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

An Extended Framework for Adaptive Playback-Based Video Summarization

K.A. Peker, A. Divakaran

TR2003-115 September 2003

Abstract

In our previous work, we described an adaptive fast playback framework for video summarization where we changed the playback rate using the motion activity feature so as to maintain a constant pace. This method provides an effective way of skimming through video, especially when the motion is not too complex and the background is mostly still, such as in surveillance video. In this paper, we present an extended summarization framework that, in addition to motion activity, uses semantic cues such as face or skin color appearance, speech and music detection, or other domain dependent semantically significant events to control the playback rate. The semantic features we use are computationally inexpensive and can be computed in compressed domain, yet are robust, reliable, and have a wide range of applicability across different content types. The presented framework also allows for adaptive summaries based on preference, for example, to include more dramatic vs. action elements, or vice versa. The user can switch at any time between the skimming and the normal playback modes. The continuity of the video is preserved, and complete omission of segments that may be important to the user is avoided by using adaptive fast playback instead of skipping over long segments. The rule-set and the input parameters can be further modified to fit a certain domain or application. Our framework can be used by itself, or as a subsequent presentation stage for a summary produced by any other summarization technique that relies on generating a sub-set of the content.

SPIE Internet Multimedia Managmeent Systems IV

Published: SPIE-ITCOM 2003-Internet Multimedia Management Systems IV, August 2003

An Extended Framework for Adaptive Playback-Based Video Summarization

Kadir A. Peker, Ajay Divakaran
Mitsubishi Electric Research Laboratories
{peker, ajayd}@merl.com
201 Broadway, Cambridge, MA 02139
Tel: 617-621-7500 Fax: 617-621-7550

ABSTRACT

In our previous work, we described an adaptive fast playback framework for video summarization where we changed the playback rate using the motion activity feature so as to maintain a constant "pace". This method provides an effective way of skimming through video, especially when the motion is not too complex and the background is mostly still, such as in surveillance video. In this paper, we present an extended summarization framework that, in addition to motion activity, uses semantic cues such as face or skin color appearance, speech and music detection, or other domain dependent semantically significant events to control the playback rate. The semantic features we use are computationally inexpensive and can be computed in compressed domain, yet are robust, reliable, and have a wide range of applicability across different content types. The presented framework also allows for adaptive summaries based on preference, for example, to include more dramatic vs. action elements, or vice versa. The user can switch at any time between the skimming and the normal playback modes. The continuity of the video is preserved, and complete omission of segments that may be important to the user is avoided by using adaptive fast playback instead of skipping over long segments. The rule-set and the input parameters can be further modified to fit a certain domain or application. Our framework can be used by itself, or as a subsequent presentation stage for a summary produced by any other summarization technique that relies on generating a sub-set of the content.

Keywords: Video summarization, adaptive playback, visual complexity.

1.Introduction

Video content is becoming more and more pervasive. As the technological infrastructure, from storage to transmission, processing power, display technologies, etc. increases in power, availability and effectiveness, video content becomes available in larger volumes and extent. Video is an opaque medium, due to its serial nature. Unlike images, or even text, a video content is not open to grasp with a quick glimpse. Research in recent years on browsing, indexing and summarization of video has been aimed at making video content a more transparent medium. The goal is to make navigating through video and finding a desired target easier and more effective.

In recent years, several video summarization approaches have been introduced. One of the approaches is based on reducing redundancy by clustering video frames and selecting representative frames from clusters [1,2,3]. Another approach is using a measure of change in the video content along time, and selecting representative frames whenever the change becomes significant [4,5]. Finally, there have been approaches based on assigning some significance measure to the parts of the video -- usually based on criteria inspired from the human visual system -- and subsequently filtering less significant parts [6]. Sundaram uses Kolmogorov complexity of a shot as its visual complexity and establishes a relationship with the complexity and the comprehension time. He uses analysis of film syntax along with the visual complexity, to construct video skims [7].

In terms of the presentation style, we can identify two main categories of video summaries: still image-based summaries, and motion summaries. Many of the above approaches can be used to generate either of these types of summaries. Various other visualization options such as video mosaics have also been proposed.

