Attention-based Video Summarisation in Rushes Collection

Reede Ren, Punitha Puttu Swamy, Jana Urban, Joemon Jose
IR Group, University of Glasgow
17 Lilybank Gardens, Glasgow
UK, G12 8QQ
reede,punitha,jana,jj@dcs.gla.ac.uk

ABSTRACT

This paper presents the framework of a general video summarisation system on the rushes collection, which formalises the summarisation process as an $0-1$ Knapsack optimisation problem. Three stages are included, namely content analysis, content selection and summary composition. Content analysis is the pre-processing step, consisting of shot segmentation, feature extraction, raw video discrimination and shot clustering. Content selection weights the importance of video segments by an attention model. A greedy approximation approach is employed in the composition of summary videos with a cost function, which balances the video importance gain and the duration cost. The average content coverage achieved on the rushes test collection is about 29%, while the average score on readability is 3.13 with the redundancy credit at 4.08.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Content-based Video Summary

General Terms

Algorithms, Experimentation

Keywords

attention analysis, redundancy detection, video summarisation

1. INTRODUCTION

Video summarisation is a form of content-based video compression, which produces a video abstract by removing the redundancy in content presentation. The literature can be roughly categorised into two classes, content abridgement and event summary. Speeding up the playback [7] is a straightforward approach of content abridgement. Although it does not abridge video contents, this approach condenses the temporal duration of presentation. Generally, content abridgement simulates video context and identifies the semantic redundancy with proposed content models, i.e. language model and domain ontology. The InforMedia system [5] gathered text data inside videos, such as caption, video script and automatic speech recognition(ASR), to produce a meaningful video skimming based on text context. Ma et al. [4] proposed a psychological *attention* model to estimate the "attractiveness" of general contents. Hua et al. [2] employed this model in the home video summarisation and matched their video abstracts to the background music. However, it is clearly evident that there is no single best model for video content description and that the success of video summarisation systems depend not only on the method used but also on statistical properties of data. For instance, the efficiency of the InforMedia system strongly relies on extracted text information and the robustness of employed text summarisation algorithms. The appearance of an unknown keyword or concept can decrease system performance seriously. Utilising prior domain knowledge, event summary approaches [6] define a set of semantically important video moments as events. They regard the collection of events as a semantically meaningful video summary. Several applications have been well developed for some specific video genres, i.e. news abstraction and sports highlights [9].

To assess the quality of video summary, there are two fundamental requirements, (1) the summary should cover the most important content topics in the original video; (2) the video summary should be easily understood by viewers. Note that the coverage of the most important contents is not equivalent to that of general content topics. There is a trade-off between the content coverage and the readability. Given the limitation on summary duration, too many content topics will shorten the average length of topic presentation so as to break the integrity of story content. For example, a slide of key frames can cover all content topics but is too difficult to be understood. For a summarisation system, these requirements will be specified as: (1) allocate the most important content topic reasonably; (2) discover the content redundancy efficiently; (3) compose the summary properly according to the content importance and context.

2. RUSHES COLLECTION

The rushes collection builds up a common test bed for content-based video summarisation. It consists of multiple unstructured video genres from different content domains, such as children television program, travel tour video, and

TVS'07, September 28, 2007, Augsburg, Bavaria, Germany.

raw video data. This generality not only evaluates the robustness of summarisation systems, but also brings many challenges against current domain-dependent techniques. For instance, though matching ASR keywords is helpful in the discovery of video context in children television programs, such an audio-content mapping can hardly be employed in the competition because it is domain specific. Moreover, the inclusion of raw video introduces extra redundancy. The raw video is a direct record of the production process, which is made up by a sequence of repeating shots. Although such a redundancy is easy to be identified, the problem is how to select a video clip among a group of highly similar shots. The summarisation of raw videos is an automatic video editing [2] rather than the content-based abridgement. Nevertheless, the absence of editing effects indicates: (1) the appearance sequence of shot is random and might not follow any context; (2) there is much unnecessary information both in audio and visual streams. It is inefficient to simulate video context by modelling concept sequence or matching audio key words in raw video [11]. A preprocessing step is developed to discriminate raw video and other video genres. In this paper, we present our summarisation system for the rushes summarisation task. We regard the task of video summarisation as a typical $0-1$ Knapsack problem and propose a greedy approximation solution, which relies on a cost function to find an optimised answer. Three processing steps are taken, namely content analysis, content selection and summary composition.

