Hierarchical Video Summarization Based on Video Structure and Highlight

Yuliang Geng, De Xu, and Songhe Feng

Institute of Computer Science and Technology,
Beijing Jiaotong University, Beijing, 100044, China
gengyuliang@hotmail.com

Abstract. Video summarization is a significant scheme to organize massive video data, and implement a meaningful rapid navigation of video. In this paper, we propose a hierarchical video summarization approach based on video structure and highlight. We extract video structure unit, and measure the unit (frame, shot and scene) importance rank based on visual and audio attention models. According to the unit importance rank, the skim ratio and key frame ratio are assigned to the different video units. Thus we achieve a hierarchical video summary. Experimental results show the excellent performance of the approach.

1 Introduction

With recent advance in digital video technologies, the amount of video data has grown enormously, so quick browsing a video and getting its main content becomes a crucial problem. Video summarization is a significant scheme to organize massive video data, and implement a meaningful rapid navigation of video. Video summarization technique has attracted attention of many researchers in recent years. There are two fundamentally different approaches for video summarization: static summary and dynamic skimming. Static summary is a collection of key frames selected from video sequence, many approaches are proposed to extract and organize key frames, such as video table of contents [1], storyboard [2], and pictorial video summary [3]. Dynamic skimming consists of a collection of video clips selected from video sequence. There are two main approaches for video skimming extraction: one is the predefined event-based approach in which the events are detected and ranked to create video skimming. For example, in sport video [4,5], goal, foul, and touchdown are detected as important events and composite video skimming. The other is a bottom-up approach, which employs special features to analyze the video content [7,8,9]. In [7], authors use audio and video tempo to simulate human's emotion feeling and extract meaningful skim. Literature [8] constructs a user attention curve based on visual, audio attention model to abstract video skimming. In [9], each scene is modeled as a graph, and its optimal skimming is created with graph dynamic programming.

As mentioned above, the static summary based on key frames covers the total video content, but it cannot reflect video semantic content effectively because it loses audio and temporal attributes. The dynamic skimming emphasizes video

D.-Y. Yeung et al. (Eds.): SSPR&SPR 2006, LNCS 4109, pp. 226–234, 2006.
© Springer-Verlag Berlin Heidelberg 2006

highlight and preserves audio and temporal attributes, but it sacrifices the content integrity. In this paper, we integrate the advantages of static summary and dynamic skimming, and propose an effective approach for multilevel video summarization based on video structure and highlight. First we extract the video structure and measure the unit (frame, shot and scene) importance rank based on visual and audio attention models. According to the unit importance rank, the skim ratio and key frame number of video summary are assigned to different video units. Thus a hierarchical video summary is generated. The block diagram of the video summarization approach is shown in Fig. 1, which gives a 3-level video summary that consists of scene level summary, shot level summary and sub-shot level summary from bottom to up. The hierarchical video summarization approach can provide viewers a multilevel summary with different granularity. In the scene level summary, the viewers can obtain an overview of a video, and can grasp the highlight plots rapidly. In the next level summary, the viewers can further obtain more concise video highlight scenes. In general, our approach not only maintains the content integrity but also emphasizes highlight scenes that may attract viewers' attention.

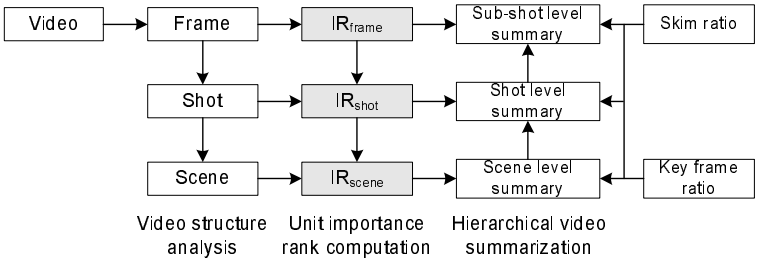

Fig. 1. Block diagram of the hierarchical video summarization approach

The organization of the paper is as follows. Section 2 gives an overview of video structure analysis. We present, in Section 3, the unit importance rank computation based on attention models in detail. Then a hierarchical video summarization approach is proposed in Section 4. Section 5 and 6 give the experimental results and draw the conclusions.

