Video Shrinking by Auditory and Visual Cues

Qianqian Xu[1], Huiying Liu[1,2,3], Shuqiang Jiang[2,3], Qingming Huang[1,2,3,*], and Yu Gong[1,2,3]

[1] Graduate University of Chinese Academy of Sciences, Beijing, 100049, China
[2] Key Lab of Intell. Info. Process., Chinese Academy of Sciences, Beijing 100190, China
[3] Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China
{qqxu,hyliu,sqjiang,qmhuang,yug}@jdl.ac.cn

Abstract. Video content is growing at an explosive rate nowadays. How to consume them efficiently is an important research point for years. Although the widely investigated video summarization solution can generate the main content of a video, it cannot ensure the coherence and apprehensibility of the original video. In this paper, we present a new framework called video shrinking to remove the video's redundant information while keeping the integrality of the video content. Firstly, speech detection is performed to extract Candidate Deletion Shots (CDS), which have the property of low speech-ratio. Then, by combining the attention analysis and continuity analysis, CDS are refined to obtain the final temporal shrinking output. Subsequently, we further shrink the video spatially to adapt for the small screens of mobile devices. Experimental results demonstrate the effectiveness and efficiency of the proposed method.

Keywords: Video Shrinking, Speech Detection, Visual Attention.

1 Introduction

With the rapid development and wide application of digital media devices, video is becoming an important part of daily life. This trend brings a challenge for the multimedia researchers. On one hand, both the amount and the resolution of video are increasing rapidly. On the other hand, users' spare time is decreasing with the speeding up of the pace of city life. Thus, there appears a gap between the huge video amount and limited watching time. At the same time, more and more users enjoy videos through mobile devices, of which the size is decreasing for portability at the sacrifice of small screens. Thus, another gap appears, between the high resolution and the limited screen size. In this paper, we aim to bridge the two gaps to provide users convenience to enjoy videos.

Video summarization is a widely used method to extract the important content of a video. For sports video, which is a popular video type, domain knowledge is usually used to extract and rank the video highlights to generate a summary [1]. For general video types, audience attention [2] and bi-directional similarity [3] can be adopted to summarize videos. Video summary facilitates users to fast grasp the main content of a video. But it can't ensure the viewing experiences for the following reasons. 1) Speech,

[*] Corresponding author.

