MITSUBISHI ELECTRIC RESEARCH LABORATORIES
http://www.merl.com

Blind Summarization: Content-Adaptive Video Summarization Using Time-Series Analysis

Ajay Divakakran, Regunathan Radhakrishnan, Kadir A. Peker

TR2006-026 January 2006

Abstract

Severe complexity constraints on consumer electronic devices motivate us to investigate general-purpose video summarization techniques that are able to apply a common hardware setup to multiple content genres. On the other hand, we know that high quality summaries can only be produced with domain-specific processing. In this paper, we present a time-series analysis based video summarization technique that provides a general core to which we are able to add small content-specific extensions for each genre. The proposed time-series analysis technique consists of unsupervised clustering of samples taken through sliding windows from the time series of features obtained from the content. We classify content into two broad categories, scripted content such as news and drama, and unscripted content such as sports and surveillance. The summarization problem then reduces to finding either finding semantic boundaries of the scripted content or detecting highlights in the unscripted content. The proposed technique is essentially and event detection technique and it thus best suited to unscripted content, however, we also find applications to scripted content. We thoroughly examine the trade-off between content-neutral and content-specific processing for effective summarization for a number of genres, and find that our core technique enables us to minimize the complexity of the content-specific processing and to postpone it to the final stage. We achieve the best results with unscripted content such as sports and surveillance video in terms of quality of summaries and minimizing content-specific processing. For other genres such as drama, we find that more content-specific processing is required. We also find that judicious choice of key audio-visual object detectors enables us to minimize the complexity of the content-specific processing while maintaining its applicability to a broad range of genres.

SPIE Conference Multimedia Content Analysis, Management and Retrieval

Blind Summarization: Content-Adaptive Video Summarization Using Time-Series Analysis

Ajay Divakakran, Regunathan Radhakrishnan, Kadir A. Peker

TR2006-026 January 2006

Abstract

Severe complexity constraints on consumer electronic devices motivate us to investigate general-purpose video summarization techniques that are able to apply a common hardware setup to multiple content genres. On the other hand, we know that high quality summaries can only be produced with domain-specific processing. In this paper, we present a time-series analysis based video summarization technique that provides a general core to which we are able to add small content-specific extensions for each genre. The proposed time-series analysis technique consists of unsupervised clustering of samples taken through sliding windows from the time series of features obtained from the content. We classify content into two broad categories, scripted content such as news and drama, and unscripted content such as sports and surveillance. The summarization problem then reduces to finding either finding semantic boundaries of the scripted content or detecting highlights in the unscripted content. The proposed technique is essentially and event detection technique and it thus best suited to unscripted content, however, we also find applications to scripted content. We thoroughly examine the trade-off between content-neutral and content-specific processing for effective summarization for a number of genres, and find that our core technique enables us to minimize the complexity of the content-specific processing and to postpone it to the final stage. We achieve the best results with unscripted content such as sports and surveillance video in terms of quality of summaries and minimizing content-specific processing. For other genres such as drama, we find that more content-specific processing is required. We also find that judicious choice of key audio-visual object detectors enables us to minimize the complexity of the content-specific processing while maintaining its applicability to a broad range of genres. We will present a demonstration of our proposed technique at the conference.

SPIE Conference Multimedia Content Analysis, Management and Retrieval

Blind Summarization: Content-Adaptive Video Summarization Using Time-Series Analysis

Ajay Divakakran, Regunathan Radhakrishnan, Kadir A. Peker

TR2006-026 January 2006

Abstract

Severe complexity constraints on consumer electronic devices motivate us to investigate general-purpose video summaarization techniques that are able toa pply a common hardware setup to multiple content genres. On the other hand, we know that high quality summaries can only be produced with domain-specific processing. In this paper, we present a time-series analysis based video summarization technique that provides a general core to which we are able to add small content-specific extensions for each genre. The proposed time-series analysis technique consists of unsupervised clustering of samples taken through sliding windows from the time series of features obtained from the content. We classify content into two broad categories, scripted content such as news and drama, and unscripted content such as sports and surveillance. The summarization problem then reduces to finding either finding semantic boundaries of the scripted content or detecting highlights in the unscripted content. The proposed technique is essentially and event detection technique and it thus best suited to uncripted content, however, we also find applications to scripted content. We thoroughly examine the trade-off between content-neutral and content-specific processing for effective summarization for a number of genres, and find that our core technique enables us to minimize the complexity of the content-specific processing and to postpone it to the final stage. We achieve the best results with unscripted content such as sports and surveillance video in terms of quality of summaries and minimizing content-specific processing. For other genres such as drama, we find that more content-specific processing is required. We also find that judicious choice of key audio-visual object detectors enables us to minimize the complexity of the content-specific processing while maintaining its applicability to a broad range of genres. We will present a demonstration of our proposed technique at the conference.

