Automatic Video Scene Extraction by Shot Grouping

Tong Lin[1]

National Laboratory on Machine Perception
Peking University
Beijing 100071, China
Tonysc007@263.net

Hong-Jiang Zhang

Microsoft Research, China
No.5 Zhichun Road
Beijing 100084, China
hjzhang@microsoft.com

Abstract

For more efficient organizing, browsing, and retrieving digital video content, it is important to extract video structure information at both scene and shot levels. This paper presents an effective approach to video scene segmentation based on a pseudo-object-based shot correlation analysis. A new measure of the semantic correlation of consecutive shots based on dominant color grouping and tracking is proposed. A new shot grouping method named expanding window is designed to cluster correlated consecutive shots into one scene. Evaluations based on real-world sports video programs validate the efficiency and effectiveness of our shot correlation measure and scene structure construction.

1. Introduction

Efficient and automatic content organization and management of digital video is a key to the success of future video libraries, and various video applications on Internet has highlighted the need for smart content filtering and selective content delivery. Video structure parsing is the process to extract construction units of video programs. The structural information resulted from video parsing, especially semantically defined structures which appear to be more meaningful to human perception, is essential to automatic and content-based organization and retrieval of video,

There are usually two layers of construction units in video: shots and scenes (also often referred as story units). A shot consists of a sequence of frames recorded contiguously and representing a continuous action in time or space. A video scene consists of a sequence of semantically correlated shots. Earlier work in video structure analysis mostly focused on shot boundary detection, and numerous techniques have been proposed for parsing video streams into individual shots. After detecting shot boundaries, corresponding key frames can then be extracted to perform image analysis techniques for understanding shot content. While shot-based video analysis approaches provide users with better access than unstructured raw video stream, they are still not sufficient for meaningful video browsing and retrieval. Extracting scene structure information of videos will also facilitate hierarchical video abstraction, indexing and browsing.

There are a number of algorithms developed to detect scene boundaries in video sequences. Following the idea of examining boundary heterogeneities used in edge detection, an approach to scene boundary detection was proposed that determines a shot boundary being a scene boundary if color, motion and audio change simultaneously [1]. However, this approach does not consider a shot as a whole, and the beginning and ending frames contribute more to scene segmentation.

Most of others approaches attempt to merge similar and consecutive shots into scenes [2][3][4][5][6]. These approaches explore the internal homogeneity of a scene. Different shot similarity measures have been proposed, such as matching blocks between key frames [2], comparing color histograms between key frames [3] or mean color histogram [4], and comparing color histograms between any frame in two shot [6]. In other words, these approaches rely heavily on similarities between frames, either key-frames or individual frames in video shots. However, shot similarities defined by key-frames do not represent the temporal information completely. Also, comparing every pair of frames in two shots is very expensive computationally. More importantly, it is often because they are semantically correlated rather than visually similar when a sequence of shots is considered a scene. Therefore, scene detection approaches based on visual similarity of frames between two shots often do not produce good results, and what needed is a quantitative measure of semantic correlations between shots.

In this paper, we present an effective approach to video scene extraction, which consists of two significant new features. First, in contrast to previous works, we use shot correlation rather than similarity in grouping shots into scenes. For this, we have developed a new scheme to measure the semantic correlation of consecutive shots using dominant color grouping and tracking. The

[1] The work presented in this paper was performed while this author was working as a summer intern at Microsoft Research, China.

correlation measure depends not only on dominant colors of individual frames, but also their temporal variation. Therefore, this measure meets the nature of video as a temporal media. Also, a new shot grouping method named expanding window is designed to group correlated consecutive shots into scenes.

The rest of this paper is organized as follows. In Section 2, we first describe focus of our work, together with the definition of semantic correlation and considerations. Then, we present the new framework for measuring shot correlation, and the approach for scene structure construction using expanding window shot grouping method. In Section 3, we present experimental evaluation of the proposed approaches based on a data set real world sports programs. Concluding remarks are given in Section 4.

