A Web-Enabled Video Indexing System

Jian Zhou
Department of Electrical and Computer Engineering
Ryerson University
350 Victoria Street, Toronto, Ontario
Canada, M5B 2K3

jzhou@ee.ryerson.ca

Xiao-Ping Zhang
Department of Electrical and Computer Engineering
Ryerson University
350 Victoria Street, Toronto, Ontario
Canada, M5B 2K3

xzhang@ee.ryerson.ca

ABSTRACT

Video parsing and indexing is an important early stage of content-based video analysis. In this paper, we present a new web-enabled video indexing system that integrates Synchronized Multimedia Integration Language (SMIL) standard. New algorithms are proposed for video temporal segmentation. Sharp transition detection is achieved by an enhanced histogram-based method that is robust to illumination changes. For gradual transition detection, new features are introduced for dissolve detection. The proposed dissolve detector is based on a combined analysis of mean-variance-skewness of intensity. Compared with existing variance-based approaches, the introduced new features improve the discrimination ability on shot boundaries. We also describe methods for eliminating false positives. Experimental results show that the proposed algorithms can effectively detect shot boundaries. Detected scenes and other cinematic attributes are structured and organized by integrating HTML and SMIL. For each video file, the system generates a table-of-contents indexing file. The user-friendly interface provides web-based interaction, browsing and previewing of video content.

Categories and Subject Descriptors

H.5.1 [**Information Interfaces and Presentation**]: Multimedia Information Systems- *Video*; H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing.

General Terms

Algorithms, Design, Experimentation, Performance, Theory.

Keywords

Shot boundary detection, video temporal segmentation, video indexing, SMIL.

1. INTRODUCTION

Content based video analysis and retrieval has become an area of active research in recent years. The first step of video content analysis is to segment video into its constituent shots, and high-

level scenes or episodes. As the building blocks of video projects, video shots need to be identified effectively and efficiently. The research of video parsing focuses on the detection of both sharp shot transition (cuts) and gradual shot transition (fade/dissolve). Many automatic techniques have been developed to detect video shot boundaries in both the compressed and the decompressed domains. Zhang *et al.* [1] proposed pair-wise pixel comparison, likelihood ratio and histogram comparison for sharp transition detection. Edge changes can also be used as a good feature for shot detection [2]. Yeo and Liu [3] detected scene changes by using pixel difference and luminance histograms based on DC-images in compressed domains. In [4], recently developed methods for shot detection were reviewed in detail, and a statistical detector was proposed based on motion compensation. Shot detection techniques can be categorized into feature-based [2], model-based [5] and statistical [6] methods. Most of the above techniques can achieve good performance on hard cut detection.

However, compared with sharp transition detection, gradual transition still remains a challenging problem. When dealing with gradual transition, Zhang *et al.* [7] used a twin threshold mechanism based on histogram difference metric. Frame differences were accumulated when inter-frame difference was above the lower threshold but smaller than the higher threshold. When the accumulated difference exceeded the higher threshold, a gradual transition was defined. In [2], edge and contour changes were used for gradual transition detection. Another feature that is commonly used for dissolve detection is intensity variance. During a dissolve transition, the intensity variance curve forms a downwards-parabolic shape. The variance-based approach was first introduced by Alattar [5], and many other researchers have used this feature to build their dissolve detectors [4] [6]. Alattar [5] proposed to take the second order difference of intensity variance, and then check two large negative spikes. However, such pattern might not be pronounced due to object/camera motion and noise. Truong *et al.* [6] proposed an improved version by adding more constraints. Lienhart [8] introduced a somewhat different approach that includes a transition synthesizer and a neural network classifier, and dissolves are detected by a multi-resolution search.

Most of the existing techniques require careful selection of thresholds to achieve good performance. Some key factors that affect the performance of shot detection include illumination changes and object/camera motion. Since histograms do not carry spatial information, they are expected to be robust to object and camera motion. However, illumination changes can cause serious problems. Some feature-based methods, for example, those

algorithms based on the appearance of intensity edges, are less sensitive to illumination changes, but they are not robust to the motions of large objects and extra computations are also introduced.

In this paper, new algorithms are proposed for shot detection. We present a cut detection algorithm that is robust to illumination changes. Dissolve detection is achieved by a combined analysis of intensity moments. In addition to the variance feature, the introducing of mean and skewness adds more constraints and improves the discrimination ability of frame distances on shot boundaries, and thus provides a more robust way for shot detection Experimental results show that the proposed mean-variance-skewness approach can capture those dissolves whose intensity patterns are not so obvious if using only variance feature.

To structure and describe video content, media description scripts or template need to be developed. In [9], MPEG-7 and MPEG-21 were used to design a video personalization and summarization system. While MPEG-7 defines several levels of abstraction and provides a standard set of tools for describing multimedia content, its scope dose not focus on the storage, delivery or presentation of digital video. Considering the growing amounts of online digital video with the success of the Internet, it is desirable to incorporate web-based technologies in content based video analysis projects. In this work, we describe the use of the Synchronized Multimedia Integration Language (SMIL) standard for building the web-enabled video indexing system. In the proposed system, SMIL bridges the gap between the structure of video content and its web-based presentations. User friendly web-enabled interaction, integration and synchronization of video segments are realized by hybrid documents combining HTML and SMIL.