2 Video Structure Analysis

Shot and scene are usually two basic temporal units in video structure analysis. A shot is defined as a single continuous recording made by a camera. A scene consists of a series of related shots (in time, space, etc.), which is a higher-level semantic unit and reflects the narrative structure of a film. We employ singularity detection with wavelet to detect shot boundary [10]. Then we exploit the cinematic rules as a guideline to identify the video scenes [11]. In this step, three main scene categories are identified: dialogue scene, action scene and dialogue with action scene. Thus, we achieve the hierarchical structure of video data.

3 Unit Importance Rank Computation Based on Attention Model

In this section, we compute video unit (frame, shot and scene) importance rank based on visual and audio attention models. And the unit importance rank is regarded as an effective measurement for the highlight of video unit.

3.1 Audio Attention Model

As loudness is a fundamental component of film sound, it plays an important role in defining the overall sonic texture of film. A film usually startles the viewers by exploiting abrupt and extreme shifts in loudness, which is called changes in dynamics [12]. A rough analysis of the loudness can be gained by the square of signal amplitude. In order to stabilize the signal, a Gauss filtering of loudness amplitudes is performed. The loudness amplitude is normalized to archive comparability. Then we utilize the difference between loudness peak $E^{\mathrm{A}}_{\mathrm{peak}}$ and loudness mean $E^{\mathrm{A}}_{\mathrm{mean}}$ to measure loudness dynamic change. Meanwhile, the loudness mean is another important factor in loudness attention measurement. So we define the loudness attention as

$$M_{\mathrm{loud}} = E^{\mathrm{A}}_{\mathrm{mean}} \cdot (E^{\mathrm{A}}_{\mathrm{peak}} - E^{\mathrm{A}}_{\mathrm{mean}}) \tag{1}$$

This metric is similar to the audio saliency attention model proposed by Ma [8], but we more emphasize the dynamic change of loudness. In experiment, the audio signal is sampled at 22.05KHz and each audio frame contains 512 samples shifted by 128 samples from the previous audio frame. The audio feature extraction is based on the audio frame. Here one-second sliding window is used to compute loudness attention M_{loud}.

From the viewpoint of human aural perception, various sounds usually play different roles in attracting the audience attention. So we first classify the audio stream into four classes of semantic segments: silence, speech, music, and environment sound [13]. We assign a weight for each audio semantic segment according to its semantic class.

Obviously, speech usually gives audience more meaningful narrative content, but a long speech scene with low loudness may not attract viewers' attention. While an excellent action scene often accompanies the environment sound with high loudness. Here we unify the sound events, such as explosion, whistle and collision, into the environment sounds, and don't identify them respectively. There are two music effects: harmonic sound and inharmonic sound. Harmonic sounds are perceived as more comfortable, and often are accompanied with mild scene content. While inharmonic sounds often implicate that an unpredictable event may happen, or a worrying event is happening, and can more arouse audience's attention. In the scene construction, the length of the harmonic sound is longer than that of the inharmonic sound.

With above analysis, we define the weights of various audio semantic segments. The weight of a speech segment at time t is defined as

$$w_{\mathrm{s}}(t) = \begin{cases} -(t - t_{\mathrm{start}})/t_{\mathrm{Th}} + 2 & if \quad t - t_{\mathrm{start}} < t_{\mathrm{Th}} \\ 1 & else \end{cases} \tag{2}$$

where t_{Th} is a given threshold, and t_{start} is the start time of the speech segment.