2. The Proposed Approach

A scene is defined as one or more consecutive shots that they are semantically correlated [1], or they all share the same "content" in terms of action, place and time [3]. While shots are marked by physical boundaries, scenes are marked by semantic boundaries, so scene boundary detection is a far more difficult task compared with shot boundary detection. **Figure 1** shows two examples of video scenes, each consists of a sequence of shots taken from the same place and in a successive order of time, and present an event.

(a)

(b)

Figure 1: Two examples of video scenes, each consists of a sequence of shots taken from the same place and in a successive order of time, and present an event.

In this paper, we define a scene as a sequence of consecutive shots that they are semantically correlated, or they share the same semantics in terms of time, place, objects or events. However, our task is to detect scene boundaries rather than semantic contents of scenes; thus, the key problem in scene segmentation is to determine if two shots are semantically correlated.

To solve this problem, we have focused on developing a measure of shot correlation without solving the difficult image understanding problems, such as object segmentation, recognition and tracking. In general, we can classify shots into two types: focusing on the environment, such as a street, without dominant foreground objects; or focusing on static or moving objects, such as a car or person. When a scene is composed by either one of these types of shots, or a combination of the two with a transition, there will be at least one aspect (dominant objects or background) in common between the shots. We define such a common aspect between two shots as their correlation. We find that color is an effective yet computational inexpensive feature to be used in representing such a correlation. As described in detail in the following, we have designed a measure based on dominant color grouping and tracking in a shot. The correlation measure depends not only on dominant colors of individual frames, but also their temporal variation. Therefore, this measure meets the nature of video as a temporal media. This new measure is distinctive from that proposed in previous works and is one of the two significant contributions presented in this paper.

2.1 Shot Correlation Measure

The color-based correlation measure between two shots, *a* and *b*, denoted as *cor (a, b)*, is calculated by dominant color object comparison and tracking between the two shots as following. This is achieve by first calculate the color histogram of each frame, from which dominant colors of the frame are identified, as described in the following.

We use the *HSV* color space in calculating color histograms since the *HSV* color space is natural and approximately perceptually uniform. Also, we can define a quantization of *HSV* to produce a collection of colors that is compact and complete. In our method, the HSV color space is quantized by a 3D Cartesian coordinate system with 10 values for X and Y, 5 values for Z (the lightness), respectively, as shown in **Figure 2**. This is because the *HSV* space is cylindrical, and the similarity between two colors given by indices $(h1,s1,v1)$ and $(h2,s2,v2)$ is given by the Euclidean distance between the color points $(x1,y1,z1)$ and $(x2,y2,z2)$, *respectively*, in the cylindrical *HSV* color space. The fineness of the color quantization will influent the extraction of dominant objects. A fine quantification will be able to discriminate more objects, while it may also cause the extraction of dominant objects being sensitive to lighting dominant objects between frames, which may result in loss of tracking of dominant objects.

To determine dominant colors of a video shots, pixels of each frame, or DC blocks in I frames when MPEG1/2 video are used, of the shot are projected into the quantized HSV color space. The normalized distribution of these pixels in the 3-D color space thus forms a normalized 3D color histograms of the frame. All dominant local maximum points in the 3-D color histogram are identified; and a sphere surrounding each local maximum point within a small neighborhood (with diameter of 3 quatization units) in the color space is defined as a color object. These colors objects (top 20 in our implementation) with the largest numbers of pixels are identified as dominant objects. These dominant objects capture the most significant color information of a frame and are more resilient to noise. We then form a 3-D *dominant color histogram*, $hist_d(k, x, y, z)$, for each frame by counting only pixels included in dominant color objects, where k denotes the frame number, and (x, y, z) denotes a color bin. It is worth noticing that we do not perform object segmentation in the spatial domain though the segmentation in *HSV* color space could be mapped back to a frame image, leading to a spatial segmentation; rather, we consider pixels falling into a dominant regions in the color space an object, which may (often) not represent a spatial object in a frame.

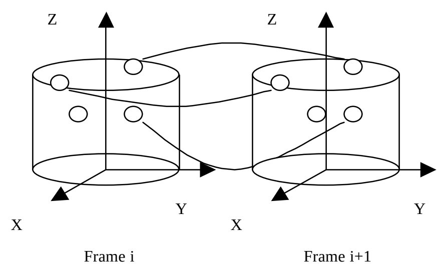

Figure 2: Color object segmentation and tracking.