2. SYSTEM OVERVIEW

As shown in Figure 1, the web-enabled video indexing system includes frame-level playback, feature extraction, cut detection, dissolve detection, a SMIL generator and Graphic User Interface (GUI).
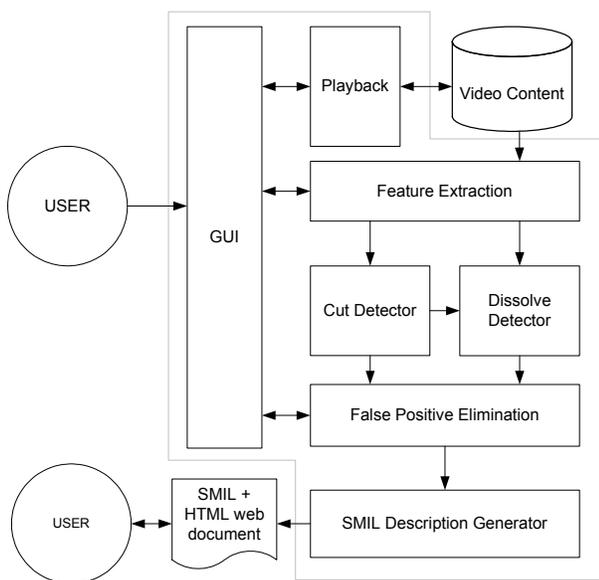

Figure 1: Architecture of a web-enabled video parsing system

After features are extracted from video content, cut detection is performed, followed by dissolve detection. Then, false positives are eliminated by a validation process. SMIL and HTML are used to describe the structures of video content including shot boundaries, video format, frame rate and other cinematic information. For each video clip, the system generates a web document as table-of-contents, which can be used for previewing and browsing video content online.

3. SHARP TRANSITION DETECTION

3.1 Improved Cut Detection

Histogram based methods are widely used for cut detection. Compared with other techniques, a histogram is robust to object motion, and it is simple and fast to calculate intensity or color histogram difference between two consecutive frames. Given a video sequence with N frames, denote the n-th frame as $f_n, n = 1, \ldots, N$. Let $H(f_n, k)$ denote the value of the k-th bin of the histogram for the frame f_n. Suppose there are total K bins in the histogram. Then the histogram difference at time n, denoted by $D_h(n)$, can be defined as follows:

$$D_h(n) = \frac{1}{A} \cdot \sum_{k=1}^{K} \left| H(f_{n+1}, k) - H(f_n, k) \right|, \quad n = 1, 2, \ldots N, \quad (1)$$

where A is the normalization factor. Histogram difference is a measure of dissimilarity between two consecutive frames. For cut transitions, visual characteristics are expected to change sharply at short boundaries, and thus the histogram difference can capture the visual discontinuities between shots. In practice, histogram based methods are the most common approach to shot detection, since they provide a good trade-off between accuracy and computational efficiency [1] [10] [11]. However, most histogram based methods are very sensitive to lighting changes. An example is shown in Figure 2. The luminance component is used to calculate the histogram difference. The first peak at frame 120 represents a real shot cut, and other peaks are caused by camera flashlight scenes. It can be seen that these illumination changes cause serious problems for cut detection.

In this paper, we propose a new histogram-based algorithm that is robust to lighting changes. The three channels in RGB color space are converted to the opponent color space. Only the red-green *(R-G)* component is used to compute the histogram difference. The opponent color representation of RGB color space is defined as [12]:

$$(R+G+B, R-G, 2 \cdot B - R - G), \quad (2)$$

where R, G and B are red, green and blue channels respectively. By choosing the opponent color space, the proposed cut detection algorithm is less sensitive to lighting changes. As shown in Figure 3, the performance is significantly improved. Experimental results show that our method performs better, compared with working only with chromatic color components in YCbCr color space.
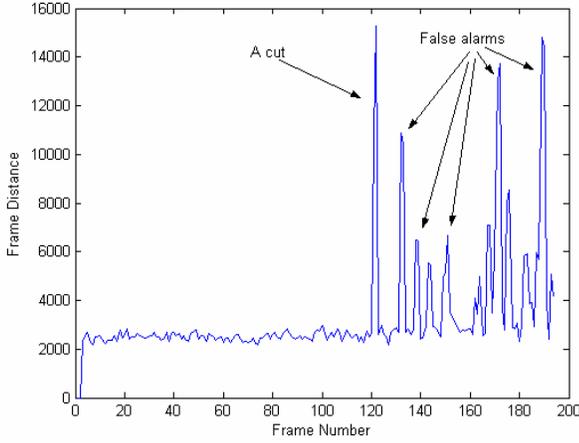
Figure 2: Histogram difference of luminance component in YCbCr color space.
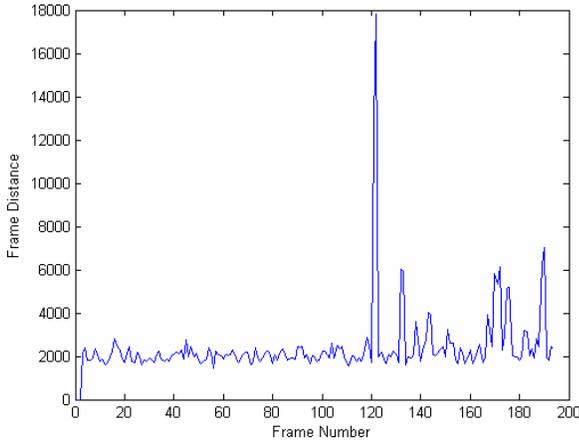
**Figure 3: Histogram difference of (R-G) component in the opponent color space.**

The conversion from RGB color space to its opponent color representation is computationally efficient. The advantage of this representation is that the last two chromaticity axes are invariant to changes in illumination intensity and shadows. The same video sequence is used to calculate the histogram difference in Figure 2 and Figure 3. It can be readily seen that the effects caused by flashlight scenes are significantly reduced.

