to automatically detect captions on MPEG compressed video. The algorithm has been tested on CCTV news, and the accuracy 96.6% and the recall 100% are achieved. Details of the algorithm have been given in Ref. [14]. The caption detection module implements the algorithm, and can supply the caption appearance and disappearance information for the kernel module.

1.3 Audio content analysis

International standard MPEG-1, 2[15,16] exploit the perception-based high performance encoding schemes to compress audio signals. For audio, the standards specify three layers of encoding schemes. Higher the layer is, more complex the encode/decode computation is, and higher compression ratio is gained. Since audio in digital TV programs is commonly encoded with layer II, we assume the audio involved in the paper is also the case. The audio elementary stream consists of a sequence of audio frames, and each frame contains a fixed number of samples, such as 1152 samples for layer II. For each audio frame, its shot time average magnitude can be calculated using the following expression,

$$M_m = \frac{1}{N}\sum_{n=0}^{N-1}|x(n)|, \tag{1}$$

where M_m represents the short time average magnitude of the frame m, $x(n)$ is the sample values in the frame, N is the number of samples in each frame. Since the synthesis of sub-band filtered data in decoding process of MPEG audio is a linear computation, the average magnitude can be approximated as the following expression,

$$M_m \approx \frac{1}{32*K}\sum_{i=0}^{32}\sum_{n=0}^{K-1}|s_i(n) , \tag{2}$$

where $s_i(n)$ is the n^{th} sample of sub-band i, K gives the number of samples in each sub-band. If $M_m < \lambda$, then the frame m is considered in the short time silence state, where λ is a threshold. For an audio stream with the sample rate of 48 KHz/s, each audio frame lasts about 24ms. To identify various types of clips, we select a relatively long interval and consider its features. Let T represent the length of the interval, and each audio frame lasts τ, thus the long interval contains about $\lceil T/\tau \rceil$ audio frames. In the system, T is chosen about 1 second.

When an anchorperson reports news, pauses with different lengths exist between words, sentences, as well as news items. Some music clips or some clips with music background also exist in the TV news program. Accurate identification of them is very helpful to parse TV news. For the purpose, two features are calculated from the relatively long intervals, i.e., the pause rate and the silence ratio. Assume $AC_i(i=0,1,\dots)$ is a long interval, af_{ij} $(j=0,1,\dots,L-1)$ represents the jth audio frame in it and $M(af_{ij})$ is the short time average magnitude of af_{ij}, we define $Tag(i,j)$ as follows.

$$Tag(i,j)=\begin{cases}1, & \text{if } M(af_{ij})\geqslant\lambda \\ 0, & \text{if } M(af_{ij})<\lambda\end{cases} \tag{3}$$

Then the pause rate and the silence ratio for the long interval can be calculated using the following expressions.

$$C(i,j)=\begin{cases}1, & \text{if } Tag(i,j)-0 \text{ and } Tag(i,j-1)=1, j\geqslant 1 \\ 0, & \text{else}\end{cases} \tag{4}$$

$$PauseRate(i)-\sum_{j=0}^{L-1}C(i,j) \tag{5}$$

$$SilenceRatio(i)=1-\Big(\sum_{j=0}^{L-1}tag(i,j)\Big)\Big/L \tag{6}$$

Based on the two features, the music clips or the clips with background music in TV news can be identified. For an audio clip, if each $AC_i(i=s,s+1,\dots,t-1,t)$ satisfies $PauseRate(i)-0$, $SilenceRatio(i)=0$, then the clip is labeled as music type. After that, the following algorithm is applied to identify pauses in speech for the remaining clips, and the pauses are considered as potential silence intervals between news items.