SPIE Conference Multimedia Content Analysis, Management and Retrieval

Blind Summarization: Content-Adaptive Video Summarization using Time-Series Analysis

Ajay Divakaran, Regunathan Radhakrishnan and Kadir. A. Peker
Mitsubishi Electric Research Laboratory,
Cambridge, MA 02139
E-mail: {ajayd,regu,peker}@merl.com

ABSTRACT

Severe complexity constraints on consumer electronic devices motivate us to investigate general-purpose video summarization techniques that are able to apply a common hardware setup to multiple content genres. On the other hand, we know that high quality summaries can only be produced with domain-specific processing. In this paper, we present a time-series analysis based video summarization technique that provides a general core to which we are able to add small content-specific extensions for each genre. The proposed time-series analysis technique consists of unsupervised clustering of samples taken through sliding windows from the time series of features obtained from the content. We classify content into two broad categories, scripted content such as news and drama, and unscripted content such as sports and surveillance. The summarization problem then reduces to finding either finding semantic boundaries of the scripted content or detecting highlights in the unscripted content. The proposed technique is essentially an event detection technique and is thus best suited to unscripted content, however, we also find applications to scripted content. We thoroughly examine the trade-off between content-neutral and content-specific processing for effective summarization for a number of genres, and find that our core technique enables us to minimize the complexity of the content-specific processing and to postpone it to the final stage. We achieve the best results with unscripted content such as sports and surveillance video in terms of quality of summaries and minimizing content-specific processing. For other genres such as drama, we find that more content-specific processing is required. We also find that judicious choice of key audio-visual object detectors enables us to minimize the complexity of the content-specific processing while maintaining its applicability to a broad range of genres. We will present a demonstration of our proposed technique at the conference.

1. INTRODUCTION

The goals of multimedia content summarization are two-fold. One is to capture the essence of the content in a succinct manner and the other is to provide top-down access into the content for browsing. Towards achieving these goals, signal processing & statistical learning tools are used to generate a suitable representation for the content using which summaries can be created. For content that is carefully produced & edited (scripted content) such as news, movie, drama, etc., a structure is imposed during the production of the content in terms of semantic units such as news stories, scenes etc. Therefore, a representation that captures the sequence of semantic units that constitute the content would be useful. The user can browse the content using abstractions of each of the semantic units such as keyframes, skims etc. Hence, past work on summarization of scripted content has mainly focussed on coming up with a Table of Contents (ToC) representation as shown in Figure 1. It has shown that the representation units starting from the "keyframes" up to the "groups" can be detected using unsupervised analysis. However, the highest level representation unit requires content-specific rules to bridge the gap between semantics & the low/mid level analysis[1][2][3].[4]

In unscripted content such as sports & surveillance, interesting events happen sparsely in a background of usual events. Therefore, if the analysis is focused on detecting specific events of interest, a summary can be generated using a combination of what is typical in the content and what are "interesting" events in the content. Hence, we proposed a hierarchical representation for unscripted content as shown in Figure 2.[5]

Past work has shown that the play/break representation for sports can be achieved by an unsupervised analysis by bringing out repetitive temporal patterns.[6] However, the rest of the representation units require