The rest of this paper is organised as follows. After formulating the summary problem in Section 3, Section 4 overviews the framework of our rushes summarisation system. Video genre classification and perceptual shot weighting are introduced in Section 5 and Section 6. Section 7 defines the cost function and presents the greedy algorithm to optimise the process of summary composition. Conclusions and ideas for future extensions of this system are stated in Section 8.

3. PROBLEM FORMULATION

The objective of video summarisation is to find a short presentation of the original video, while fulfilling the requirements or a set of rules. Given the difficulty in the rephrasing of video stories, most approaches select a particular set of shots from a long video sequence. To clearly describe our algorithm, we define a series of symbols which will be used in this paper. A video v consists of a series of shots as,

$$Shot = \{Shot_i, 0 \leq i < K^{SH}\} \quad (1)$$

$$Frame_{(j,i)} = \{Frame_j \in Shot_i\} \quad (2)$$

where K^{SH} is the number of shots in the video. These shots are grouped according to their visual similarity.

$$ShotClass = \{SC_n, 0 \leq n < K^{SC}\} \quad (3)$$

where K^{SC} is the number of shot clusters. For a shot, several features are extracted to represent content and time relation, such as the attention index, start moment and end moment. These features are denoted as follows,

$$Attention = \{att_i, 0 \leq i < K^{SH}\} \quad (4)$$

$$Start = \{s_i, 0 \leq i < K^{SH}\} \quad (5)$$

$$End = \{e_i, 0 \leq i < K^{SH}\} \quad (6)$$

Therefore, the problem of video summarisation can be formulated as to select M elements from the K^{SH}-element set, while the overall length of selected elements $L = \sum_{i=0}^{M} \|e_i - s_i\|$ is smaller than the given length limitation. Let θ stand for the M-element subset and Θ the set of all subsets of this form. The content selection thus can be viewed as a Knapsack problem [1], which tries to maximise or minimise given costs or objective functions $F(Attention, \theta), \theta \in \Theta$, while keeping the capacity limitation.

$$maximise \sum_{i}^{K^{SH}} p_i F_i(att_i), \quad (7)$$

$$subject: \sum_{i}^{K^{SH}} p_i \|e_i - s_i\| \leq 0.04 \sum_{i}^{K^{SH}} \|e_i - s_i\| \quad (8)$$

where $p_i \in \{0, 1\}$ is the decision array on the shot selection. Note that the Knapsack problem is NP-complete.

4. SYSTEM OVERVIEW

We improved our football video summarisation system [9] for the rushes competition. There are three processing stages as illustrated in Fig 1. The first stage is content analysis, which decomposes the original video into shots, and extracts low level audio and visual features, such as edge histogram and silence pitch ratio, for content identification. Video shots are clustered according to their visual similarity and video genres are discriminated by modelling the temporal distribution of shot classes. In the second stage, we employ an attention model [2][9] to assume the content importance. Video shots are ranked according to their attention intensity and the number of embedded content concepts, i.e. human face. The last stage is summary composition, which renders selected video segments and finally composes the summary video. Hence this stage focuses on how to ensure the readability of the summary and the integrity of selected content stories. A greedy approximation approach is followed to find an optimised solution for this Knapsack problem: (1) we select the shot with highest attention intensity among each shot classes; (2) we rank all selected shots by their cost; (3) we add the entire shot into summary according to their rank until the summary length exceeds predefined length threshold. The insertion of whole shot guarantees the integrity of content presentation and the ranking on attention intensity ensures only the most important video clips will be included in the summary.