Algorithm 1. Selection Candidates of the Silence Intervals between News Items

```
IF ($PauseRate(i)\neq 0$)
    IF ($SilenceRatio(i)/PauseRate(i)>\alpha$)
        $AC_i$ will be chosen as a candidate silence clip
    ELSE
        $AC_i$ is not a candidate silence clip
ELSE
    IF ($SilenceRatio(i)>\beta$)
        $AC_i$ will be chosen as a candidate silence clip
    ELSE
        $AC_i$ is not a candidate silence clip
```

In the algorithm 1, α,β are thresholds and $\beta>\alpha$. Their values are related with $L*\tau$. The larger $L*\tau$ is, the less α,β become. Since we choose 1s for the length of the long interval $L*\tau$, correspondingly $\alpha=0.27$, $\beta-0.85$ are chosen experientially. For a long interval, Fig. 2 gives several typical wave images of $Tag(i,j)$, as well as the corresponding values of $PauseRate(i)$ and $SilenceRatio(i)$. After the candidates of the silence intervals are identified, the silence segment $SG(n)$ can be further determined through clustering the intervals, i.e., $SG(n)=\langle s_n,e_n\rangle$, $n,s_n,e_n=1,2,\dots$, and $s_n\leqslant e_n$, which means in the audio clip $AC_{s_n-1}AC_{s_n}\dots AC_{e_n}AC_{e_n+1}$, only AC_{s_n-1}, AC_{e_n+1}, are not the candidates of the silence intervals. We consider the clip $SG(n)$ is where the transition between news items occurs with the most probability.

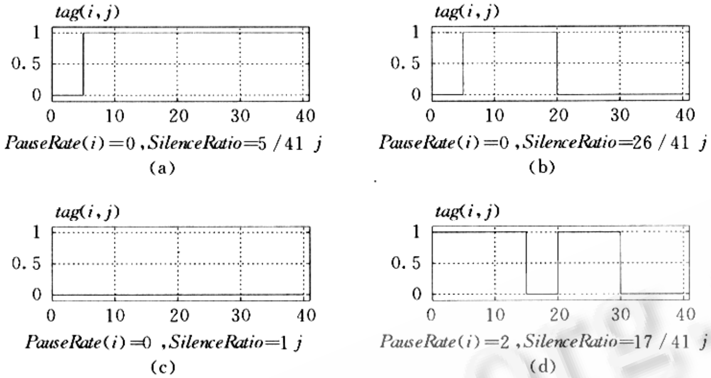


Fig. 2 Typical wave images of *Tag(i,j)*

1.4 Integration of audio, visual and text information

Through the processes described by Subsections 1.1, 1.2, 1.3, the following important information are obtained, including the temporal locations of start and end frames for shots, anchor shot clips, clips overlapped by captions, as well as music clips, silence segments etc. in the audio stream.

The whole parsing process involves three steps. First, the whole program is segmented into content units. The news video generally has some similar temporal structure pattern. Figure 3 gives a typical temporal structure for CCTV news. The accompanying audio of program head and tail are both music clips. Therefore, they can be accurately identified through audio analysis. The frames in the abstract clip all have the same spatial structure as in Fig.4. The clip can be accurately located by exploiting the similar algorithm as in Ref.[14] to verify the existence of captions in the region A. Then the rest of the program consists of two types of clips, i.e. anchor shot clips and new shot clips. They are interlaced and can be determined by identifying the anchor shot clips. The process of detecting anchor shots breaks the whole program into a sequence of clips, i.e. *Clip(k)*, $k=0,1,2,...$, and categorizes them into five classes, including program heads, abstracts, anchor shot clips, new shot clips, as well as program tails. Most of the clips are anchor shot clips or new shot clips.