Because the distribution of frame distances varies from video to video, adaptive threshold is more reasonable than a globally fixed threshold. It is worth mentioning that the normalization factor A defined in equation (1) is crucial for robust adaptive threshold, since the video data may have different image sizes. A temporal sliding window of the length $2w+1$ with $w=8$ and centered at current frame n is used to capture the local characteristics. A hard cut candidate is detected at frame n if the following conditions are satisfied:

1. $D_h(n)$ has the maximum value inside the sliding window, i.e., $D_h(n) \ge D_h(k), \forall k \in [n-w, n+w]$.

2. The difference between $D_h(n)$ and the median value of the sliding window is larger than a given threshold T_1.

An example based on the proposed cut detection algorithm is shown in Figure 4.
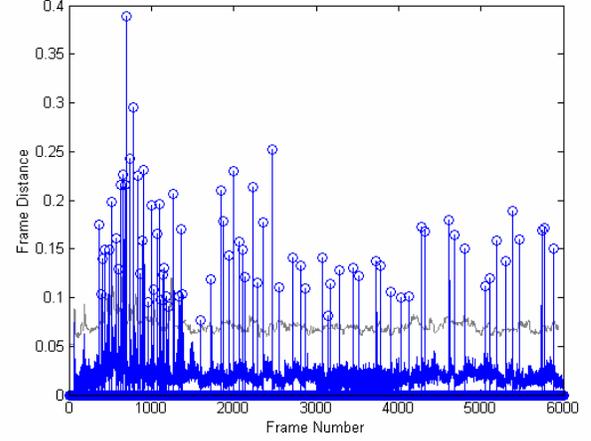
Figure 4: Cuts detected by adaptive threshold in the opponent color space (TV show "*Friends*").

3.2 False Positive Elimination

Due to camera/object motion and noise, some frames might be mistakenly identified as shot boundaries. Thus, a false positive elimination process is necessary. The validation process includes two criteria. First, a hard cut candidate at time n is declared as a false positive if frame $(n+2)$ is similar as frame $(n-2)$, i.e.:

$$\sum_{k=1}^{K} \left| H(f_{n+2},k) - H(f_{n-2},k) \right| < T_2, \tag{3}$$

where T_2 is a given threshold. Second, excluding the maximum value $D_h(n)$ of the sliding window, we calculate the left half maximum denoted as M_L and the right half maximum M_R. Let M_V denote the average of the two maximums. If the difference between $D_h(n)$ and M_V is less than a given threshold T_3, the cut candidate $D_h(n)$ is deemed as a false positive; that is:

$$M_L = \max\{\forall \, D_h(i), i \in (n-1-w, n-1)\}, \tag{4}$$

$$M_R = \max\{\forall \, D_h(i), i \in (n+1, n+1+w)\}, \tag{5}$$

$$M_V = (M_L + M_R)/2, \tag{6}$$

$$\left| D_h(n) - M_V \right| < T_3. \tag{7}$$

The above criteria can effectively eliminate the false positives. The first criterion define in (3) is to measure the difference of visual characteristics between two shots. The second criterion compares the maximum value at the center with the second and third peaks, and such criterion can remove the false positives caused by dissolves.

4. GRADUAL TRANSITION DETECTION

4.1 Mean-Variance-Skewness

The detection of gradual transitions is less mature, compared with cut detection. One of the main reasons is that it is difficult to define and capture the visual discontinuities for gradual transitions. Most of the recent research work focuses on dissolve detection, since dissolve is the dominant editing style in gradual transitions. In this paper, we also focus on dissolves instead of other types of gradual transitions. During a dissolve, intensity variance has been proven to show a parabolic shape [5]. However, such pattern might not be so obvious due to motion and noise. In the proposed algorithm, in addition to intensity variance, mean and skewness are introduced as new features. The first order (mean - μ), the second order (variance - σ) and the third order (skewness - s) intensity moments for the *n-th* frame are defined as:

$$\mu(n) = \frac{1}{M \cdot N} \sum_{i=1}^{M} \sum_{j=1}^{N} f(i,j,n), \qquad (8)$$

$$\sigma(n) = \left\{ \frac{1}{M \cdot N} \sum_{i=1}^{M} \sum_{j=1}^{N} [f(i,j,n) - \mu(n)]^2 \right\}^{1/2}, \qquad (9)$$

$$s(n) = \left\{ \frac{1}{M \cdot N} \sum_{i=1}^{M} \sum_{j=1}^{N} [f(i,j,n) - \mu(n)]^3 \right\}^{1/3}. \qquad (10)$$