5. CONTENT ANALYSIS

Content analysis consists of three components, shot segmentation, shot clustering based on the visual similarity and video genre classification. Note that the rushes collection is absent of editing effects. Most shot boundaries are simple cuts. A two-threshold algorithm [13] is employed to allocate shot boundary. In the audio track, we segment sentences

[1]The knapsack problem is a problem in combinatorial optimization. It derives its name from the maximisation problem of choosing possible essentials that can fit into one bag (of maximum weight) to be carried on a trip. Given a set of items, each with a cost and a value, then determine the number of each item to include in a collection so that the total cost is less than some given cost and the total value is as large as possible.

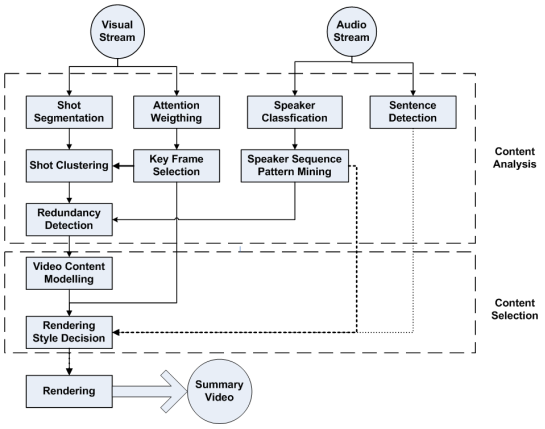

Figure 1: Video Summarisation System Framework

by detecting silent pitches with the base band energy, and try to discriminate speakers by a Gaussian mixture model [10]. Some high level concepts are extracted, too. We use an Adaboost detector to search human face in the video.

5.1 Visual shot clustering

Clustering similar shots is an efficient approach to identify visual redundancy and discover content novelty [12], especially for the raw video data. In a shot, visual frames are sampled one for every 25 visual frames (1/25). Their color layout and edge histogram are computed and combined to set up the feature space, in which each row corresponds to the representative feature vector of a frame. Euclidean distance is used to compute the distances between these feature vectors. A hierarchical cluster tree, called as dendrogram, using complete linkage is constructed to group these feature vectors. The clustering process starts with individual frames, which are represented as feature vectors, and subsequently groups them into clusters based on the maximum distance between the feature vectors. Formally, the process of complete linkage can be stated as follows.(Eq 9),

$$D(R,S) = \max_{i \in R, j \in S}(d(i,j)) \tag{9}$$

where $D(R,S)$ is the distance between two clusters R,S and i,j are a data point in the cluster R,S, respectively. At each stage of complete linkage, the clusters r and s, whose $D(r,s)$ is minimum among all, are merged into one class. Since our aim is to get similar frames grouped into some acceptable number of clusters, we terminate the process of clustering using experimentally chosen thresholds, which is 0.5 in our experimentation.

5.2 Video Genre Discrimination

The discrimination of video genre is essential for the later content modelling. However, we only identify raw video, given the genre generality in the rushes collection. Two features are used, the average member number in shot classes and the average time distance ratio. The time distance between two shots (i,j) is defined as,

$$D(i,j) = \frac{\|s_i - s_j\| + \|e_i - e_j\|}{2} \tag{10}$$

The distance between two shot classes R,S is the minimum distance between their members (Eq 11).

$$DC(R,S) = \min_{i \in R, j \in S} D(i,j) \tag{11}$$

For a shot class SC_i, the time distance ratio is the intra-class distance over that of inter-class distance (Eq 12).

$$DR_{SC_i} = \frac{K^{SC}}{K^{SC_i}} \frac{\sum_{n,m \in SC_i; n \neq m} D(n,m)}{\sum_{j \neq i; j=0}^{K^{SC}} DC(i,j)} \tag{12}$$

where K^{SC_i} is the member number of shot class SC_i. We manually mark three raw videos in the development collection to train a linear classifier, which discriminates raw video vs. non-raw video. The precision in the development collection is above 93%.