The weight of a music segment at time t is defined as

$$w_{\mathrm{m}}(t) = \exp(MinLM - L_{\mathrm{music}}) + 1 \tag{3}$$

where $MinLM$ denotes the minimum length among all the music segments, and L_{music} is the length of the current music segment.

The weight of a silence segment $w_{\mathrm{z}}(t)$ is set as 1, and the weight of an environment segment $w_{\mathrm{e}}(t)$ is set as 1.5.

Thus, the audio attention value at the tth second is computed as

$$M_{\mathrm{audio}}(t) = w(t) \cdot M_{\mathrm{loud}}(t) \tag{4}$$

where $M_{\mathrm{loud}}(t)$ is the loudness attention value at the tth second, and $w(t)$ is the weight of the corresponding semantic segment. For example, if the audio segment at the tth second is music, $w(t)$ is set as $w_{\mathrm{m}}(t)$.

3.2 Visual Attention Model

As motion is an intrinsic nature of video and implicates some semantic cues in visual perception, we combine the camera motion and local motion to compute visual attention value.

First, we employ a qualitative method to estimate the camera motion category, which employs motion vectors mutual relationship to implement camera motion classification [14]. As the camera motion continuity, we utilize a sliding window to filter abnormal camera motion. Similar to camera attention weighted strategy [8], we assign different weight w_{c} for a given video frame according to its camera motion category.

Then, the visual attention value of the ith frame is represented as

$$M_{\mathrm{visual}}(i) = w_{\mathrm{c}}(i) \cdot E^{\mathrm{M}}(i) \tag{5}$$

where $E^{\mathrm{M}}(i)$ is the motion activity of the ith frame, which is defined as the standard deviation of motion vector magnitudes because it can measure local motion intensity effectively.

3.3 Unit Importance Rank Computation

Because the visual attention value is a metric based on video frame, and the audio attention value is a metric based on second, we first unify the measurement units to frame according to the video frame rate. Then the visual and audio attention values are normalized by using Gauss normalization formula, and are denoted as $\bar{M}_{\mathrm{visual}}(i)$ and $\bar{M}_{\mathrm{audio}}(i)$. The attention value at the ith frame is defined as a linear combination of the audio and visual attention values.

$$IR_{\mathrm{frame}}(i) = \alpha \cdot \bar{M}_{\mathrm{visual}}(i) + \beta \cdot \bar{M}_{\mathrm{audio}}(i) \tag{6}$$

where α and β are the preassigned weights and used to be a tradeoff between the visual and audio attention values.

The shot importance rank of the shot j is defined as

$$IR_{\text{shot}}(j) = \sum_i IR_{\text{frame}}(i)/N_{\text{frame}}(j) \tag{7}$$

where $N_{\text{frame}}(j)$ is the video frame number of the shot j.

We employ three main components to determine the scene importance rank, namely, shot cut frequency, visual and audio attention values. In the film editing, filmmaker often uses a series of short shots to create tense or strong atmosphere. The shot cut frequency of shot j is defined as the inverse of shot length and is normalized as $SF(j)$. We define the scene importance rank of the scene k as

$$IR_{\text{scene}}(k) = a \cdot \sum_j (IR_{\text{shot}}(j) \cdot N_{\text{frame}}(j))/(\sum_j N_{\text{frame}}(j)) + b \cdot \sum_j SF(j)/N_{\text{shot}} \tag{8}$$

where N_{shot} is the shot number of the scene. a and b are the weight values.