P. Muneesawang et al. (Eds.): PCM 2009, LNCS 5879, pp. 788–796, 2009.

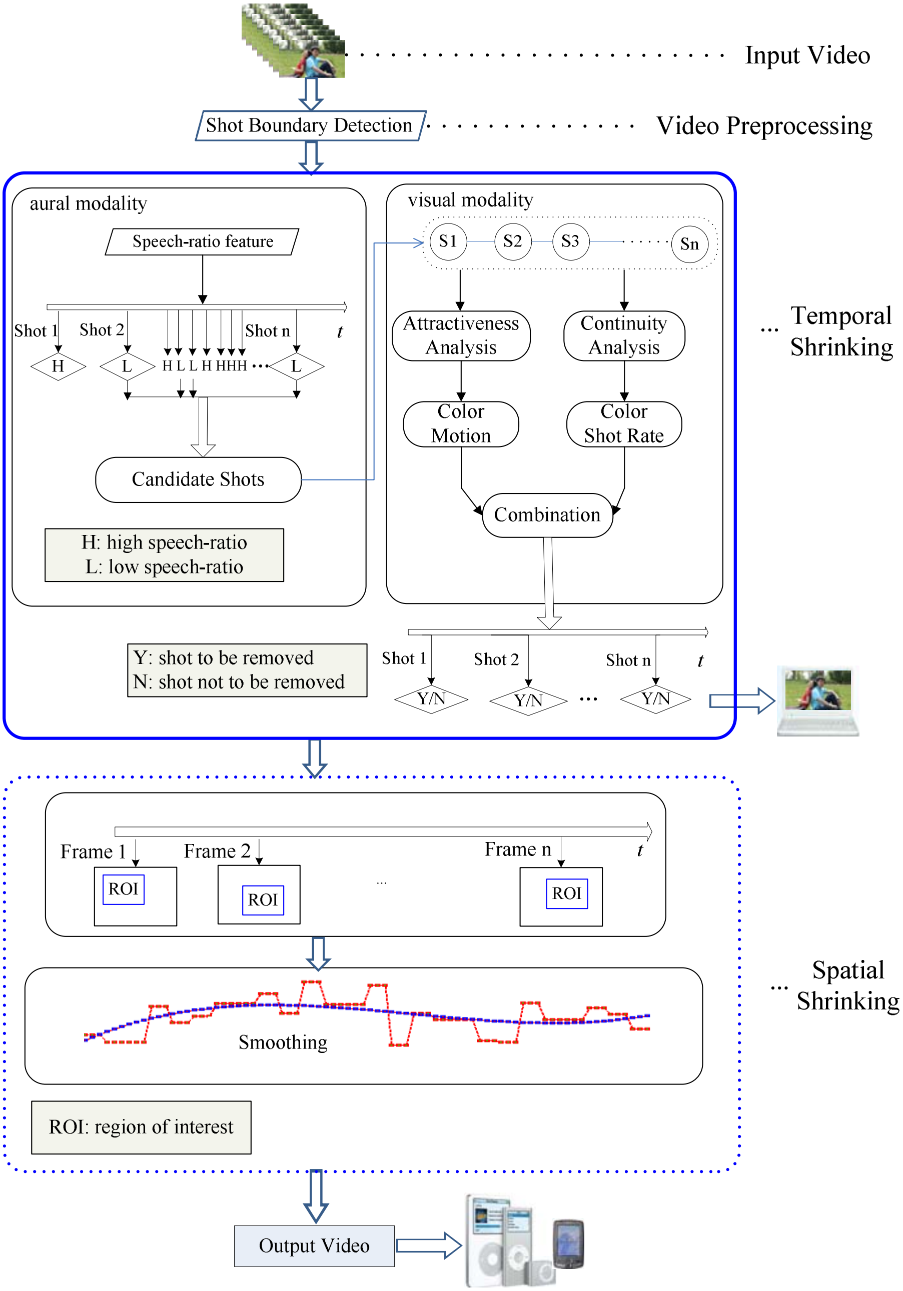

Fig. 1. The overview of our method

which is an important part of a video, is not kept complete. 2) Contextual fluency is destroyed. 3) The rhythm of the video is changed. So, in this paper, considering these factors, we propose an approach to remove temporal redundancy of a video clip while keep it informative, continual, and fluent.

While adapting a video to a smaller screen, several criteria should be considered. 1) The resulting video should contain the main content of each frame. 2) The camera motion of the resulting video should be continual to ensure audiences' viewing experiences. Researchers present several approaches to meet these qualifications. Video retargeting generates a virtual pan to scan the main content of each video frame [4]. The bi-directional summarization method [3] compresses the spare space of each video frame to reduce its resolution. These two methods reserve the video's main content but can't maintain its rhythm. The method in [2] detects the Region of Interest (ROI) of each video frame and adopts a curve fitting method to generate a fluent camera motion.

In this paper, we propose an approach, video shrinking, to facilitate users enjoying videos by reducing both temporal and spatial redundancy caused by recording or editing. The framework of the approach is illustrated in Fig.1. It is composed of three stages, video preprocessing, temporal shrinking, and spatial shrinking. We choose shot as our basic unit, as it usually contains much more information than a frame and the content within a shot is of spatial and temporal continuum. The shot boundary detection method proposed in [5] is adopted for video segmentation.

In the temporal shrinking stage, both vision and audio information are considered. In the aural part, speech is detected firstly because the shot containing speech usually has important information. In the visual part, audience attention is adopted to ensure the attractiveness of the result video. To ensure the fluency of final output, we calculate the similarity of adjacent shots using color information and shot pace differences. Based on the above analysis results, we obtain the temporal shrinking result. In the spatial shrinking stage, the method proposed in [2] is adopted to remove the spatial redundancy.

The rest of this paper is organized as follows. The main components of video shrinking, i.e. temporal and spatial shrinking are presented in section 2. In section 3, experimental results on video clips are reported. In the end, we conclude the paper in section 4.

2 Video Shrinking

In this section, we present the proposed approach of video shrinking. It contains two parts: temporal shrinking and spatial shrinking.

2.1 Temporal Shrinking

We aim to remove the temporal redundancy of a video without interfering with the viewing experience of the original video. The resulting video contains the important content of the original video, and can be watched as an independent one. Generally, the resulting video should satisfy three criteria. 1) It should contain complete sentences. 2) It should contain the attentive shots of the original video. 3) It should be contextually fluent. So, the following four modules: speech detection, attention analysis, continuity analysis and combination are included in the framework.

2.1.1 Speech Detection

Speech is an important part of a video. In our work, we present a “speech-ratio” feature to describe how much speech existing in a shot. Shots with lower speech ratio are considered as ones that have little audio information, which will be chosen as Candidate Deletion Shots (CDS) likely to be removed. We adopt SVM to train a speech detector.