Here $f(i,j,n)$ is the intensity value of the image pixel at location (i,j) for the *n-th* frame. M and N are image width and height respectively. The above moments are functions of time. Assuming a frame at time t is defined as $f(x,y,t)$, and *x, y, t* are continuous variables. A dissolve transition with duration of T can be considered as a mixture of two shots $f_1(x,y,t)$ and $f_2(x,y,t)$. During dissolve transition, the intensity of one shot decreases, and the intensity of the other increases. The dissolve editing style can be approximated by choosing two linear scaling functions $g_1(t)$ and $g_2(t)$ as:

$$g_1(t) = \frac{T-t}{T}, g_2(t) = \frac{t}{T}, \text{ where } t \in [0,T]. \qquad (11)$$

The dissolve sequence $D(x,y,t)$ for $t \in [0,T]$ can be defined as [13]:

$$D(x,y,t) = g_1(t) \cdot f_1(x,y,t) + g_2(t) \cdot f_2(x,y,t). \qquad (12)$$

Assuming two shots $f_1(\cdot)$ and $f_2(\cdot)$ are statistically independent and roughly ergodic random processes [5] [13]. The intensity variance is given by [5]:

$$\sigma^2(t) \equiv Var(D(x,y,t))$$
$$= B_2 \cdot t^2 + B_1 \cdot t + B_0, \qquad (13)$$

where coefficients, B_2, B_1 and B_0, are independent of t. Thus, ideally, before and after a dissolve transition, the variance is roughly constant, and during the transition, the variance curve forms a parabolic shape. Based on the same assumptions, the mean can be calculated as:

$$\mu(t) \equiv E[D(x,y,t)]$$
$$= E[g_1(t) \cdot f_1(x,y,t) + g_2(t) \cdot f_2(x,y,t)]$$
$$= g_1(t) \cdot \mu_1(t) + g_2(t) \cdot \mu_2(t) \qquad (14)$$
$$= \left(\frac{T-t}{T}\right) \cdot \mu_1 + \left(\frac{t}{T}\right) \cdot \mu_2$$
$$= A_1 \cdot t + A_0,$$

where coefficients, A_1 and A_0, are independent of t. Equation (14) shows that mean curve forms a line during a dissolve. Similarly, the skewness is given by

$$s^3(t)$$
$$= Skew(D(x,y,t))$$
$$= E[(D(x,y,t) - \mu_D)^3] \qquad (15)$$
$$\equiv E\{[g_1(t) \cdot f_1(x,y,t) + g_2(t) \cdot f_2(x,y,t) - \mu_1(t) - \mu_2(t)]^3\}$$
$$= C_3 \cdot t^3 + C_2 \cdot t^2 + C_1 \cdot t + C_0,$$

where the coefficients, C_3, C_2, C_1 and C_0 are independent of t. It can be seen that the skewness forms a cubical curve during a dissolve. Skewness characterizes the degree of asymmetry of the distribution around its mean. Two shots with different visual characteristics are expected to have different skewnesses, which are connected by a cubical curve during the dissolve. Another interpretation is that we can easily make connections between moments and distance. The first moment actually defines the *1-norm*, and the *k-th* moment is related to the *k-norm*. Thus, if we only consider the absolute value, the skewness feature is nothing but a number that measures the difference between the distribution and its mean based on a real metric, i.e.:

$$s = d_3(f,\mu) = \|f - \mu\|_3 = [E(|f - \mu|^3)]^{1/3}. \qquad (16)$$

The skewness that provides higher order information can be used as a feature for analyzing shot transitions. Figure 5 shows how variance and skewness are affected during a dissolve transition.
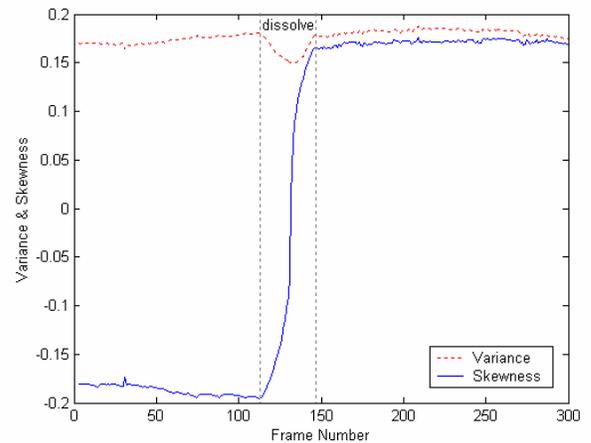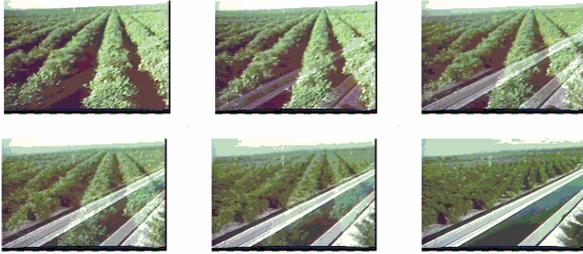

Figure 5: Variance and skewness curves during a typical dissolve.

It can be seen that intensity variance forms a parabola while skewness forms a cubical curve. Both features can capture the dissolve. But, numerically speaking, skewness provides higher

discrimination ability at the shot boundary since its value changes from -0.19 to 0.16 while variance only changes from 0.17 to 0.14.