Then, color objects defined as above in consecutive frames are tracked in the HSV color space to identify dominant objects of a shot. If the centers of two color objects in two consecutive frames are sufficient close, these two color objects are recognized as the same color object. Such a color tracking process will continue until all frames in the shot are tracked. After tracking, only the color objects that have longer duration in a shot are retained as dominant objects. In the words, we form an overall dominant color histogram for each shot, $hist_d^a(x, y, z)$ (a denotes a shot), consisting of only dominant color objects that are not only dominant in a frame, but also dominant across the entire shot. To give more weight to color objects with longer duration in a shot since they are more dominant, the histogram bins, corresponding to each dominant objects are weighted by its relative duration in a shot as,

$$hist_d^A(x, y, z) = hist_d^a(x, y, z) \times d_1/d_0 \qquad (1)$$

where d_0 is the duration of the shot, and d_1 is duration of the dominant color object with color (x, y, z). Also, $hist_d^A(x, y, z)$ is normalized by normalizing the mean size of each dominant color object within the shot. Therefore, the dominant color histogram of a shot represents both structural content in a frame and temporal content in a shot. Also, these dominant color objects often represent dominant objects or background in a shot and the correlation between these color objects in two shots is a good representation of correlations between the two shots. The correlation score between two shots, a and b, is calculated by performing the histogram intersection between two dominant color histograms of the two shots. That is,

$$\mathrm{Cor}(a, b) = \Sigma_x \Sigma_y \Sigma_z min[hist_d^A(x, y, z), hist_d^B(x, y, z)] \qquad (2)$$

This correlation score has the following properties:

1) $0 <= cor(a, b) <=1, cor(a, a) = 1$
2) $cor(a,b) = cor(b,a)$

2.2 Shot Grouping

A new method named expanding window is designed to group correlated consecutive shots into one scene based on the correlation scores as defined above. With this method, there is no need to compare many shot pairs and constructing complex links, as in [2, 6].

Considering the temporal constraints, i.e. shots that are closer to each other in time is more likely to belong to the same scene, the correlation score between two shots is weighted by temporal attraction factor:

$$w = 1/ (1+d / C) \qquad (3)$$

where d is the minimum distance between the two shots (from the ending frame of the previous shot to the beginning frame of the current shot) and C is a constant, determined by the average shot length.

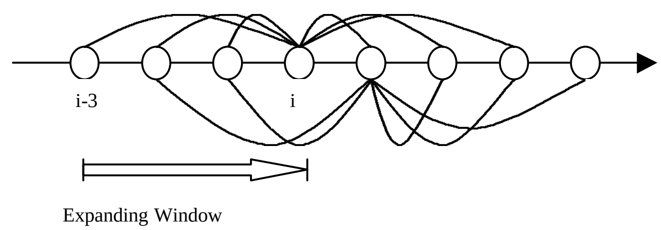

Figure 3. Expanding window shot grouping method, where shot *i* is the current new shot.

Assume every scene should contain at least 3 shots. Initially, the first 3 shots form a new scene, i.e. the size of expanding window is set to 3. Every time a new shot comes in, its correlation scores with the last 3 shots in the window is calculated and the maximum v, among the three correlation scores, is determined. Then, if

$$v > mean\text{-}var \qquad (4)$$

this shot is absorbed by the expanding window into the current scene. In (4), *mean* and *var* are the mean and variation of maximum correlation scores between shots contained in the current expanding window, respectively. Otherwise, we consider a few more subsequent shots for more confidence, as shown in **Figure 3**, because it is often that a scene may contain a shot that is uncorrelated with either previous or next shot. That is, we define an attraction ratio of the current shot *i* toward a new scene as

$$R(i) = (right(i)+right(i+1))/(left(i)+left(i+1)) \quad (5)$$

where

$$left(i)=max\{cor(i,i\text{-}1), cor(i,i\text{-}2), cor(i,i\text{-}3)\}$$
$$left(i+1)=max\{cor(i+1,i\text{-}1), cor(i+1,i\text{-}2)\}$$
$$right(i)=max\{cor(i,i+1), cor(i,i+2), cor(i,i+3)\}$$
$$right(i+1)=max\{cor(i+1,i+2), cor(i+1,i+3), cor(i+1,i+4)\}$$

If

$$R(i) > T \text{ and } R(i) > R(i\text{-}1) \text{ and } R(i) > R(i+1) \quad (6)$$

(where T is a threshold and we set T=1.5), then, the attraction to shot *i* from right side is greater than from left side, thus, the current shot *i* starts a new scene. Otherwise, the current scene absorbs this shot.