Firstly, some low-level audio features are extracted from each audio frame, including spectrum power, Zero Crossing Rate (ZCR), intensity, bandwidth [6], pitch and MFCC. To all of these features, we compute the means and standard deviations in a one second long audio clip. Also, spectral flux and high zero cross rate ratio [7] are computed for each audio clip.

Based on those low-level audio features, a one-against-all SVM is trained to detect speech from other audio effects in the audio stream, and the approach in [8] is exploited to make SVM yield a probability output. The SVM outputs a confidence $C_n \in [0,1]$ for each audio clip. For shot k , let N_k be the number of audio clips in shot k , the speech ratio is extracted as:

$$R_k^S = \frac{1}{N_k}\sum_{n=1}^{N_k} C_n \tag{1}$$

2.1.2 Attention Analysis

While watching a video, audiences may pay different amount of attention to the video content at different time because the attractiveness level of each shot is different. A great deal of research has been done on estimating human attention. In [9], temporal context is taken into account because the more different a frame/shot is to the preceding ones, the more attention it will attract. In our work, we detect the attentive shots of each video by the scheme presented in [9]. In this method, the attention value of each video shot is calculated as the weighted sum of its difference to the preceding ones. The features adopted include color and motion. The attention value of a shot is:

$$AV(S_t) = \sum_{k=t-l}^{t-1} D(S_t, S_k) w(k) \tag{2}$$

where l is the length of context window which is set as 5 in our work. $w(k)$ is the weight of the shot for the consideration that the nearer shots may act more. In our work, we simply adopt linear weight $w(k) = (t-k)/(1+2+...+l)$, which describes the relative distance between the two shots. $D(S_t, S_k)$ is the feature distance, a weighted sum of color distance and motion distance:

$$D(S_t, S_k) = \alpha \times D_{color}(S_t, S_k) + (1-\alpha) \times D_{motion}(S_t, S_k) \tag{3}$$

Since histogram is a widely used global feature, it is adopted to calculate feature distance. The color distance and motion distance are calculated as:

$$D_{color}(S_t, S_k) = dis(HC_t, HC_k) = 1 - \sum\nolimits_{r,g,b} \min\left(HC_t(r,g,b), HC_k(r,g,b)\right) \tag{4}$$

$$D_{motion}\left(S_t, S_k\right) = dis\left(HM_t, HM_k\right) = 1 - \sum\nolimits_{m,d} \min\left(HM_t\left(m,d\right), HM_k\left(m,d\right)\right) \qquad (5)$$

where HC and HM are the normalized color and motion histograms of a shot respectively.

2.1.3 Continuity Analysis

Another issue that should be taken into consideration in the course of video shrinking is video continuity. In this paper, we analyze the video continuity through the shot similarity analysis. There has been lots of work on video similarity analysis. In [10], a Best-First Model Merging (BFMM) method is proposed for scene segmentation in which visually similar shots are gradually merged. In our work, we adopt the same method with BFMM to calculate the similarity between shots since the merge order in BFMM can be regarded as the degree of continuity. Here we select two kinds of features, color and shot pace. Firstly, five key frames whose attractiveness values are top five in a shot are picked out to represent the shot content. Then color is used again as the feature of each shot to obtain the differences between adjacent shots. Secondly, by means of the shot boundary detection method proposed in [5], we calculate the length of each shot, which can be regarded as the pace of the shot. Then we can obtain the shot pace differences between adjacent shots which is an essential factor for analyzing the video continuity. After the shot merge step, we normalize the merge order to interval (0, 1]. The smaller the value is, the more similar the two adjacent shots are.

2.1.4 Combination

After the detailed analysis mentioned above, our system steps into a crucial stage which is called combination to remove the unnecessary shots.

The output of speech detection is the speech ratio of each shot. So a threshold is set to extract CDS, which contain little or no speech. The number of CDS provides us a maximum shrinking degree. The redundant shots to be removed will be chosen from these shots.

For each candidate shot, we calculate its attention value and similarity to the preceding one. However, the attractiveness level and the degree of the continuity of a shot is a contradiction. The reason is, on one hand, the more different a shot is to the preceding ones, the more attention it will attract; on the other hand, the more different between adjacent shots, the lower the degree of the continuity is. So it is necessary to acquire an appropriate balance between the two. Here we calculate the value of each candidate shot as:

$$\alpha \times SAV(n) - (1-\alpha) \times SSV(n) \qquad (6)$$

where $SAV(n)$ is the attention value of the n^{th} shot, $SSV(n)$ is the similarity value between the n^{th} shot and its preceding one.