Another example is shown in Figure 6. This transition contains some extreme factors such as fast camera motion and similar scenes between two shots. Variance and skewness features extracted from the video (see Figure 6) are plotted in Figure 7. In this case the parabola pattern of variance shown in Figure 7 is not obvious due to motion and noise. But skewness is still a good feature to identify the dissolve. When a dissolve joins two similar scenes, at the beginning of the dissolve, the intensity of one shot decreases, but at the same time, it is compensated by similar intensities from the other shot. Such situation can cause serious problems for variance-based approaches. But by exploiting higher order (such as the skewness curve) feature, it is still possible to capture such dissolve transitions.

Figure 6: An example of dissolve transition (frames 995, 1000, 1005, 1010, 1015 and 1025 of "*The Miracle of Water*").
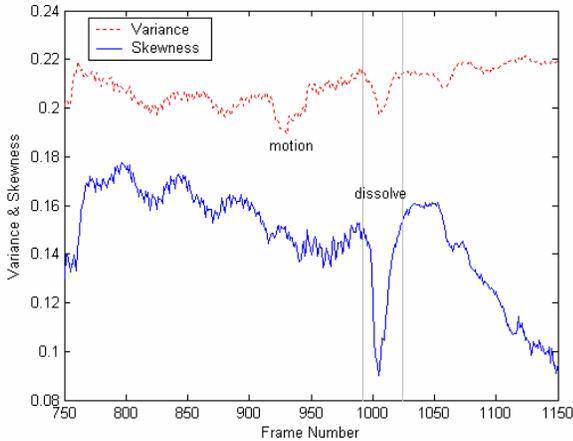
Figure 7: Variance and skewness curves during a dissolve.

In Figure 7, the cubical curve becomes a parabola-like curve when the cubical coefficient is close to zero.

4.2 Dissolve Detection

For dissolve detection, a new method based on a combined analysis of mean-variance-skewness is proposed in this paper. The dissolve detector takes the output of cut detector as input. Therefore, we assume there are no sharp transitions within each input video segment. A temporal sliding window of size *2w+1* with *w=16* and centered at current frame is chosen to adaptively detect the features. Dissolve detection is achieved by three parts consisting of a variance based detection, a mean-skewness

detection and a validation phase that is used to remove false positives.

During variance-based detection, a median filter with a length of three is applied to variance curve to remove noise. And then a dissolve transition at time *n* is detected if all the following conditions are satisfied:

1. Variance $\sigma(n)$ has the minimum value in the sliding window; that is

$$\sigma(n) \le \sigma(k), \forall\, k \in [n-w, n+w]. \quad (17)$$

2. Let $\sigma_{mean}[i,j]$ denote the mean value of $\sigma(k)$ between the interval $[i, j]$, i.e.:

$$\sigma_{mean}[i,j] \equiv mean\{\forall\, \sigma(k), k \in [i,j]\}. \quad (18)$$

In order to match the downwards-parabolic shape of variance curve within the sliding window, the left half of the curve should be decreasing while the right half should be increasing. From this, we have the following two conditions:

$$\sigma_{mean}[n-w, n-w/2] > \sigma_{mean}[n-w/2, n], \quad (19)$$

$$\sigma_{mean}[n, n+w/2] < \sigma_{mean}[n+w/2, n+w]. \quad (20)$$

3. To make sure the curve has deep "valley" and strong "shoulders", the following two conditions should be satisfied:

$$\sigma_{mean}[n-w-w, n-w] - \sigma(n) > T_4, \quad (21)$$

$$\sigma_{mean}[n+w, n+w+w] - \sigma(n) > T_4, \quad (22)$$

where T_4 is a given threshold.

4. A curve fitting with a degree of two is used to further match the parabola pattern in the sliding window. The performance is evaluated by the estimated quadratic coefficient $\hat{B}_2$ (see (13)) and the fitting error. We have

$$\hat{B}_2 > T_5, \quad (23)$$

$$\sum_{i=n-w}^{n+w} \left| \sigma(i) - \hat{\sigma}(i) \right| < T_6, \quad (24)$$

where T_5 and T_6 are given thresholds, and $\hat{\sigma}(i)$, $i \in [n-w, n+w]$, is the value of the estimated curve evaluated at point *i*.

For variance-based detection, the above four conditions are used to match the parabolic shape. A dissolve is detected if all conditions are satisfied.

The mean-skewness detection combines mean and skewness features for shot detection. First, a median filter with a length of three is applied to the skewness curve to remove noise. And then, the first order difference of skewness curve is calculated and its absolute value is used as input. Dissolve detection now becomes measuring the input and finding the large positive spikes. A sliding window with a length of *2w+1* with *w=16* is used to adaptively calculate the local properties. Mean curve is used to

validate the detected transitions. Let $s(k), \forall k \in [1, N]$, denote the skewness curve after being median-filtered. The first order difference $S(k), \forall\, k \in [2, N]$, is given by:

$$S(k) = \left| s(k) - s(k-1) \right|, \forall\, k \in [2, N].\qquad(25)$$

The frame at time *n* is declared as a dissolve boundary if all the following conditions are satisfied:

1. $S(n)$ has the maximum value in the sliding window; that is:

$$S(n) \geq S(k), \forall\, k \in [n - w, n + w].\qquad(26)$$

2. Denote $S_{median}(n)$ as the median value of the sliding window centered at current frame *n*; that is:

$$S_{median}(n) \equiv median\{\forall\, S(k), k \in [n - w, n + w]\}.\quad(27)$$

The difference between the median value and $S(n)$ should be greater than a given threshold T_7; that is:

$$S(n) - S_{median}(n) > T_7.\qquad(28)$$

3. A regression line is used to fit the mean curve $\mu(k), \forall\, k \in [n - w, n + w]$, inside the sliding window. This condition requires that the fitting error should be less than a given threshold T_8. From this, we have:

$$\sum_{i=n-w}^{n+w} \left| \mu(i) - \hat{\mu}(i) \right| < T_8,\qquad(29)$$

where $\hat{\mu}(i), \forall\, i \in [n - w, n + w]$ is the value of the estimated curve evaluated at point *i*.