To remove redundant shots in the raw video, we select the last shot in the time sequence to stand for its shot class. These selected shots are linked to produce a middle-level video, which will replace the original video in the later content selection stage.

6. CONTENT SELECTION

As we have mentioned, approaches of automatic semantic understanding are inefficient in the rushes collection, especially in the case of unstructured videos. An alternation is the *attention* assumption [4][2][9], which detects "attractive" segments while avoiding the full understanding on video context. As a psychobiological concept, *attention* is widely used in computing psychology to measure the intensity of reflection against stimulus, i.e. a flash in darkness or a clip of music in the silence. Note that the stimulus is a joint effect of visual, audio, and linguistic issues. We modified our attention model in the sports video summarisation [9] to assume the attention intensity or "attractiveness" of rushes videos. Three low level audio features are kept, the base band energy [3], speech pitch ratio [14], and the first order derivatives of Mel-frequency Cepstral Coefficients (MFCC) [1]. Four visual features are used, including visual harmony, shot duration, and motion area and domain colour ratio. Visual harmony is proposed for the measurement of static spatial contrast. Given the block-based encoder in commercial standards, i.e. MPEG-1(8×8 blocks), block mean hue (Eq 13) and block hue covariance (Eq 14) for $n \times n$ image block with the centre at (i,j),

$$mean(i,j) = \frac{1}{n^2} \sum_{x=1}^{n} \sum_{y=1}^{n} C(i \times n + x, j \times n + y) \tag{13}$$

$$cov(i,j) = \frac{1}{n^2} \sum_{x=1}^{n} \sum_{y=1}^{n} (C(i \times n + x, j \times n + y) - mean(i,j)) \tag{14}$$

where C is the pixel colour. We use an 256-bin histogram to count the block covariance distribution. Then the visual harmony of a frame is,

$$Vh = \arg\max_{N} \sum_{n=0}^{N} (-P_n \log(P_n)) \tag{15}$$

where P_n is the portion of bin n over all histogram. The visual harmony Vh is the block covariance value at the bin N. Shot duration Vt is the length of a shot. Motion area Vd

is the average number of pixel difference between adjacent frames in a shot. The domain colour ratio Vc measures the size of image area with uniform colour.

In our sports video summary system, we normalised these feature signals by their self-entropy [2] and developed a multiresolution autoregressive model to fuse them and assume a unified attention curve, which is robust against noise. However, the rushes cannot support such a statistics model because of the small data size and the generality in game genres. Nevertheless, some videos in the collection only contain three shots. We just normalised all feature intensity into $[0, 1]$ and adopted a linear combination to combine these stimulus (Eq 17). The visual attention M_v is the average of normalised visual stimulus(Eq 16), the same as audio attention M_a.

$$M_v = \frac{\overline{Vh} + \overline{Vt} + \overline{Vd} + \overline{Vc}}{4} \tag{16}$$

The overall attention is the linear combination of visual and audio attention intensity.

$$att_i = w_v M_v + w_a M_a \tag{17}$$

where w_a, w_v are combination weights. In the experiments, w_a is 0.25 while w_v 0.75.

To match audio to the summary, we used the feature of audio base band energy to detect salient periods before and after shot boundary so as to segment audio stream. However, we found that it made the video more difficult to be understood if the summary contained the audio segment rather than discarding them. It is partially caused by the discontinuity in the linguistic stream of audio. This observation is further supported by the evaluation report of competition [8]. As a result, our summary videos do not contain audio track.