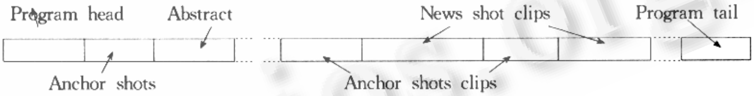
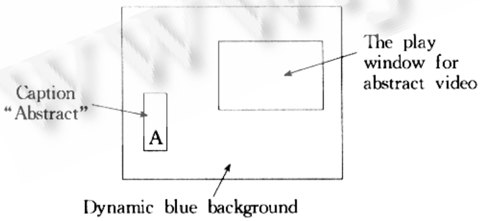
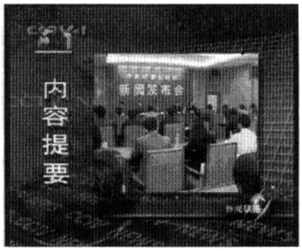


Fig. 3 The temporal structure of a typical CCTV news



(a) Sketch map



(b) An example

Fig. 4 The spatial structure model for the frames in abstract

The last two steps are used to extract news items from the anchor shot clips and the news shot clips. In the second step, some news items are determined based on caption, anchor shot and shot segmentation information. We specify introductory anchor shots and following news shots involving the same topic belong to different items. From the observation of CCTV news, we obtained the following priori knowledge, which also holds for other news video. Each news item has a corresponding caption text in news shot clips. In anchor shot clips, different anchor shots present different news items. These result in the rules 1 and 2 for extracting news items.

Rule 1: let the function $CapNum(c)$ represent the number of captions contained by a clip c. For a news shot clip $N=(cs,ct)$, where cs,ct are the start and end frame of the clip. if $CapNum(N)\leqslant 1$, then the clip forms a news item.

Rule 2: let the function $CapNum(c)$ represent the number of captions contained by a clip c. For an anchor shot clip $A=(as,at)$, where as,at are the start and end frame of the clip, if it consists of a sequence of shot $S_1S_2\ldots S_n,n=1,2,\ldots$, and for each shot $S_i,i=1,2,\ldots,n$, if $CapNum(S_i)\leqslant 1$, then the shot S_i forms a news item.

Applying the rules 1 and 2 to the output of the step 1, some clips are determined as news items. After that, the rest are clips which contain more than one caption. So the last step is to locate the news items' boundaries between them. Some characteristics can be summarized to extract news items. For news shot clips, news item transitions occur on the boundaries between shots, and the boundaries between news items are also between consecutive caption events. Furthermore, there are relatively long silence segments that cover the boundaries. For anchor shot clips, the information of captions and audio forms the essential cues. Based on the above analysis, the rules are constructed to deal with the case just mentioned. For each silence segment $SG(n)=\langle s_n,e_n\rangle$, there is a synchronized video segment $\langle V_{sn},V_{en}\rangle$, where V_{sn},V_{en} are the temporal location of the corresponding start frame and end frame.

Rule 3: let $\langle s^i,e^i\rangle,\langle s^{i+1},e^{i+1}\rangle$ are respectively the appearance and disappearance location of two consecutive caption events in an anchor shot. If the silence segment $SG(n)$'s total silence ratio $\eta=\sum_{i=S_n}^{e_n}SilenceRatio(i)$ satisfies $\eta>\zeta$, ζ is a threshold, and the corresponding video clip $\langle V_{sn},V_{en}\rangle$ satisfies $e^i\leqslant V_{sn},V_{en}\leqslant s^{i+1}$, then the frame at the position $INT((V_{sn}+V_{en})/2)$ is chosen as a boundary of news items.

Rule 4: let $\langle s^i,e^i\rangle,\langle s^{i+1},e^{i+1}\rangle$ are respectively the appearance and disappearance location of two consecutive caption events in a news shot clip. If the silence segment $SG(n)$'s total silence ratio $\eta=\sum_{i=S_n}^{e_n}SilenceRatio(i)$ satisfies $\eta>\sigma$, σ is a threshold, and the corresponding video clip $\langle V_{sn},V_{en}\rangle$ satisfies $e^i\leqslant V_{sn},V_{en}\leqslant s^{i+1}$ and covers two different shots, represented by $shot(k)$ and $shot(k+1)$, then $shot(k)$ is considered as the last shot of the foregoing news items, and $shot(k+1)$ as the first one of the following news item.