We have previously presented an adaptive fast playback-based video summarization framework [8]. The playback rate was modified so as to maintain a constant "pace" throughout the content. We assumed the motion activity descriptor, which is the average magnitude of the motion vectors in mpeg video, to provide a measure of the "pace" of the content. This approach can be viewed as a bandwidth allocation scheme, where we are given a measure of the "visual bandwidth" of the video and a channel bandwidth defined by the human visual system. Since the motion activity measure we use is linearly proportional with the playback rate of the video, we can linearly increase or decrease the visual bandwidth of the video by changing the playback rate. Hence, we achieve the optimum time-visual bandwidth allocation by adaptively changing the playback rate so as to have a constant motion activity during the playback of the video. As opposed to keyframe based summaries or those constructed by concatenating selected clips, adaptive fast playback is based on playing the whole content, but in a shorter time, so as to maximize the intake of visual information in the shortest time possible. In that form, it is based more on low level visual characteristics of video and the early vision processes in the perception, than in semantic content and cognitive processes in perception.

In this paper, we extend the visualization-based method of fast playback with a number of semantic features, and "summarization templates" fitted for specific classes of video segments. We use camera motion, face detection, and audio classes as the new features in our framework. The audio feature in our current setup consists of classifying each short segment of audio (.5 - 1sec) as silence, applause, male/female speech, speech with music, or music. These are the classes that we use in our general framework for detection of highlights in sports video [9] and news video browsing [10]. However, the class set can be adapted to include the key semantic classes for the target application, such as laughter detection for comedy-sitcom programs or explosion detection for action movies. We also introduce a novel measure of visual complexity that improves on motion activity that we previously used. We mainly consider the implementation of our summarization methods on consumer devices, hence we limit our selection of features and the algorithms we use to those that are low in complexity and can be implemented within the constraints of current consumer device platforms.

One general requirement for a video summary can be formulated as that it should have an effect on the viewer as close as possible to the original video. This requirement can be constrained based on domains or applications, as to which effect is of primary concern. We can recognize information transfer and invoking of certain emotions as two salient categories of effects that a piece of video content has on the viewer. Accordingly, we can imagine a summary trying to convey the essence of the information presented in the original video and/or arouse the emotional states on the viewer that the original video does. A similar but somewhat different categorization may be as (information) content and (presentation) style. A news program summary would be primarily concerned about conveying the main informative points whereas a movie teaser would be more of a stylistic sample. The summarization method we present here allows to a certain degree the flexibility of adjusting how much of these elements the final summary will carry.

2.Features

The selected features have different semantic levels; motion activity is a low-level feature whereas audio classes and face detection have a clearer semantic meaning, and camera motion is at a level between the two. Also, the features have different time scales or resolutions. For example, color is usually regarded as a short-range feature -- better for shot segmentation -- and has a correlation range of a few minutes, whereas audio -- especially speech -- is more of a scene level feature. Since we do not carry out speech recognition or face recognition/matching, our features mostly work at the shot level. We use these features to recognize shots as transition shots, talking-head shots etc. and mark certain semantically significant points in the shot. This particular selection of features is well suited for drama, for instance, where faces have an important place. However, we can also imagine the use of the same features for, for example, sports video summarization where camera motion would mark certain key actions and face detection marks the breaks and the introduction of players, etc.

We continue using low complexity measures, as was the case in our work on using motion activity. Our target platform is consumer devices, hence the complexity is of prime importance. Our features can be computed using the compressed domain data. Although we use full MPEG-1 frames for face detection, the DC image of HDTV or even MPEG-2 signals can be successfully used for face detection, since we are primarily interested in larger size faces. Although the features are computationally inexpensive, they are robust. For the sake of robustness and simplicity, we opt for face detection alone, without face matching at this time. In the same way, we use audio classification, which is more general and robust then speech recognition is.

2.1. Motion activity and visual complexity

In this paper, we introduce a novel measure of visual complexity based on motion activity and spatial complexity of a scene. The novel visual complexity measure is based on the limitation of the human visual system in perceiving spatial and temporal frequencies beyond a certain range [11]. It is essentially an estimate of the temporal bandwidth of a segment of an MPEG video computed using the DCT coefficients and motion vectors. The lower the temporal bandwidth, the faster the segment can be played within the perceptual limitations of the human visual system, and without aliasing due to temporal sub-sampling.

Assume a sinusoid $\cos(2\pi fx)$ is translating in x direction in time. If it is moving with velocity v, then the sinusoid can be expressed as $\cos(2\pi f(x - v \cdot t))$. Then, the frequency along the temporal axis is $f \cdot v$. If we look at a fixed point on the x-axis while the sinusoid is moving, we will see a sinusoidal change in the value of the function at that point over time. If the velocity v is high, the change will be faster, hence the frequency $f \cdot v$ of the sinusoid. The frequency in temporal dimension of a moving 2-D sinusoid is computed as the dot-product of the velocity vector **v** and the spatial frequency vector $\mathbf{f} = (f_x, f_y)$ (See figure 1). If the motion is perpendicular to the spatial frequency vector, in other words, if the sinusoid is moving parallel to the wavefront, there will be no change at a given point in time. If the motion is in the direction of the spatial frequency vector, which is the direction of fastest change, the change at a given point will be the maximum possible.