4 Hierarchical Video Summarization

4.1 Scene Level Summary

Once the skim ratio SR and the key frame ratio KFR are given, we may assign them to each scene according to the scene importance rank. Before assigning the key frame number, we set the minimum of the key frame number of the various scene categories that are extracted in Section 2. For the dialogue scene, dialogist number, which can be archived in scene analysis, is used to determine the key frame number. The action scene should have three key frames at least to represent the attack, sustain and release of action scene. Here we use $MinKF(i)$ to represent the minimum of key frame number of the scene i. So the key frame number of the scene i is assigned as

$$KFN_{\text{scene}}(i) = \min(KFN_{\text{video}} \cdot IR_{\text{scene}}(i)/\sum_j IR_{\text{scene}}(j), MinKF(i)) \tag{9}$$

where $IR_{\text{scene}}(i)$ is the scene importance rank of scene i. KFN_{video} is the total number of key frames in the video sequence and is set as the nearest integer to $KFR \cdot L_{\text{video}}$. L_{video} is the total number of video frames in the video sequence.

For every scene, we utilize the C-mean clustering algorithm to locate the key frames according to its key frame number KFN_{scene}. Thus, we obtain the scene level summary that consists of a group of key frames.

Then, we select the first K scenes with the greatest scene importance ranks as skimming scenes according to the skim ratio. K is the maximum integer, which satisfies the inequality: $\sum_{k=1}^{K} L_{\text{scene}}(k)/L_{\text{video}} \leq SR, k \in \{\text{skimming scenes}\}$. $L_{\text{scene}}(k)$ is the total number of video frames in the scene k. The other scenes with low scene importance ranks are regarded as common scenes. Thus we obtain the scene level summary that consists of a group of skimming scenes.

4.2 Shot/Sub-shot Level Summary

The approach for shot level summarization is similar to the approach for scene level summarization. In this step, we need reset the minimum of the key frame number for each shot according to its camera motion category. Here the minimum of key frame number for still shot is set as 1, and other shot types are set as 2. We also need reassign the skim ratio for each scene, $SR_{\text{scene}}(i)$, according to its scene importance rank as Eq. (10) depicted. If $SR_{\text{scene}}(i)$ is less than a given threshold T_{SR}, we will discard this scene. Thus we obtain the shot level summary according to $SR_{\text{scene}}(i)$ and $KFR_{\text{scene}}(i)$. $KFR_{\text{scene}}(i)$ is the key frame ratio of the scene i and is set as $KFN_{\text{scene}}(i)/L_{\text{scene}}(i)$.

$$SR_{\text{scene}}(i) = \min\left(\frac{SR \cdot L_{\text{video}}}{L_{\text{scene}}(i)} \cdot \frac{IR_{\text{scene}}(i)}{\sum_j IR_{\text{scene}}(j)}, 1\right) \tag{10}$$

Next, we construct the sub-shot level summary. For a given shot, we reassign its key frame ratio and skim ratio as the same way. Then we extract its sub-shot around the maximum of attention value curve $IR_{\text{frame}}(i)$. The length of the sub-shot is determined by its skim ratio. The key frames are also extracted to represent the skimming shot according to its key frame ratio.

Thus we get a hierarchical and scalable video summary that is composed of static key frame sequence and dynamic skimming. As the video hierarchical structure is the basic element for filmmaker to construct story plots, the summary based on the video structure and unit important rank can provide a good tradeoff between the content integrity and content compactness. Additionally, users may adjust the summary by tuning the key frame ratio and skim ratio.

5 Experimental Results

The video summary is the logical layer of representation based on subjective semantics, and there are still no objective definition and evaluation criterion. So how to evaluate video summary is a difficult issue. In experiment, we invite test users including naive users and experienced users (engaged in video retrieval) to assess the performance of the proposed video summarization approach. We collect the test dataset from five various movie videos, namely, *Rain man*, *Ghost* are dramas; *Leon* and *The Shaolin Temple* are action movies; and *Shrek* is a cartoon movie. The total length of test dataset is about 75 minutes, which is composed of 878 shots and 53 scenes. All the video data is in MPEG-1 format with a frame rate of 30 fps, and the audio track was sampled at 22.05 KHz.