Usually the contents of TV news are compact. The silence segment $SG(n)$ in the audio channel is very short, generally less than 3.5 seconds. On the other hand, to impress the viewer and make them comfortable visually, two shot transitions within the 3.5 seconds do not occurs generally. Therefore we can assume the segment $\langle V_{sn},V_{en}\rangle$ covers two shots at most. Our observation of CCTV news has also verified the characteristics.

2 Experiments and Evaluation

We implement the system described in Section 1, and randomly choose 4-day MPEG CCTV news from video program database as test data to validate the algorithm. The whole experiment is conducted on the PC with PIII-450 CPU and 64M memory. The frame rate of test data is 24f/s with frame size of 720 * 576 pixels. The test data set contains 184 100 frames, and lasts two hours or so in total. Before testing, we manually label all the news items in the 4-day news, as a standard reference to evaluate the performance of the algorithm. When labeling, introductory anchor shots and following news shots involving the same topic are considered as different items.

Compared to accuracy, we pay more attention to recall, since the latter means less manual effort for a user to correct the results generated automatically by the system. Therefore, the following parameters values are chosen, $\alpha=0.27$, $\beta=0.85$, $\zeta=1.6$, $\sigma=1.31$, and the corresponding experimental results are tabulated in Table 1.

According to the statistics in Table 1, the accuracy of news items segmentation $P=1-\frac{E}{D}=1-\frac{7}{105}=93.3\%$ and the recall $R=1-\frac{U}{S}=1-\frac{5}{103}=95.1\%$ can be calculated. In the experiment, the less ζ, σ are chosen, the higher recall can be obtained. But the accuracy becomes less, since the higher recall also brings more false boundaries.

Table 1 The experimental results of news item segmentation

Sequences	Actual shot number	Actual item boundaries (S)	Output of item boundaries (D)	False (E)	Missed (U)
NewsA	276	33	32	0	1
NewsB	243	24	28	5	1
NewsC	305	28	26	0	2
NewsD	274	18	19	2	1
Total	1098	103	105	7	5

The analysis of the experiment results found that the missed boundaries are mainly caused by background sound occurring on the boundaries, which confuse the detection of silence segments. False segmentations also result from existence of scene sound in news shots. For instance, in an interview scene, the speaker altercation between a reporter and an interviewee can bring some silence segments in audio channels. At the same time, if a shot transition accompanies, false claims occur.

The two cases, which can not be processed by Ref. [6], are ubiquitous in CCTV news. The experiment results related to them are summarized and tabulated in Table 2. The statistics in it shows our algorithm is effective and it can accurately detect most of the boundaries in the two cases. It should be pointed out, almost all false claims in table 1 result from the effort to identify them.

Table 2 The experimental results of the two complex cases

Sequences	Actual item boundaries (S)		Boundaries detected accurately (U)	
	A	B	A	B
NewsA	3	3	3	2
NewsB	2	5	2	4
NewsC	2	3	1	2
NewsD	0	6	0	5
Total	7	17	6	13

Annotation: A——multiple items in an anchor shot
B——news items with no introductory anchor shots

3 Conclusions

The paper explores integration of audio, visual and text cues to parse the news video. The system can effectively identify news items, as well as other elements, including program tail, abstract, program tail. The frames overlapped by captions can also be accurately located, which can form a picture catalogue to support fast browsing of news content. Our algorithm of new item segmentation overcomes the limitation of the approach in Ref. [6]. That makes our approach have wider adaptation to variable existence situations of news items. The experiment results show the algorithm is valid and effective. The recall 95.1% and the accuracy 93.3% have been achieved in detecting the boundaries between news items. The experiments also imply multimodal analysis is an effective approach to parse high-level structures of video. Though the method is designed specifically for parse TV news, its analysis of audio signal, as well as integration strategy of audio-visual cues can also be applied to the scene analysis of other video classes.

References:

[1] Yeo, B.L., Liu, B. Rapid scene analysis on compressed videos. In: IEEE Transactions on Circuits and Systems for Video Technology, 1995,5(6):533～544.