Finally, under given expected shrinkage, the shots of lower values should be removed.

2.2 Spatial Shrinking

After shrinking the video in temporal dimension, viewers may spend less time on browsing these videos on common personal computers. But for mobile devices, because of the small window sizes, it is necessary to go on shrinking the video in spatial dimension. In this section, we adopt the method proposed in [2] to shrink the video in spatial dimension. Firstly, each frame's ROI is extracted according to the saliency map [11]. Then, because of the latent drastic change of ROIs' positions and sizes, in order to offer users more visually pleasing viewing experiences, a curve fitting based smoothing method is proposed to mitigate this issue.

3 Experimental Results

In order to verify the effectiveness of the proposed approach, we perform experiments on different teleplay series with different lengths. The details of the testing data are presented in Table 1.

Table 1. Testing data

No.	Video	Shot	Length
1	Strive1.mpg	540	46'48''
2	Strive2.mpg	621	47'25''
3	Snow Queen1.mpg	902	65'47''
4	Snow Queen2.mpg	863	62'08''
5	Our Ambitious and Restless Youth1.mpg	184	40'07''
6	Our Ambitious and Restless Youth2.mpg	228	39'33''

The most commonly used measures to test the performance are: precision P , recall R and F1-measure $F1$. They are defined as follows:

$$P=\frac{n_{tp}}{n_p},\quad R=\frac{n_{tp}}{n_t},\quad F1=\frac{2*P*R}{P+R} \tag{7}$$

Theoretically, n_{tp} is the number of correctly detected shots to be reserved, n_p represents the number of shots declared as the ones to be reserved, and n_t is the total shots manually labeled as the ones not to be removed. F1-measure is a harmonic mean of precision and recall, which is often used to measure the overall performance of the method.

However, in our experiment, n_p and n_t are always equivalent because of the presence of the input shrinkage value. So we choose another evaluation criterion, called False Positive Rate (FPR), to evaluate the performance of our method alternatively.

$$FPR=\frac{FP}{N} \tag{8}$$

where FP is the number of shots that are falsely deleted, N represents the total number of shots that have been removed.

The specific assessment procedure is as follows. According to our former statistical analysis, the shrinkage of teleplays usually lies between 5% and 15%. Thus, in the testing procedure, the expected shrinkages are set by users as 5%, 6%, 7%, 8%, 9% and 10% respectively, which are all smaller than the maximum shrinkages obtained by speech detection. Then, after the videos are shrunk in temporal dimension, 7 individuals are invited to watch the resulting clips. They are all postgraduate students aging from 23 to 28. We adopt a voting strategy to obtain the FPR for each clip by counting those shots that are falsely deleted. A shot is considered to be falsely deleted if four or more people think it should not be done.

The testing results following our proposed video temporal shrinking method are shown in Table 2.

Table 2. Experimental results in temporal dimension

Video	Expected Shrinkage	False Positive Rate
Strive1.mpg	5%	25.9%
Strive2.mpg	6%	24.3%
Snow Queen1.mpg	7%	15.9%
Snow Queen2.mpg	8%	21.7%
Our Ambitious and Restless Youth1.mpg	9%	25.0%
Our Ambitious and Restless Youth2.mpg	10%	26.1%

Two aspects may influence the FPR of our method. 1) The accuracy of speech detection is not ideal enough. 2) The shot boundary detection still lacks precision. Although it has the above limitations, our shrinking approach is proved feasible and consistent with audience's subjective understanding by experiments.

Fig. 2 illustrates some examples after shrinking the video in spatial dimension.