First, we carry out an experimental comparison to evaluate the performance of the key frame sequence of video summary between our approach (denoted as HVS) and storyboard technique (denoted as ST) [2]. Here we design two evaluation criteria, content compactness and content integrity, to evaluate the performance of these two approaches. For the content compactness, test users give an assessment of being too much, much, good, few and too few to key frame sequence, corresponding to quantitative scores: 0.1, 0.5, 1, 0.5 and 1. The content

integrity means whether test users can capture the story plot from the key frame sequence by answering the questions, such as, "who", "where", "when", and "what". According to the accuracy of answers, the score of the content integrity is obtained. All the questions are selected from the user investigation report. For example, for the static summary, users pay more attention to whether they can get the information about the protagonists, location, and coarse events, which is the reason that these questions are provided in our evaluation scheme.

Fig. 2 gives the performance curves of the content integrity and the content compactness. As Fig. 2 illustrates, our approach can maintain the content integrity at different key frame ratio very well. When the key frame ratio is increasing, the content compactness is decreasing. Our approach (sub-shot level summary) got the best performance when the key frame ratio is set as 0.02. Our proposed method can provide a meaningful representation of video content because the key frames assignment and location are based on the semantic content of the video unit, while the storyboard based on the hierarchical clustering method cannot ensure the extracted key frames have semantic structure.

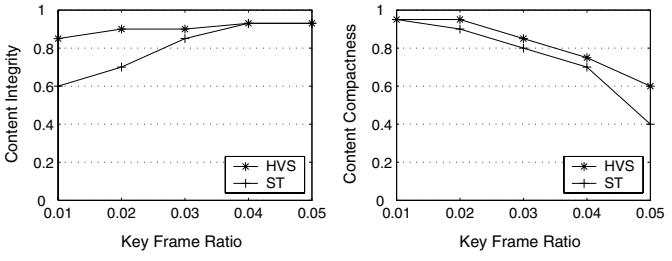

Fig. 2. Experimental comparison between our approach and storyboard method. Left: content integrity curve; right: content compactness curve.

Next, we evaluate the quality of the video skimming from two criteria: comprehensibility and highlight degree. For a good video skimming, like the trailer of a movie, the users more care whether its content is comprehended easily, and whether the summary is composed of the most excellent video clips. Because it is still a subjective problem to evaluate the video skimming, we only assess the video skimming by analyzing test users' answers to the test questions. Here we carry out an experimental comparison between our approach and the method (denoted as SAGO) proposed in [9]. Video skimming assessment is complex process. We first let the test users look through the video skimming from low to high skim ratio in turn. When the test users finished viewing the video skimming with a certain skim ratio, they need assess the video skimming according to the two criteria. Then the users continue their assessment with another skim ratio, and so on. After they finished all the video skimming, they may reassess these video skimming. The assessment is quantified to score from 0 to 1. Fig. 3 gives the experimental results.

As the comparison results shown, our proposed approach has a good performance. One important reason is that we extract video skimming under the

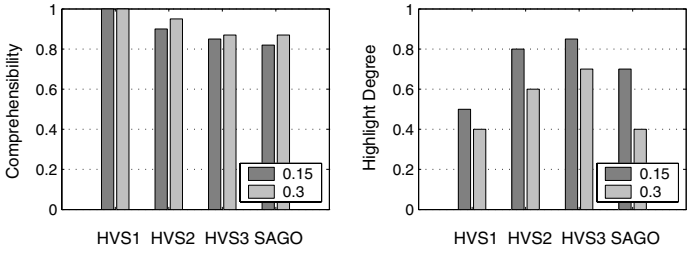

Fig. 3. Experimental comparison between our approach and the SAGO [9]. Left: comprehensibility assessment score; right: highlight degree assessment score. Notes: HVS1 denotes scene level summary, HVS2 denotes shot level summary, HVS3 denotes sub-shot level summary.

guideline of user attention value, which ensures the highlight degree of video summary. Another reason is that the hierarchical structure of video skimming keeps the integrity of semantic content. From Fig. 3 we can see, the video skimming with skim rate of 0.3 has higher comprehensibility score, and the video skimming with skim rate of 0.15 has higher highlight degree. In general, when the skim rate is set as 0.15, sub-shot level summary can archive the best experimental results.