[2] Boreczky, J.S., Rowe, L.A. Comparison of video shot boundary detection techniques. In: Proceedings of the SPIE Conference Storage and Retrieval for Video Databases IV. San Diego/La Jolla, CA, USA, 1996,2670:170～179.

[3] Zabih, R., Miller, J., Mai, K. A feature-based algorithm for detecting and classifying production effects. Multimedia Systems, 1999,(7):119～128.

[4] Yeung, M.M., Yeo, B.L. Time-constrained clustering for segmentation of video into story units. In: Proceedings of the International Conference Pattern Recognition (ICPR'96). Vienna, 1996. 375～380.

[5] Hanjalic, A., Lagendijk, R.L., Biemond, J. Automatically segmenting movies into logical story units. In: Proceedings of the 3rd International Conference VISUAL'99. Amsterdam (NL), 1999. 229～236.

[6] Zhang, H.J., Tan, S.Y., Smoliar, S.W., *et al.* Automatic parsing and indexing of news video. Multimedia Systems, 1995,(2):256～266.

[7] Liu, Z., Huang J., Wang, Y., *et al.* Audio feature extraction & analysis for scene classification. In: Proceedings of the IEEE Signal Processing Society 1997 Workshop on Multimedia Signal Processing. Princeton, New Jersey, USA, 1997. 343～348.

[8] Zhang, T., Jay Kuo C.-C. Heuristic approach for generic audio data segmentation and annotation. In: Proceedings of the ACM Multimedia Conference. Orlando, 1999. 67～76.

[9] Boreczky, J.S., Wilcox, L.D. A hidden markov model framework for video segmentation using audio and image features. In: Proceedings of the ICASSP'98. Seattle, 1998. 3741～3744.

[10] Nam, J., Tewfic, A.H. Combined audio and visual streams analysis for video sequence segmentation. In: Proceedings of the ICASSP'97. Munich, Germany, 1997,4:2665～2668.

[11] Huang, J., Liu, Z., Wang, Y. Joint video scene segmentation and classification based on hidden markov model. In: Proceedings of the ICME'2000. New York, NY, 2000,3:1551～1554.

[12] Smith, M.A., Kanade, T. Video skimming and characterization through the combination of image and language understanding techniques. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Puerto Rico, 1997. 775～781.

[13] Yeo, B.L., Liu, B. Visual content highlighting via automatic extraction of embedded captions on MPEG compressed video. In: SPIE Digital Video Compression: Algorithms and Technologies, 1996,2668:38～47.

[14] Wang, W.Q., Gao, Wen, Li, J.T., *et al.* News content highlight via fast caption text detection on compressed video. In: Proceedings of the International Conference on Intelligent Data Engineering and Automated Learning. Hong Kong, 2000. 443～448.

[15] International Standard ISO/IEC, 11172. Information Technology-Coding of Moving Pictures and Associated Audio for Digital Storage media at up to about 1.5Mbits/s, 1993.

[16] International Standard ISO/IEC, 13818. Information Technology-Generic Coding of Moving Pictures and Associated Audio. 1995.

基于多模式分析自动解析新闻视频

王伟强[1]， 高 文[1,2]

[1](中国科学院 计算技术研究所,北京 100080);

[2](哈尔滨工业大学 计算机科学与工程系,黑龙江 哈尔滨 150001)

摘要：提出一种结合视觉、声音、文字等多种模式信息自动解析新闻视频的方法.并对音频特征的提取以及综合多种模式信息解析新闻视频的算法进行了详细的探讨.多种模式信息的使用有效地弥补了仅基于图像分析技术分割新闻条目的不足,从而使该方法对不同方式存在的新闻条目在分割时具有更广泛的适应性.在包含184 100帧的测试数据集上,对于新闻条目边界点的检测,系统获得了95.1%查全率,93.3%的正确率.实验结果证明了该方法的有效性、强壮性.

关键词：MPEG-2视频;新闻条目自动分割;音视频信息分析;播音员镜头;标题文字

中图法分类号：TP391 **文献标识码**：A