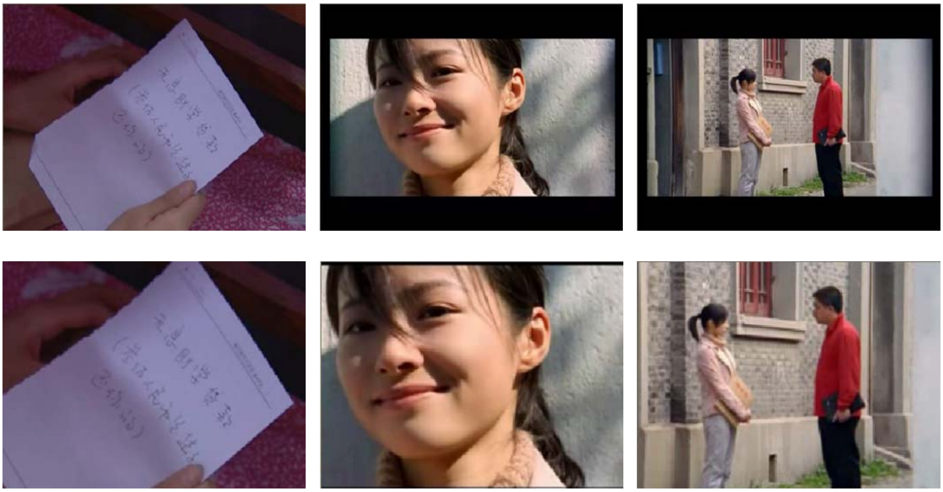

Fig. 2. Examples of spatial shrinking

4 Conclusions

In this paper, we present a video shrinking method to remove video's redundant information in both temporal and spatial dimensions. In temporal dimension, speech detection is first exploited to assign a label to each shot, which shows whether this shot has the possibility to be cut off or not. Then, we could obtain the temporal shrinking result by combining the attractiveness analysis and the continuity analysis of CDS. In spatial dimension, firstly, each frame's ROI is extracted according to the saliency map. Then, a curve fitting based smoothing method is adopted to offer users more pleasing viewing experiences. The performance of the method has been tested on several teleplays with different lengths. The experimental results have verified the effectiveness of the method and indicated that it is a promising solution which greatly reduces the computational complexity.

However, the system can still be improved in some aspects. Firstly, the work of this paper detects the redundant shot. However, a shot can also contain redundant information. So, in our future work, we will try to shrink each shot to further reduce the video length. Secondly, features adopted are only color, motion and shot pace, which are all vision information. We believe that integrating of more information which is more consistent with human understanding, such as text, can improve the performance of the system. Besides, a design of interactive user interface may perfect the art of our system. These are all the tasks what we will do in the future.

Acknowledgements

This work was supported in part by National Natural Science Foundation of China: 60702035 and 60833006, in part by National Hi-Tech Development Program (863 Program) of China: 2006AA01Z117, in part by National Basic Research Program of China (973 Program): 2009CB320906, and in part by Beijing Natural Science Foundation: 4092042.

References

1. Zhu, G.Y., Huang, Q.M., Xu, C.S., Xing, L.Y., Gao, W., Yao, H.X.: Human Behavior Analysis for Highlight Ranking in Broadcast Racket Sports Video. IEEE Transactions on Multimedia 9(6), 1167–1182 (2007)
2. Qiu, X.K., Jiang, S.Q., Huang, Q.M., Liu, H.Y.: Spatial-temporal video browsing for mobile environment based on visual attention. In: 2009 IEEE International Conference on Multimedia and Expo., New York (2009)
3. Simakov, D., Caspi, Y., Shechtman, E., Irani, M.: Summarizing visual data using bidirectional similarity. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2008. CVPR 2008, Anchorage, pp. 1–8 (2008)
4. Liu, F., Gleicher, M.: Video retargeting: automating pan and scan. In: 14th annual ACM international conference on Multimedia, Santa Barbara, pp. 241–250 (2006)
5. Liu, C.X., Liu, H.Y., Jiang, S.Q., Huang, Q.M., Zheng, Y.J., Zhang, W.G.: JDL at Trecvid 2006 Shot Boundary Detection. In: TRECVID 2006 Workshop (2006)

6. Wold, E., Blum, T., Keislar, D., Wheaten, J.: Content-based classification, search, and retrieval of audio. IEEE Multimedia 3(3), 27–36 (1996)
7. Lu, L., Zhang, H.J., Jiang, H.: Content analysis for audio classification and segmentation. IEEE Transactions on Speech and Audio Processing 10(7), 504–516 (2002)
8. Platt, J.C.: Fast training of support vector machines using sequential minimal optimization. In: Advances in kernel methods: support vector learning, pp. 185–208. MIT Press, Cambridge (1998)
9. Liu, H.Y., Jiang, S.Q., Huang, Q.M., Xu, C.S.: A generic virtual content insertion system based on visual attention analysis. In: 16th annual ACM international conference on Multimedia, Vancouver, pp. 379–388 (2008)
10. Li, Z., Wei, Q., Wang, Y.J., Yang, S.Q., Zhang, H.J.: Video shot grouping using best-first model merging. In: SPIE conference on Storage and Retrieval for Media Database, San Jose, pp. 262–296 (2001)
11. Cheng, W.H., Wang, C.W., Wu, J.L.: Video adaptation for small display based on content recomposition. IEEE Transactions on Circuits and Systems for Video Technology 17(1), 43–58 (2007)