6 Conclusions

We have addressed the main issues of the video summarization from the video structure analysis, unit importance rank computation to video summarization. As the video hierarchical structure is the basic element for filmmaker to construct story plots, and the unit importance rank is an effective measurement for the highlight of video unit, the approach for video summarization based on video structure and highlight can give us a better tradeoff between the content integrity, comprehensibility and the content compactness, highlight degree. Additionally, users can also adjust video summary by tuning the key frame ratio and skim ratio. In general, our proposed approach can provide us a multilevel and flexible video summary with different granularity. Experimental results have been reported in detail.

Acknowledgements

This research was supported by Science Foundation of Beijing Jiaotong University (Grant No. 2004SM013).

References

1. Rui, Y., Huang, T.S., Mehrotra, S.: Constructing Table-of-Content for Videos. ACM Multimedia Systems Journal, Special Issue Multimedia Systems on Video Libraries, Vol. 7, No. 5 (1999) 359-368

2. Hasebea, S., Mustafa M.S.: Constructing Storyboards Based on Hierarchical Clustering Analysis. In: Proceedings of Visual Communications and Image Processing, SPIE Vol. 5960 (2005) 437-445
3. Ma, Y.F., Zhang, H.J.: Video Snapshot: A Bird View of Video Sequence, In: Proceed-ings of the 11th International Multimedia Modelling Conference (2005) 94-101
4. Tjondronegoro, D.W., Chen, Y.P.P., Pham, B.: Classification of Self-Consumable Highlights for Soccer Video Summaries. In: Proceedings of IEEE ICME, Vol. 1 (2004) 579-582
5. Noboru, B., Yoshihiko, K., et al.: Personalized Abstraction of Broadcasted American Football Video by Highlight Selection, IEEE Transactions on Multimedia, Vol. 6, No. 4, (2004) 575-586
6. Rapantzikos, K., Tsapatsoulis, N., Avrithis, Y.: Spatiotemporal Visual Attention Architecture for Video Analysis. In: Proceedings of Multimedia Signal Processing (2004) 83-86
7. Lee S.H., Yeh, C.H., Kuo, C.C.J.: Video Skimming Based on Story Units via General Tempo Analysis. In: Proceedings of IEEE ICME, Vol. 2 (2004) 1099-1102
8. Ma, Y.F., Lu, L., Zhang, H.J., Li, M.J.: A User Attention Model for Video Summarization. In: Proceedings of ACM Multimedia, (2002) 533-542
9. Lu, S., King, I., Michael, R.L.: Video Summarization by Video Structure Analysis and Graph Optimization. In: Proceedings of IEEE ICME, Vol. 3 (2004) 1959-1962
10. Geng, Y.L., Xu, D.: A Unified Framework for Shot Boundary Detection. Journal of Image and Graphics (Chinese) Vol.10, No.5 (2005) 650-655
11. Geng, Y.L., Xu, D., Wu, A.M.: Effective Video Scene Detection Approach Based on Cinematic Rules. In: Proceedings of KES, LNCS, Vol. 3682 (2005) 1197-1204
12. Ohm, J.R, Multimedia Communication Technology Representation, Transmission and Identification of Multimedia Signals. Springer, Berlin Heidelberg (2004)
13. Lu, L., Jiang, H. and Zhang, H.J.: A Robust Audio Classification and Segmentation Method. In: Proceedings of ACM Multimedia, Vol. 9 (2001) 203-211
14. Zhu, X.Q., Xue, X.Y.: Qualitative Camera Motion Classification for Content-Based Video Indexing. In: Proceedings of IEEE PCM, LNCS, Vol. 2532, (2002) 1128 - 1136