A dissolve transition is detected if all three conditions are satisfied.

4.3 False Positive Elimination

In the false positive elimination phase, the shot lists obtained from variance-based method and mean-skewness method are merged into one list for further analysis. If the distance between two consecutive dissolves is less than the length of the sliding window *2w+1*, duplicate entries are defined. In that case, we merge the overlapped dissolves into one dissolve. For each dissolve in the shot list, the histograms from frame *(n+w)* and *(n-w)* are compared to validate the results. If their difference is less than a given threshold T_9, the dissolve is considered as a false positive.

$$\sum_{k=1}^{K} \left| H(f_{n+w}, k) - H(f_{n-w}, k) \right| < T_9.\qquad(30)$$

The elimination criterion is based on the assumption that the visual characteristics from two shots are expected to be different.

5. EXPERIMENTAL RESULTS FOR SHOT DETECTIONS

Extensive experiments are performed to test the proposed shot detectors. Two TV shows, "Friends" and "Sex and the city" were selected, and documentary video data were collected from Carnegie Mellon University's The Informedia Project at "The Open Video Project" [14]. Experimental results are presented in Table 1. Precision and recall for hard cuts were obtained as 93.4% and 97.4% respectively. For dissolve detection precision and recall were 73.7% and 82.4%. In TV show "Friends", the three false alarms are caused by object motion, and fade in/out effects. In the show "Sex and the city", fast motion blur is used to connect two scenes. In fact, all three false alarms for hard cuts were caused by such special editing effects. Even though they were counted as errors in our tests, we could argue that they are actually shot boundaries. The documentary video data contain many water scenes and camera motions from close-up to establishing shot. Editing effects such as zoom-ins, zoom-outs, and camera panning are also used extensively. As it can be seen from the table, the general performance of documentary is not as good as TV shows, especially for cut detection. Part of the reason is that some transitions join similar outdoor scenes.

Table 1: Detection results for hard cuts (H) and dissolves (D)

Test Data	Total		Missed		False	
	(H)	(D)	(H)	(D)	(H)	(D)
Friends	73	4	0	1	3	0
Sex & the City	53	3	0	1	3	4
Documentary	64	44	5	7	7	11
Total	190	51	5	9	13	15

In one video clip from the documentary data, among the six dissolve transitions, only one dissolve can be detected if using only variance feature. But all six dissolves can be successfully identified if skewness feature is added. To compare the results with other works, we refer to [4] [6] [13]. For dissolve detection, a precision of 75.1% and recall of 82.2% is reported in [6], and Lienhar [8] obtained a precision of 82.4% and recall of 75% by using Neural Networks. Hanjalic [4] reached a precision of 79% and recall of 83% with a smaller test set containing only 23 dissolves. Best results for dissolve detection still use intensity variance feature [4] [6]. Even though video data collected in the above works were carefully selected to contain as many effects as possible, the performance evaluation from different researchers is still based on different materials. However, by introducing mean-variance-skewness and the combined analysis of these new features, we present new patterns and criteria for analyzing dissolve transitions. The experimental results show that the proposed algorithms are effective for shot boundary detection. Also, the methods are computationally efficient.

6. A WEB-ENABLED INTEGRATED SYSTEM

A system tool is developed to integrate the proposed shot detection algorithms. The graphic user interface of the system is shown in Figure 8. The system provides a frame-level playback. Both manual shot detection and automatic shot detection are supported. After shots are automatically detected, users can edit the shot list, for example, to merge or split shots.

In the proposed system, Synchronized Multimedia Integration Language (SMIL) standard is chosen as multimedia content descriptor. SMIL is a web multimedia format developed by the

World Web Consortium (W3C) and released in 1998. SMIL provides a cross-industry support for synchronized multimedia integration [15]. It is built on Extensible Markup Language (XML) and allows users to write and publish interactive multimedia online. SMIL syntax and semantics can also be incorporated into other XML-based languages for multimedia timing and synchronization. A simple example of hybrid document combing HTML and SMIL is shown below.

```
<html xmlns:t="urn:schemas-microsoft-com:time">
<head>
<?import namespace="t" implementation="#default#time2">
</head>
<body>
...
<input id="button1" type="button"
value="preview" fill="freeze" />
<t:video style="width:100; height:80px;"
   src="./aquarium1.mpeg" clipBegin="00:00:00.000"
   clipEnd="00:00:08.068" begin="button1.Click"
   type="mpeg" />
...
</body>
</html>
```

After shot boundaries are detected, the shot list and other cinematic attributes are managed by a SMIL-based web document. The table-of-contents web-enabled indexing file generated by the tool is shown in Figure 9. Web users can browse and preview the video segments, and jump to a specified location from frame-level. During the implementation, we found that SMIL standard is an effective media description for video structuring and indexing, and its close connection to web makes it very convenient to build and present structured web-enabled multimedia content. Also, keywords and conceptual attributes can be embedded in the SMIL-based indexing file that could be used by existing text-based search engines to realize video web search request.