7. VIDEO SUMMARY COMPOSITION

In Section 3, we formulate the composition of a video summary as an optimisation problem of $0-1$ Knapsack question, namely selecting a subset of shots which maximises the "attractiveness" while keep the overall length less than the given duration. There are three independent objectives identified in Section 1:

1. Select "important" shots;
2. Remove "redundancy" contents;
3. Keep "integrity" of story.

The cost function of including a shot i could be

$$Cost(i) = \frac{F_a(att_i)F_r(i)F_i(i)}{\|e_i - s_i\|} \tag{18}$$

where F_a, F_r, F_i are evaluation functions on shot content importance, redundancy and integrity. However, it is difficult to quantify these evaluations. Since the shot is regarded as the fundamental content unit of video story, F_i will be maximised by adding the whole shot into the summary. F_r assumes the algorithm efficiency of redundancy removal. After the clustering of visual similarity and content concepts, F_r would be a constant in the stage of summary composition, if all content redundancy had already been identified. Finally, the cost function for the greedy approximation solution is,

$$Cost(i) = \frac{att_i}{\|e_i - s_i\|} \tag{19}$$

[2]Self-entropy is the measurement of information gain and directly propositional to attention intensity. In the view of information theory, *attention* is the ability of consuming information. The pan-out speed of message will decide the distribution of *attention* in a neutral situation that people keep neutral or feel interested or uninterested in all active information sources. In such a case, self-entropy can be used to measure the intensity of attention.

8. EXPERIMENT AND DISCUSSION

The processing time of our system is about 1/2 real time on a Windows desktop with 2.4GHz P4 CPU and 1G memory, including shot segmentation, content analysis, shot selection and summary composition. Additionally, our shot segmentation is about 1/3 real time. The average content coverage is about 29%, while the average qualification score on readability is about 3.13 and that on redundancy at about 4.08. The qualification evaluation on "easy to understand" stands for the readability of summary videos, which ranges from 1 to 5, and 5 is strongly agree. The higher the score the easier for evaluators to understand the video. The evaluation of "many repeated segments" tests the redundancy of summary videos. It ranges from 1 to 5, and 1 is strongly agree while 5 shows there is no redundancy in the video.

The major problem of our system is the low content coverage. There are several reasons: (1) taking the shot length as a fact in the attention estimation; (2) including the entire shot in the summary; (3) clustering shots strictly, which brings fewer classes than expected. Shot duration is an experimental feature in the attention assumption. It takes the place of shot frequency (the number of shots in the given temporal period) in the sports video summarisation system, because the computation of shot frequency was trivial in the rushes development collection. Some videos are less than 3 minutes long while some shots were as long as 5 minutes. It is difficult to set a proper size of time window. Meanwhile, a semantically important shot is usually longer than the unimportant one. The inclusion of shot length seems to be plausible in the estimation of attention intensity. However, this decision favours the selection of long shots and weakens the cost function (Eq 18), which uses the shot duration to balance the gain of content importance. As a result, too many long shots are selected and the number of content topics in a summary is decreased significantly. In later experiments, we find that the number of content topics can even be doubled if we remove the feature of shot duration from the attention model. It is a sub-optimised solution to grantee the content integrity that we add the entire shot into the summary. Some shorter temporal structure might be more efficient. For example, Hua et al. [2] proposed the sub-shot structure for video summarisation. There raises the problem how to design the integrity evaluation function F_i. A possible solution is to employ the time statistics on the audio silence and visual variation with stochastic models, such as Weibull and Erlang distribution. Additionally, the cost function needs improvement to take the assumption on the content integrity and self redundancy cost on the summary video into consideration. The choice of clustering algorithm relies on the evaluation of redundancy. But it is hard to identify

the high level semantic redundancy without the knowledge of video genre and content domain. Moreover, the low level redundancy on visual and audio similarity is not equivalent with that on high level. As we have mentioned, the major challenge in the rushes collection is the generality of video genre. It is difficult to develop some domain or video genre-related context models. Hence we have to deal with every video individually, which not only hinders the mining of video patterns, but also makes it hard to extend audio-visual patterns to other videos.