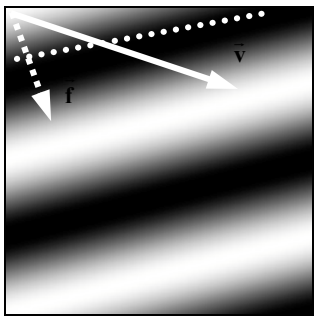

Figure 1. A 2-D sinusoid with a frequency vector **f** perpendicular to the wave front, and a motion vector **v** showing its translation velocity.

The DCT bases can be expressed as summation of two 2-D sinusoids:

$$\cos(2\pi \frac{k_x}{2N} x + 2\pi \frac{k_x}{4N}) \cdot \cos(2\pi \frac{k_y}{2N} y + 2\pi \frac{k_y}{4N})$$
$$= \frac{1}{2}\left[\cos\left(2\pi \frac{k_x}{2N} x + 2\pi \frac{k_y}{2N} y + 2\pi \frac{k_x + k_y}{4N} \right) + \cos\left(2\pi \frac{k_x}{2N} x - 2\pi \frac{k_y}{2N} y + 2\pi \frac{k_x - k_y}{4N} \right) \right]$$

We use the dot-product of the block motion vectors and the DCT coefficient numbers as the visual complexity value of each DCT coefficient. Each DCT coefficient in a macroblock with an associated motion vector, then, has an associated visual complexity. We then find the cumulative energy at each visual complexity value by summing the absolute DCT values of the DCT coefficients that have the same visual complexity. This gives us a feature vector similar to a power spectrum. We compute the final visual complexity feature as the mean visual complexity of the video frame. We use the DCT coefficients of the I frames and the motion vectors of the first P frame after each I frame.

2.2. Audio Classes

We are currently using "silence", "ball hit", "applause", "female speech", "male speech", "speech and music", "music", and "noise" audio classes. We segment the audio track into 1-second segments and use GMM-based classifiers for

labeling into 8 classes. These classes are used in our general framework for sports highlights detection and audio assisted video browsing [9,10]. The selection of classes can be changed to suit the application in hand, for example, by adding a laughter class for sitcoms.

2.3. Face Detection

We use the Viola-Jones face detector based on boosting [12]. The accuracy is good on MPEG-1 frame sizes (352x288), though with some misses or false alarms in some of the frames. We apply the detector to every other frame in the video, although the sampling could be coarser for faster processing. We construct a matrix where every column corresponds to a video frame. The columns are the same length as the width of the frames. The x-positions where a face detected is set to 1, and to 2, 3, etc. where multiple faces overlap on the x-axis. So, the matrix is a projection of the 3 dimensional video data on to the x-t space. We then apply median filtering on the t-axis to eliminate false detections or misses that appear sporadically and do not continue for many frames. We use the x-size of the faces as our feature. In fact, the detected faces are almost always equal in width and height, so this is approximately equal to the square root of the detected face area.

2.4. Other features: Cut detection and camera motion

We detect cuts using the software tool Webflix. The accuracy is close to 100% on the re-encoded version of the original stream (misc2 drama from MPEG7 test content set) using TMPGEnc public domain encoder. The original version of the content has noisy motion vectors, which affected the motion activity descriptors as well as the cut detector.

The camera motion descriptor consists of the translation parameters (cx, cy) and a zoom factor s, computed using the method described in [13]. The accuracy is good for the most part. For our purposes, camera motion and close-up object motion is similar in terms of visual characteristics, thus the occasional misdetection is not important. Since the method is based on compressed domain motion vectors and the accuracy of the vectors is not high, the simplicity of the camera motion model helps increase the stability.

3.Summarization Method

Video content, especially edited content, usually has a hierarchical structure. It is also natural for a video summary to have a hierarchical structure, such as at the highest plot level, then at scenes level, and then at shots and events level. We imagine browsing of video in a similar coarse to fine fashion.