3. Experimental Results

We first tested the performance of the proposed correlation measure (SCM) in grouping similar shots, in comparison with those using mean color histogram (MCH) and keyframe color histogram (KCH). A shot database is used for comparing the proposed shot grouping method based the shot correlation measure, There are 419 shots from forty minutes TV sports news, including track and field events, swimming, soccer, basketball, etc. We pick out 8 shots each belongs to a scene as benchmark queries, and find *k* ground-truth correlated shots for each test. Also choose *k* as cut-off value, so recall is equal to precision. From Table 1, we could see that our solution outperforms the other commonly used ones.

To test the proposed shot grouping approach, we use 52 minutes soccer sequence from France98 World Cup to recognize different soccer matches. Here a scene is defined as all the consecutive shots in one match. There are 471 shots, and 40 different soccer matches, thus, there are 39 scene boundaries. Using the proposed algorithm, 61 scene boundaries are found, of which 30 are true boundaries, 9 missed, and 31 false alarms. Misses are caused by very similar grass color and lighting, which can only be recognized by even human viewer with very careful watching and with information from speech and close tracking of player's uniform colors. Alternating long shots and close-ups causes most of falsies, which are often reasonable for human visual perception. Figure 4 shows one example of corrected detected scene boundaries. It is also found that though that the proposed shot grouping method outperform other shot grouping methods, the performance will be increased significantly if audio information is integrated into the scene detection process, which is our next in developing the video scene segmentation system.

4. Conclusions

In this paper, we have presented a new measure for shot correlation and a method for applying this measure in grouping shots into scenes. The proposed method outperforms other key-frame or average color histogram based methods, though the performance needs further improvement. Since scene extraction requires more semantic information than just colors and their temporal variations, information from audio content analysis and segmentation will be helpful. How to integrate audio classification and segmentation with shot correlation analysis is our next step in developing a more robust scene extraction system.

5. References

[1] J. Huang, Z. Liu and Y. Wang, "Integration of Audio and Visual Information for Content-based Video Segmentation", *Proc. ICIP'98*, Chicago, Oct. 1998.

[2] A. Hanjalic, R. L. Lagendijk, and J. Biemond, "Automated high-level movie segmentation for advanced video-retrieval systems", *IEEE Transactions on Circuits and Systems For Video Technology*, Vol. 9, No. 4, pp. 580-588, June 1999.

[3] J. M. Corridoni and A. Del Bimbo, "Structured representation and automatic indexing of movie information content", *Pattern Recognition*, Vol. 31, No. 12, pp. 2027-2045, 1998.

[4] B. Gunsel, Y. Fu, and A. M. Tekalp, "Hierarchical temporal video segmentation and content characterization", in *Multimedia Storage and Archiving Systems II*, Proc. SPIE Vol. 3229, pp. 46-56, 1997.

[5] Y. Rui, T. S. Huang, and S. Mehrotra, "Exploring video structure beyond the shots", *Proc. IEEE Conf. on Multimedia Computing and Systems*, pp. 237-240, 1998.

[6] J. R. Kender and B. L. Yeo, "Video scene segmentation via continuous video coherence", *Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition*, pp. 367-373, 1998.

	k	*n*(SCM)	*n*(MCH)	*n*(KCH)
Soccer	8	5	6	2
Basketball	10	6	5	4
Marathon	5	3	2	1
Boat Race	4	2	1	3
Broad Jump	2	2	1	1
Swimming	9	9	9	9
Chess	9	6	6	5
Beach Volleyball	6	5	5	2
Total	53	38	35	27

Table 1. Comparison on different shot correlation queries

Figure 4; Detected scene boundary in a shot sequence caused by focus changing from long shots to close-ups. Each image is the first frame of one shot.