VIDEO SCENE EXTRACTION BY FORCE COMPETITION

Tong Lin[1+], Hong-Jiang Zhang[2], Qing-Yun Shi[1]

[1]National Laboratory of Machine Perception
Peking University
Beijing 100871, China
tonylin0602@sina.com

[2]Microsoft Research, China
49 Zhichun Road
Beijing 100080, China
hjzhang@microsoft.com

ABSTRACT

In this paper, we present a novel scheme for automatic video scene extraction. A pseudo-object-based shot representation containing more semantics is proposed to measure shot similarity and an force competition approach is proposed to group shots into scene based on content coherences between shots. Two content descriptors, color objects: Dominant Color Histograms (DCH) and Spatial Structure Histograms (SSH), are introduced. To represent temporal content variations, a shot can be segmented into several subshots that are of coherent content, and shot similarity measure is formulated as subshot similarity measure. With this shot representation, scene structure can be extracted by analyzing the splitting and merging force competitions at each shot boundary. Experiment on MPEG-7 test videos achieves promising results by the proposed algorithm.

1. INTRODUCTION

Video is a temporal media with huge amount of data, which cannot be easily organized and managed. Automatic video partitioning is the solution to segment video data into hierarchical structures, i.e. shots and scenes. A shot is a sequence of frames that are recorded contiguously, usually ended with a camera cut or an edit special effect. A video scene is referred to a group of consecutive shots taken place in the same location, or more generally, they share the same semantics in terms of time, place, objects or events.

Usually there are two steps to extract video scene structures after shot boundary detection. The first step is to represent visual content of one shot, and to define the similarity measure between two shots. The second is to construct scene structures by time-constraint shot clustering, or scene boundary detection. In most previous works on scene extraction [1][2][3][4][5], shot representation is rely on keyframe extraction and comparison, or some simple activity measure is attached to the keyframe-based shot representation scheme. Then, classical clustering algorithm or simple peak detection is used to detect scene boundaries. However, the limitation of keyframe-based shot representation is that spatio-temporal information of videos is not fully exploited. Also, when a sequence of shots is considered a scene, it is often because they are semantically correlated rather than visual similarity in term of keyframes.

In this paper, we present a scheme for automatic video scene extraction with a new shot representation scheme and a new scene boundary detection algorithm. Shots are represented with the intrinsic spatio-temporal relationships by analyzing the color and spatial content across time. Similar to [5], our approach first decomposes the temporal variations of a shot into several coherent sub-units called subshots. Subshots are indispensable for describing visual content of one shot that has significant content changes, such as panning from indoor to out of window. Unlike [5] that employs motion information to achieve this task, we utilize video color objects in a way that semantic content can be inherently embedded to describe the spatio-temporal changes of video content. To characterize visual content variations for subshot extraction and representation, two content descriptors, Dominant Color Histogram (DCH) and Spatial Structure Histograms (SSH), are developed based on extracted color objects. In addition, a new algorithm on scene boundary detection is provided by analyzing the competition of splitting and merging forces at each shot boundary.

The rest of this paper is organized as follows. In Section 2, we first introduce in detail the new shot representation, including DCH and SSH. Then we describe subshot extraction and shot similarity measure based on DCH and SSH descriptors. Section 3 presents the scene extraction approaches based on the proposed shot representation. Experimental results are given in Section 4 and the conclusion remarks are in Section 5.

2. SHOT REPRESENTATION

A new shot representation scheme that exploits the intrinsic spatio-temporal relationships and contains plentiful semantics is desirable to fulfill the task of scene extraction. This is because scene is a group of shots that are semantically correlated and is a subjective concept to reflect human perception. In this section, we present a pseudo-object-based shot representation. That is, we explore the spatio-temporal relationships of a "color object", which is defined as a color sphere in color space and a color-blob in the frame image. Color is an effective yet computational inexpensive feature, and integrating other information into color features would be preferred to improve representation capability without sacrificing efficiency. Two descriptors are introduced to describe the spatio-temporal behavior of color objects: (1) DCH that fuses temporal information into color histograms to capture the most important colors according to temporal variations; and (2) SSH that represents the spatial structural information for one individual frame, providing complementary functionality to color histograms that lack information about spatial distribution of colors. Finally, to compactly represent one shot, subshot is

[+] The work reported in this paper was mainly performed when this author was working at Microsoft Research, China as a research intern.

extracted by using DCH and SSH features, and shot similarity measure is formulated based subshot similarity measure.

2.1 Dominant Color Histogram

Color histogram is popularly used in content-based image retrieval (CBIR) for its simplicity and effectiveness. Therefore, it is natural to extend the idea of color histogram to content-based video retrieval. In [6], a set of color histograms called alpha-trimmed average histograms is introduced for one group of frames (GoF), including average histogram and median histogram. We propose dominant color histogram, achieved by dominant color extraction and tracking, to represent one GoF. Dominant color histogram for one GoF depends not only on dominant colors of individual frames, but also on their temporal variations. Therefore, this representation meets the nature of video as a temporal media. Dominant color histogram (DCH) is distinctive from previous work with incorporating temporal information and some semantic considerations.