In this paper, we primarily consider the shot level summarization and adaptive playback. We previously presented motion activity based sub-sampling and adaptive playback. Here we improve that method by considering certain higher level features such as camera motion, face detection and audio classification. Note that summarization of the program at the scene-level requires syntactic analysis and heuristics that are more domain-dependent than are the low level skimming techniques presented here. The summarization and skimming methods presented here can be augmented by other scene analysis methods to cover the whole hierarchy of video summarization and browsing.

A simple summary of a drama video can be obtained by showing the segments where the face sizes are relatively large. We implement this by detecting local maxima in the face size feature curve. We eliminate the maxima that are too close in time and value. The resulting skim shows the main characters and main dialogue scenes with well character visibility. The summary is mostly slow paced with minimal action, but has close character visibility.

A second simple summary can be generated by selecting the segments with relatively high motion and camera activity. We implement this summary in a way similar to the face detection-based summary. This second summary shows us main action segments, and transitional segments between settings or states. It also captures the dynamic aspects of the characters, and dramatic gestures and moves they make throughout the program. This summary is faster paced, especially compared to the previous one.

We combine these two aspects of a drama program through the face and motion descriptors. The face feature captures the static states of the drama such as dialogues and close-ups on main characters, where as the motion and camera

descriptors capture the transition states between the static states, as well as dramatic actions that give the program its dynamic aspect.

At the shot level, we consider summarization of each shot individually. We identify two key techniques: 1- Selection of semantically significant points in time by using face, camera motion and audio data. 2- Time allocation by using motion activity or visual complexity. Although our final output is a moving summary, we first determine key points in a given shot, and hence find keyframes. These keyframes can also be used as initial entry points into the content and then the moving summary is used for browsing back or forward in the content after that point.

In each shot, we look for local maxima in the face-size curve to find the points where the faces in the shot have the best visibility. These are static points in the final key frame set. We then look for any local maxima of the camera motion magnitude that are above a certain threshold that signals a significant action. Usually, the highest activity points have blurry and transitional images, hence not very well as key frames. However, these segments are important as they denote an action or transition. A video clip of the action segment included in the summary captures the dynamics of the characters and the action in the drama. For the keyframe-based interface, we mark the points right before and after the maxima as the keyframe points.

A shot may have both a large face and camera motion. If the local maxima of the two curves are close in time, we combine them as one segment in the summary. If the shot does not have any of the semantic features, then motion activity or visual complexity based sub-sampling is applied as a generic summarization method. Thus, a typical adaptive fast playback (or smart fast forward) will consist of visual complexity-based adaptive fast playback of the content, with instances of short duration (e.g. 1 sec) normal playback at highlight points that are detected using face or significant motion information, or any other domain specific semantic feature.

As we have mentioned before, the summarization can be aimed at providing a number of different end results, which we identify mainly as informative, and stylistic/emotional. Furthermore, the summary can be designed to convey a certain type of information (e.g. weather report in news), or certain type of emotional or style elements from a content (e.g. action scenes in a movie). In our framework, the general look and feel of the summary can be varied from slow paced and character-heavy to fast paced and action-heavy by playing with the thresholds for significant face size and significant motion magnitude. A lower threshold will result in more of that type of scene appearing in the final summary.

4.Sample Results

We apply the summarization method on a drama content from the MPEG7 test content set. The analysis and browsing tool used for visualizing the data with video, and for testing the summarization results is shown in Figure 2. Figure 3 shows the plot of face size, camera motion magnitude, and shot boundaries for a segment from the video.

A combination of the face key points and motion key points provides a balanced and fluid summary of the content. The two features overlap at certain segments. Misdetection of large faces at two points and high camera motion due to noisy content in a still image result in a few extra clips to be included in the summary. However, the errors are not distracting in the overall summary generated.

Figure 4 shows sample frames selected by the face criterion, and Figure 5 shows examples of activities captured by the motion criterion.

5.Discussions and Conclusion

We are currently considering other features for future work and for other types of content. For example the variance of the magnitude of the motion vectors can be used as an alternative motion activity feature. We observed that it mostly captures gestures, hand motion, and other activities that do not involve major translation of the central object or the camera. The tempo measure described in [14] can also be used for capturing high-paced segments of the content.

We are also investigating the use of speaker change detection for better segmentation of face segments into different speakers. A high accuracy speech detector can also help in capturing whole word utterances when selecting a segment in a shot with face appearance.

We specifically use computationally inexpensive, but robust and reliable features in our summarization methods. We use widely applicable semantic features such as faces and camera motion. We propose use of specialized summarization methods based on detection of significant semantic elements such as large faces or significant motion or camera activity. The methods are on most part content and genre independent. However, the domain knowledge can be used to make use of the available information in a more effective way; for instance, by noting that significant face frames capture the main characters in drama, or certain camera motion patterns capture ball hits in golf games, etc.