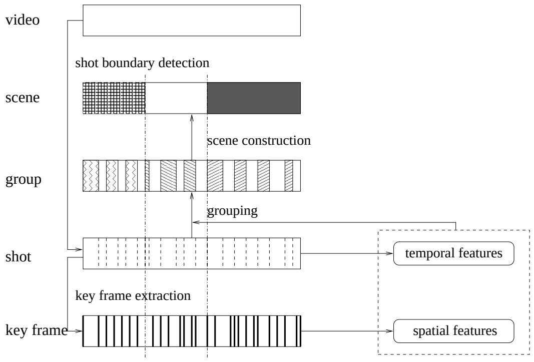

Figure 1. A hierarchical video representation for scripted content

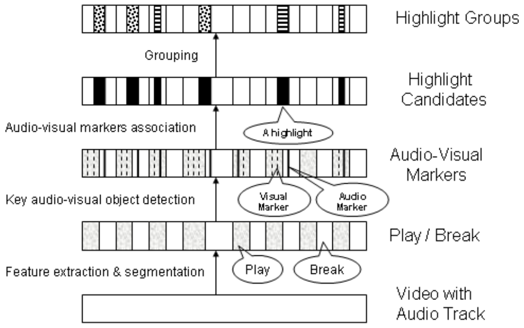

Figure 2. A hierarchical video representation for unscripted content

the use of domain knowledge in the form of supervised audio-visual object detectors that are correlated with the events of interest.

Clearly, from the past work for summarization of scripted and unscripted content one can see the need for content-specific processing at different levels. This necessitates a separate analysis framework for each domain. However, severe complexity constraints on consumer electronic devices motivate us to investigate general-purpose video summarization techniques that are able to apply a common hardware setup to multiple content genres. On the other hand, we know that high quality summaries can only be produced with domain-specific processing. Therefore, what is more desirable from this practical standpoint, is a content-adaptive analysis and representation framework that postpones content specific processing to as late a stage as possible. In this paper, we propose one such framework based on time-series analysis of audio features that provides a general core to which we are able to add small content-specific extensions for each genre.

The rest of the paper is organized as follows. In the next section, we present the motivation for the analysis framework and describe its component blocks. In section 3, we show its application to various genres and finally conclude with possibilities for further research.

2. TIME SERIES ANALYSIS BASED APPROACH

In this section, we propose a content adaptive analysis and representation framework which does not require any a priori knowledge of domain of the unscripted content. It is aimed towards an inlier/outlier based representation of the content for event discovery & summarization as shown in Figure 3. It is motivated by the observation that “interesting” events in unscripted multimedia occur sparsely in a background of usual or “uninteresting” events. Some examples of such events are:

- sports: A burst of overwhelming audience reaction in the vicinity of a highlight event in a background of commentator’s speech.

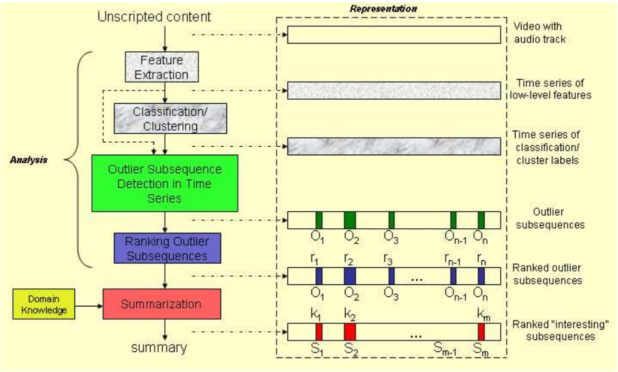
Representation
Unscripted content
Video with audio track
Feature Extraction
Time series of low-level features
Classification/ Clustering
Time series of classification/ cluster labels
Analysis
Outlier Subsequence Detection in Time Series
Outlier subsequences
O1 O2 O3 On-1 On
Ranking Outlier Subsequences
r1 r2 r3 rn-1 rn
Ranked outlier subsequences
O1 O2 O3 On-1 On
Domain Knowledge
Summarization
k1 k2 km
Ranked "interesting" subsequences
summary
S1 S2 Sm-1 Sm

Figure 3. Proposed content-adaptive analysis & representation framework for summarization of unscripted content

- **surveillance**: A burst of motion and screaming in the vicinity of a suspicious event in a silent or static background.

Based on the aforementioned observations, we focus on audio analysis for the inlier/outlier based temporal segmentation. We briefly describe the role of each component in the proposed content audio analysis framework as follows.