7. CONCLUSIONS
We have presented new algorithms for shot boundary detection. Cut detection is achieved by choosing the opponent color space that is robust to illumination changes. Dissolve detection is based on a combined analysis of mean-variance-skewness. By introducing these new features and criteria, the proposed dissolve detector has provided a new way to identify and analyze dissolve transitions. Experimental results show that the proposed algorithms can effectively detect both sharp transitions and dissolve transitions, and are computationally efficient. We also presented a system tool to structure and organize the detected shots. Shots and video information are managed and indexed by integrating SMIL web multimedia standard. That makes the system interoperable with existing web-based techniques. The generated indexing file provides functionalities like web-based user interaction, browsing and previewing of video content.

8. REFERENCES
[1] H. Zhang, A. Kankanhalli, and S. W. Smoliar. Automatic Partitioning of full-motion Video. *Multimedia Systems*, 1(1): 10-28, 1993.

[2] R. Zabih, J. Miller, and K. Mai. A Feature-Based Algorithm for Detecting and Classifying Scene Breaks. *Proc. ACM Multimedia 95*, San Francisco, CA, pp. 189-200, Nov. 1995.

[3] B. Yeo and B. Liu. Rapid scene analysis on compressed video. *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 5, No. 6, pp. 533-544, Dec. 1995.

[4] A. Hanjalic. Shot-Boundary Detection: Unraveled and Resolved? *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 12, No.2, pp. 90-104, Feb. 2002.

[5] A. M. Alattar. Detecting and Compressing Dissolve Regions in Video Sequences with a DVI Multimedia Image Compression Algorithm. *IEEE International Symposium on Circuits and Systems (ISCAS)*, vol. 1, pp. 13-16, May 1993.

[6] B. T. Truong, C. Dorai and S. Venkatesh. New Enhancements to Cut, Fade, and Dissolve Detection Processes in Video Segmentation. *Proceedings of the 8th ACM International Conference on Multimedia*, pp. 219-227, Nov. 2000.

[7] H. J. Zhang, C. Y. Low, S. W. Smoliar, and J. H. Wu. Video parsing, retrieval and browsing: An integrated and content-based solution. *Proc. of ACM Multimedia'95*, Nov. 1995.

[8] R. Lienhart. Reliable dissolve detection. *SPIE Storage and Retrieval for Media Database 2001*, vol. 4315, pp.219-230, Jan. 2001.

[9] B.L. Tseng, C.Y. Lin and J.R. Smith. Using MPEG-7 and MPEG-21 for personalizing video. *IEEE Multimedia*, vol. 11, No. 1, pp. 42- 52, Jan.-Mar. 2004.

[10] J. Boreczky and L. Rowe. Comparison of video shot boundary detection techniques. *Proc. SPIE Conference on Storage and Retrieval for Image and Video Databases*, vol. 2670, pp. 170-179, 1996.

[11] A. Nagasaka and Y. Tanaka. Automatic video indexing and full-video search for object appearances. *Visual Database Systems*, vol. 2, pp. 113-127, 1992.

[12] A.B. Poirson and B.A. Wandell. Pattern-Color Separable Pathways Predict Sensitivity to Simple Colored Patterns. *Vision Research*, vol. 36, No. 4, pp. 515-526, 1996.

[13] R. Lienhart. Reliable transition detection in videos: A survey and practitioner's guide. *International Journal of Image and Graphics (IJIG)*, vol. 1, no. 3, pp. 469-486, Aug. 2001.