The composition of summary video is an interesting research question, too. Although the inclusion of audio track seems to be unsuccessful, appending extra information into the summary video is able to improve readability. For instance, a title frame(Fig 2) will dramatically increase the understanding on sports video contents [9]. Another similar approach is to insert captions by mining keyword pattern in the text stream. Nevertheless, employing active editing effects may be an efficient method [2]. For example, we can zoom in the rectangle of interest (ROI) to emphasise the content topic. However, it is necessary to balance the gain on the readability and the loss on the information integrity carefully.

Figure 2: Title pages in abstract composition

9. ACKNOWLEDGEMENT

The research leading to this paper was partially supported by European Commission under contracts: Knowledge Space of semantic inference for automatic annotation and retrieval of multimedia content(K-Space FP6-027026), Semantic Audiovisual Entertainment Reusable Objects (SALERO FP6 -027122) and Search Environments for Media (SEMEDIA FP6 - 045032).

10. REFERENCES

[1] D. P. W. Ellis. PLP and RASTA (and MFCC, and inversion) in Matlab, 2005. online web resource.

[2] X.-S. Hua, L. Lu, and H.-J. Zhang. Optimization-based automated home video editing system. *IEEE Transactions on Circuits and Systems for Video Technology*, 14(5):572–583, May 2004.

[3] R. Lenardi, P.Migliorati, and M.Prandini. Semantic indexing of soccer audio-visual sequence: A multimodal approach based on controlled markov chains. *IEEE Trans on Circuits and System for Video Technology*, 14:634–643, May 2004.

[4] Y.-F. Ma, L. Lu, H.-J. Zhang, and M. Li. A user attention model for video summarization. In *MULTIMEDIA '02: Proceedings of the tenth ACM international conference on Multimedia*, pages 533–542, New York, NY, USA, 2002. ACM Press.

[5] M.A.Smith and T.Kanade. Video skimming and characterisation through the combination of image and language understanding techniques. In *Proc. Computer Vision and Pattern Recognition*, pages 775–781, 1997.

[6] N.Jeho and H.T.Ahmed. Dynamic video summarisation and visualisation. In *ACM Multimedia*, pages 53–56, Orlando, FL, Oct 1999.

[7] N.Omoigui, L.He, A.Gupta, J.Grudin, and E.Sanoki. Time-compression: System concerns, usage, and benefits. In *Proc. ACM ICH*, pages 136–143, 1999.

[8] P. Over, A. F. Smeaton, and P. Kelly. The TRECVID 2007 BBC rushes summarization evaluation pilot. In *Proceedings of the TRECVID Workshop on Video Summarization (TVS'07)*, pages 1–15, New York, NY, September 2007. ACM Press.

[9] R. Ren and J. Jose. Affective sports highlight detection. In *European Signal Processing Conference 2007*, Sept 2007.

[10] D. A. Reynolds and R. C. Rose. Robust text-independent speak identification using gaussian mixture speaker models. *IEEE Trans. on Speech and Audio Processing*, 3(1):72–83, Jan 1995.

[11] H. Sundaram, L. Xie, and S. Chang. A utility framework for the automatic generation of audiovisual skims. In *ACM Multimedia*, December 2002.

[12] T.Odin and D.Addison. Novelty detection using neural network technology. In *Proc. COMADEN*, 2000.

[13] N. Vasconcelos and A. Lippman. Bayesian video shot segmentation. In *NIPS*, pages 1009–1015, 2000.

[14] J. Wang, C. Xu, E. Chng, K. Wah, and Q. Tian. Automatic replay generation for soccer video broadcasting. In *MULTIMEDIA '04: Proceedings of the 12th annual ACM international conference on Multimedia*, pages 32–39, New York, NY, USA, 2004. ACM Press.