We observe that a summary that consists of largest face shots result in a slower paced summary with more character detail in certain aspects. A summary that is constructed with high camera/motion segments is a fast paced summary that conveys the motion in the drama. These complement the character face shots of the face feature with gestures. All in all, these features provide us with a complementary set of "highlight" points, with their unique significance and style. A summary of desired taste can be dynamically constructed with the right combination of these elements, based on user or application requirements.

ACKNOWLEDGEMENTS:

The authors want to thank Regu Radhakrishnan and Ziyou Xiong for the audio classification, Michael Jones for face detection software, and Lexing Xie for camera motion software.

REFERENCES

1. M.M. Yeung and B. Liu. "Efficient matching and clustering of video shots, " In ICIP '95, pages 338-341,1995.
2. D. Zhong, H. Zhang, and S.-F. Chang. "Clustering methods for video browsing and annotation, " In SPIE Storage and Retrieval for Image and Video Databases IV, pages 239-246,1996.
3. A.M. Ferman and A.M. Tekalp, "Efficient filtering and clustering methods for temporal video segmentation and visual summarization, " J. Vis. Commun. & lmage Rep., 9:336-351, 1998.
4. D. DeMenthon, V. Kobla and D. Doermann, "Video Summarization by Curve Simplification", ACM Multimedia 98, pp. 211-218, September 1998.
5. A. Divakaran, R. Radhakrishnan, and K.A. Peker, "Motion Activity based extraction of key frames from video shots," Proc. IEEE Int'l Conf. on Image Processing, Rochester, NY, Sept. 2002.
6. Y-F. Ma, L. Lu, H-J. Zhang, and M. Li, "A User Attention Model for Video Summarization, " ACM Multimedia 02, pp. 533 – 542, December 2002.
7. Sundaram, Hari, "Condensing Computable Scenes using Visual Complexity and Film Syntax Analysis, " ICME 2001, Aug. 22-27, Tokyo, Japan.
8. Peker, K.A.; Divakaran, A.; Sun, H., "Constant Pace skimming and Temporal Sub-sampling of Video Using Motion Activity", *IEEE International Conference on Image Processing (ICIP)*, Vol. 3, pp. 414-417, October 2001
9. Xiong, Z.; Radhakrishnan, R.; Divakaran, A., "Generation of Sports Highlights Using Motion Activity in Combination with a Common Audio Feature Extraction Framework", *IEEE International Conference on Image Processing (ICIP)*, To Appear September 2003
10. Radhakrishnan, R.; Xiong, Z.; Raj, B.; Divakaran, A., "Audio-Assisted News Video Browsing Using a GMM Based Generalized Sound Recognition Framework", *SPIE Internet Multimedia Management Systems IV*, To Appear September 2003
11. A. Watson, A. Ahumada, J Farrell, "Window of Visibility: a psychophysical theory of fidelity in time-sampled visual motion displays, " J. Opt. Soc. Am. A, Vol. 3, No. 3, pp. 300-307, Mar 86.
12. P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features, " In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, Kauai, HI, December 2001.
13. Yap-Peng Tan; Saur, D.D.; Kulkami, S.R.; Ramadge, P.J., "Rapid estimation of camera motion from compressed video with application to video annotation, " IEEE Trans. on Circuits and Systems for Video Technology, Volume: 0 Issue: 1, Feb. 2000, Page(s): 133 –146.
14. Brett Adams, Chitra Dorai, Svetha Venkatesh, "Finding the Beat: An Analysis of the Rhythmic Elements of Motion Pictures, " International Journal of Image and Graphics 2(2): 215-245 (2002)