- **Feature extraction:** In this step, low-level features are extracted from the input content in order to generate a time series from which events are to be discovered. For example, the extracted features from the audio stream, could be Mel Frequency Cepstral Coefficients (MFCC) or any other spectral/cepstral representation of the input audio.
- **Classification/Clustering:** In this step, the low-level features are classified using supervised models for classes that span the whole domain to generate a discrete time series of mid-level classification/clustering labels. One could also discover events from this sequence of discrete labels. For example, Gaussian Mixture Models (GMMs) can be used to classify every frame of audio into one of the following five audio classes which span most of the sounds in sports audio: Applause, Cheering, Music, Speech, Speech with Music. At this level, the input unscripted content is represented by a time series of mid-level classification/cluster labels.
- **Detection of subsequences that are outliers in a time series:** In this step, we detect outlier subsequence from the time series of low-level features or mid-level classification labels motivated by the observation that "interesting" events are unusual events in a background of "uninteresting" events.

  Given the problem of detecting times of occurrences of observations that are not from the usual background process, we propose the following time series clustering framework:

  1. Sample the input time series on a uniform grid. Let each time series sample at index 'i' (consisting of a sequence of observations) be referred to as a context, C_i.
  2. Compute a statistical model M_i from the time series observations within each C_i, as an estimate for the generative background process.
  3. Compute the affinity matrix for the whole time series using the context models and a commutative distance metric $(d(i,j))$ defined between two context models (M_i & M_j). Each element, A(i,j), in the affinity matrix is $e^{\frac{-d(i,j)}{2\sigma^2}}$ where σ is a parameter that controls how quickly affinity falls off as distance increases .
  4. The computed affinity matrix represents an undirected graph where each node is a context model and each edge is weighted by the similarity between the nodes connected by it. Then, we can use a spectral graph theory to identify distinct clusters of context models & "outliers context models" that do not belong to any of the clusters. For a detailed description, please refer[7] .

At this level, the input content is represented by a temporal segmentation of the time series into inlier and outlier subsequences. The detected outlier subsequences are illustrated in the Figure 3 as $O_i, 1 \leq i \leq n$.

- **Ranking outlier subsequences:** In order to generate summaries of desired length, we rank the detected outliers with respect to a measure of statistical deviation from the inliers. At this level, the input content is represented by a temporal segmentation of the time series into inlier and ranked outlier subsequences. The ranks of detected outlier subsequences are illustrated in the Figure 3 as $r_i, 1 \leq i \leq n$.
- **Summarization:** The detected outlier subsequences are statistically unusual. All unusual events need not be interesting to the end-user. Therefore, with the help of domain knowledge, we prune the outliers to keep only the interesting ones & modify their rank. For example, commercials & highlight events are both unusual events and hence using domain knowledge in the form of a supervised model for audience reaction sound class will help in getting rid of commercials from the summary. Furthermore, note that the training data for the sound class that correlates with the interesting event can be systematically acquired from the detected set of outliers. At this level, the input content is represented by a temporal segmentation of the time series into inlier and ranked "interesting" outlier subsequences. The "interesting" outlier subsequences are illustrated in the Figure 3 as $S_i, 1 \leq i \leq m$ with ranks k_i. The set of "interesting" subsequences ($S_i^{'}$)s is a subset of outlier subsequences ($O_i^{'}$s).

3. APPLICATION TO VARIOUS GENRES

In this section, we briefly describe the application of the proposed blind summarization framework to emphasize its content-adaptive nature. We will also specify the particular content-specific extension for each of the genres so as to generate an "interesting" summary for that genre.

3.1. Sports Highlights Extraction

We applied the proposed framework for inlier/outlier based segmentation of a total of 12 hours of Soccer, Baseball and Golf content from Japanese, American & Spanish broadcasts.[8] Since there is a burst of audience reaction in the vicinity of highlight moments in sports audio, all of the highlight moments were a part of the detected outliers. There were also other outliers that are statistically unusual but not interesting. Commercial segments and periods of the game during which the commentator is silent but the audience are cheering are some example cases which are statistically unusual and not "interesting". Therefore, after this stage one needs to use a supervised detector to pick out only the "interesting" parts for the summary. The training data for the audio class that is indicative of highlight moments was acquired from the detected set of outliers. The corresponding audio class turned out to be a mixture of audience cheering and commentator's excited speech. The only content specific processing used was a GMM trained to detect this class to prune out other types of outliers.