[14] The Open Video Project [Online]. Available: http://www.open-video.org/

[15] J. Ayers *et al*. Synchronized Multimedia Integration Language (SMIL) 2.0. World Wide Web Consortium Recommendation, Aug. 2001 [Online]. Available: http://www.w3.org/TR/smil20/
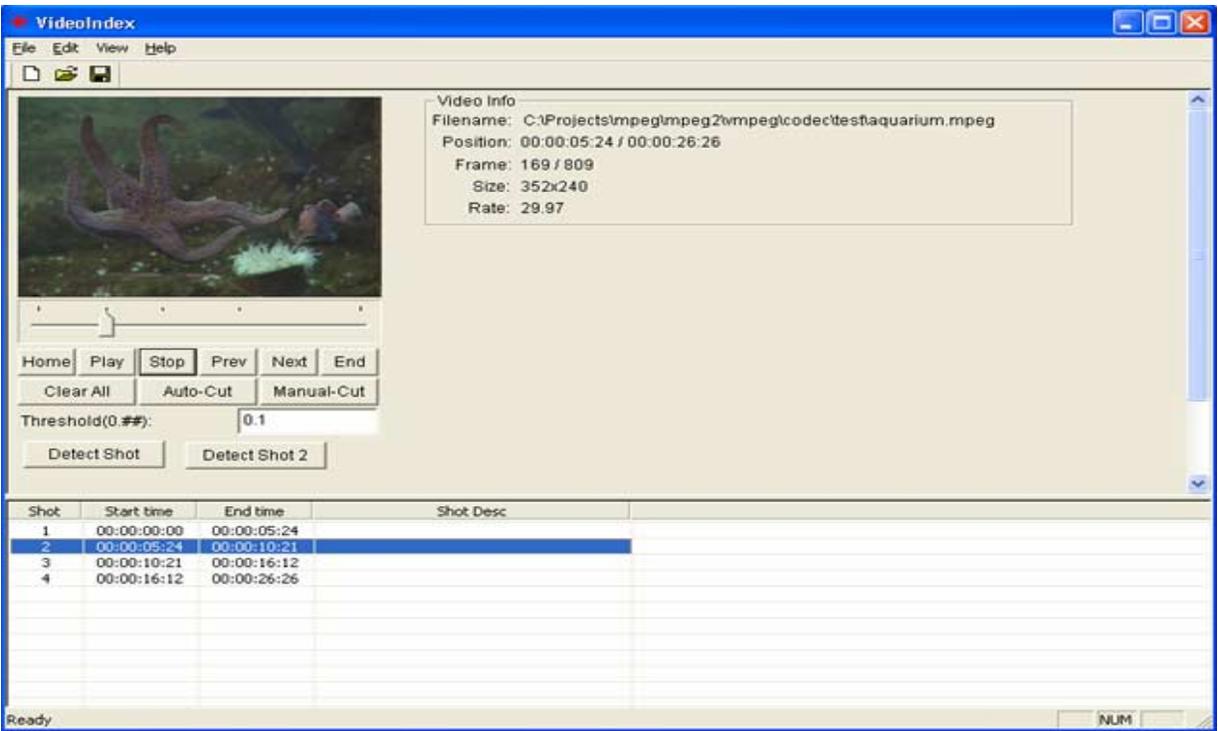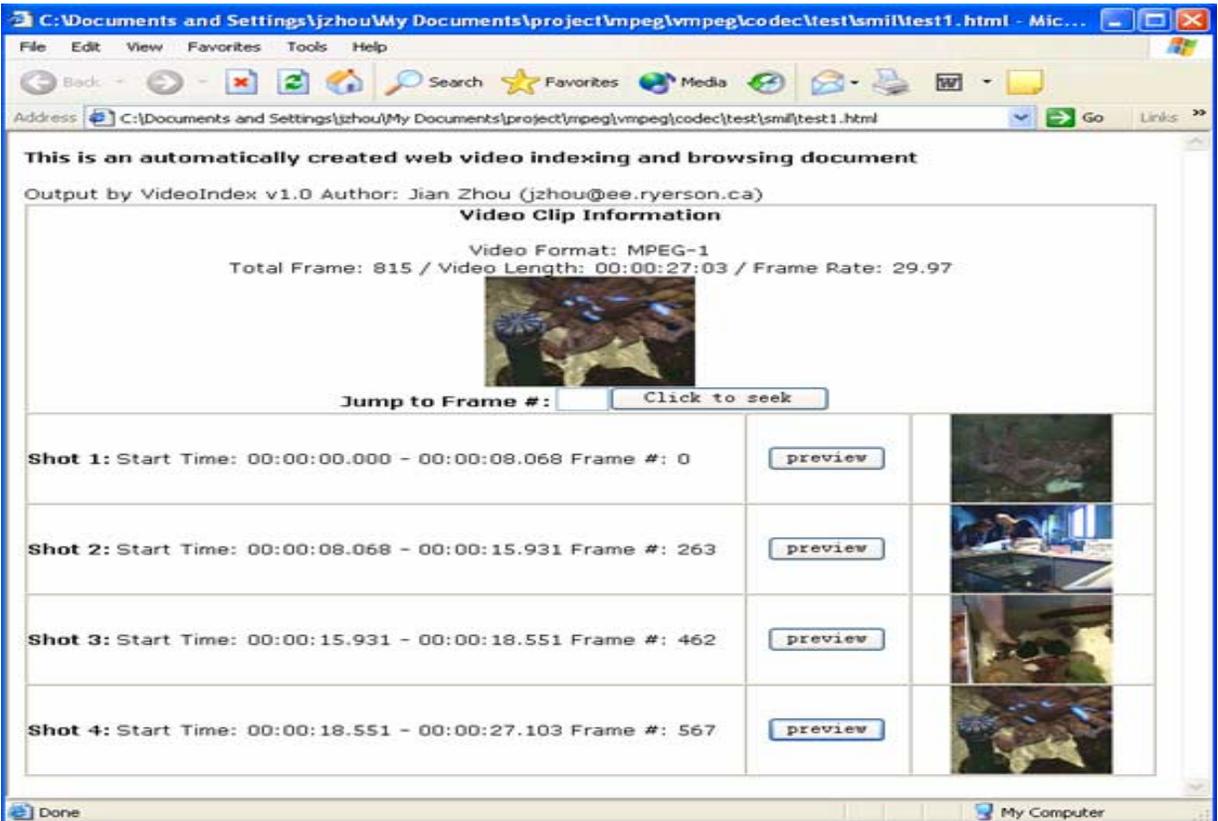
Figure 8: System Graphic User Interface.

Figure 9: Generated HTML+SMIL indexing file.