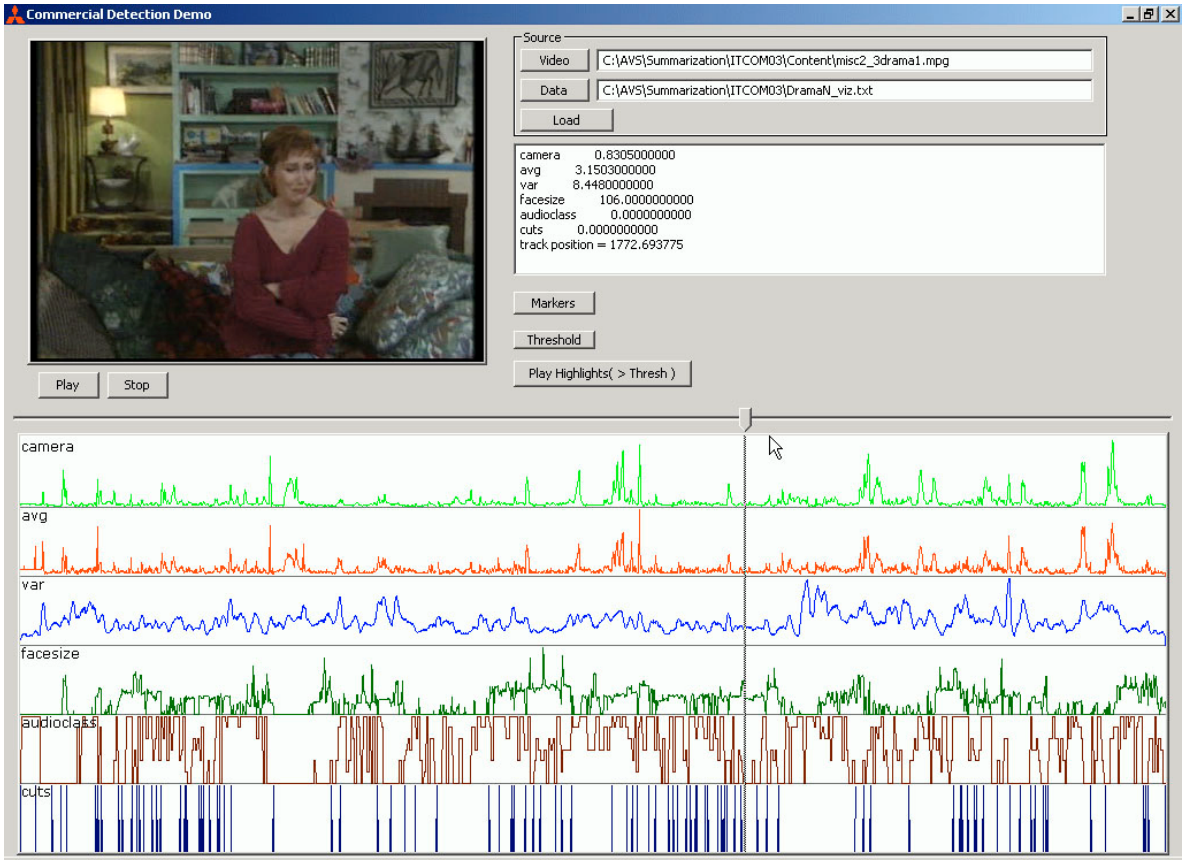

Figure 2. The analysis and browsing tool used to visualize the feature data and the video, showing the data for the drama segment from misc2 video in MPEG7 test content set.

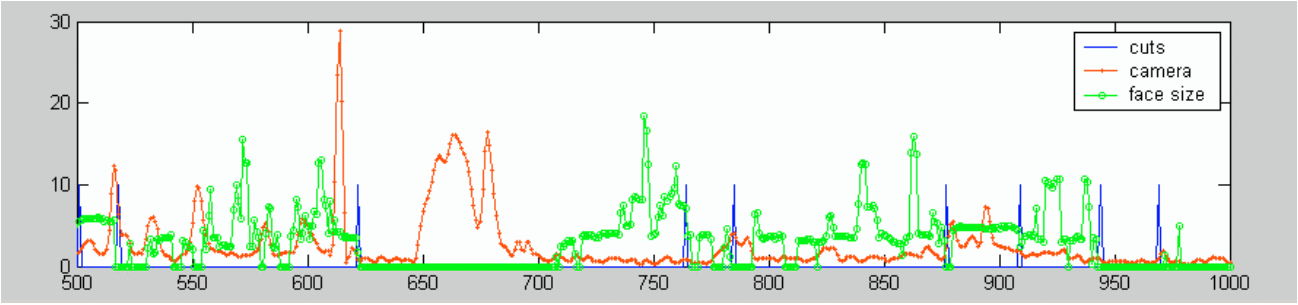

Figure 3. The plot of camera motion, face size, and shot boundaries for a segment of the drama content. We use local maxima of both the camera motion and the face curve above certain thresholds as key semantic points to be included in the summary. The proportion of character-heavy face segments and action-heavy motion shots in the summary can be adjusted by varying the respective thresholds.

Figure 4. Sample frames from the segments selected from the drama test content using the face size criterion.

Figure 5. Sample frames from the segments selected from the drama test content using the motion criterion.