3.2. Event Detection in Surveillance

We applied the proposed analysis to a collection of elevator surveillance audio data for suspicious event detection.[8] The data set contains recordings of suspicious activities in elevators as well as some event free clips. Since most of the suspicious events are outliers in a background of usual events, we were motivated to apply the proposed outlier subsequence detection framework for the task of inlier/outlier based segmentation of the surveillance content. Again, all the suspicious events were a part of the detected outliers and false alarms can be eliminated by the use of supervised detectors. In this case, the set of outliers correspond to banging, footsteps, non-neutral speech, normal speech, sound of elevator bell, sound of elevator door opening and closing etc. Of all these classes, only banging and non-neutral speech correlate with suspicious events in elevators. Hence the content-specific processing stage just has supervised detectors for these two audio classes. With our limited test-set, we are able to catch all the suspicious events with no false alarms.

3.3. Scene Segmentation in Situation Comedies

We are motivated to apply the outlier subsequence detection framework for this genre based on the observation that a music clip is played at the end of every scene in the situation comedy content. We applied the proposed analysis framework to this genre to detect all scene transitions as a part of the detected outliers.[8] Some examples of other outliers include laughter tracks. Then, by selecting outliers that only correspond to music, we have achieved successful scene segmentation. Here the music class detector serves as the content-specific processing stage.

3.4. Application to News video

The proposed analysis framework when applied to news content was able to detect commercials successfully as they are outliers in the background of whole news program. However, for news story segmentation one would need to use more content-specific processing and cues from multiple modalities as in.[2]

The aforementioned examples show the effectiveness of the proposed framework in postponing content-specific processing. Another way to achieve common processing for genres, is to carefully select some key audio-visual objects that work across different genres. For audio analysis, speech and music are examples that can cover a variety of genres. For video analysis, faces are key visual objects that can cover a number of genres.

4. CONCLUSIONS

We proposed a content-adaptive analysis framework that postpones content-specific processing to as late a stage as possible. It is based on the analysis of a time-series features extracted from the input audio. The framework provides a general core to which we are able to add small content-specific extensions for each genre. We showed its effectiveness for various genres including sports, surveillance, situation comedies and news. The results are promising and show the feasibility of content-adaptive blind video summarization.

In our future work, we will work on reducing the computational complexity of the proposed analysis framework. We would also work on the analysis of time series of features extracted from video.

5. ACKNOWLEDGEMENTS

The authors would like to thank Dr.Ziyou Xiong for his close collaborations and discussions on sports highlights extraction.

REFERENCES

1. Rainer Lienhart , "Automatic text recognition for video indexing," *Proc. ACM Multimedia*, 1996.
2. Winston Hsu and Shih-Fu Chang, "A statistical framework for fusing mid-level perceptual features in news story segmentation," *Proc. of ICME*, 2003.
3. A.Aner and J.R.Kender , "Video summaries through mosaic-based shot and scene clustering," *Proc. European Conference on Computer Vision*, 2002.
4. Y.Li and C.C.Kuo , "Content-based video analysis,indexing and representation using multimodal information," *Ph.D Thesis, University of Southern California*, 2003.
5. Z.Xiong, Y. Rui, R.Radhakrishnan, A.Divakaran, T.S.Huang , "A unified framework for video summarization, browsing and retrieval," *Handbook of Image & Video Processing, Al Bovik, Academic Press*, 2nd edition.
6. L.Xie, S.-F.Chang, A.Divakaran, H.Sun , "Unsupervised mining of statistical temporal structures in video," *Video Mining, Azriel Rosenfeld, David Doermann, Daniel Dementhon Eds, Kluwer Academic Publishers*, 2003.
7. R.Radhakrishnan, A.Divakaran, Z.Xiong and I.Otsuka, "A content-adaptive analysis & representation framework for audio event discovery from "unscripted" multimedia," *submitted to Eurasip Journal on Applied Signal Processing*, 2005.
8. R.Radhakrishnan, "A content-adaptive analysis & representation framework for video summarization using audio cues," *http://isis.poly.edu/ regu/ReguThesis.pdf*, Dec. 2004.