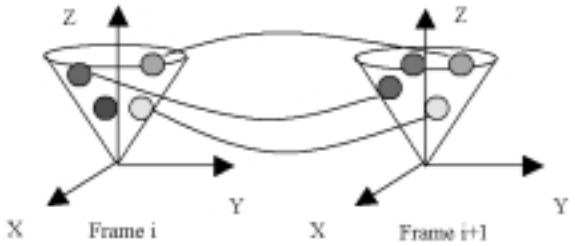

Figure 1: Dominant color extraction and tracking.

In general, we can classify one GoF (a group of frames in one shot that has single theme) 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. The focusing background or foreground objects should have longer duration. We want to capture the colors of focused objects by temporal variations and weigh them according to the temporal duration. Firstly, pixels of each frame, or DC blocks in I frames, are projected into quantized HSV cone space (shown in Figure 1). Then find out local maximum points, and a sphere surrounding each local maximum is defined as one color object in 3D color space. These color objects in consecutive frames are tracked to identify dominant color objects in one GoF, and more weight are given to the color objects with longer duration since they are more significant. Figure 1 shows this progress, and more details can be found in [7].

2.2 Spatial Structure Histogram

Spatial information would be very important to describe the global and local spatial configuration for one image. In [8] a new concept called image structural feature is introduced, which is a feature in-between texture and shape but more general. Water-Filling Algorithm is proposed to extract structural features from edge maps [8]. In this paper, we use color-blob maps to extract a new set of features called Spatial Structure Histograms (SSH) to describe spatial information for one video frame.

Firstly, color-blob maps are obtained by color quantization (examples shown in Figure 2). Each color cluster in 3D HSV cone space is extracted by K-means clustering due to its computational simplicity and efficiency. The optimal number of clusters, *k*, is obtained by using the cluster separation measure

$$\rho(k) = \frac{1}{k}\sum_{i=1}^{k} \max_{1\le j\le k, j\neq i} \{\frac{\eta_i + \eta_j}{\xi_{ij}}\} \quad (1)$$

where η_j is the intra-cluster distance of the cluster *j*, and ξ_{ij} is the inter-cluster distance of cluster *i* and *j*. Note that the cluster separation measure cannot handle the case of *k=1*. The K-means algorithm is tested for *k={1, 2, 3, …, 10}*. We choose the cluster number as *1*, if the intra-cluster distance is lower than some given threshold when *k=1*. Otherwise the cluster number is identified by the lowest value for *ρ(k), k>1*. In our implementation we use only the DC blocks of I frames, so ten color clusters would be enough to capture the color distribution. The potential problem is with texture regions that would create numerous small color regions. However, it is effectively depressed by the above cluster validity analysis that favors larger color clusters.

Figure 2: Examples of segmented color-blob maps.

$$\rho(k) = \frac{1}{k}\sum_{i=1}^{k} \max_{1\le j\le k, j\neq i} \{\frac{\eta_i + \eta_j}{\xi_{ij}}\} \quad (2)$$

where η_j is the intra-cluster distance of the cluster *j*, and ξ_{ij} is the inter-cluster distance of cluster *i* and *j*. Note that the cluster separation measure cannot handle the case of *k=1*. The K-means algorithm is tested for *k={1, 2, 3, …, 10}*. We choose the cluster number as 1, if the intra-cluster distance is lower than some given threshold when *k=1*. Or, the cluster number is identified by the lowest value for $\rho(k)$ when *k>1*. In our implementation we use only the DC blocks of I frames, so ten color clusters would be sufficient to capture the color distribution for the general DC images. The potential problem is with texture regions that would create numerous small color regions. However, it is effectively depressed by the above cluster validity analysis that favors larger color clusters.

Several distributional features are extracted from color-blob maps, including area histogram H_{area}, position histogram H_{pos}, deviation histograms in X and Y direction, H_{vx}, H_{vy}, and span histograms in X and Y direction, H_{sx}, H_{sy}. Area histogram is computed as

$$H_{area}(i) = \sum_{Rj\in\Omega i} Area(R_j), i = 0,1,...,7 \quad (2)$$
$$\Omega_i = \{R_j \mid Area(R_j) \in [A_i, A_{i+1})\}, i = 0,1,...,7$$
$$A_i = 1/2^{8-i}, i = 1,2,...,8; A_0 = 0$$

where *Area(R_j)* is the area percentage of color-blob R_j. Position histogram is defined as

$$H_{pos}(i) = \sum_{Rj\in\Omega i} Area(R_j), i = 0,1,2,...,15 \quad (3)$$
$$\Omega_i = \{R_j \mid Center(R_j) \in Block(i)\}, i = 0,1,2,...,15$$

where *Center(R_j)* is the centroid of color-blob R_j, and *Block(i)* is the i^{th} block with the image is equally divided into 16 blocks. Deviation histogram in X direction is defined as

$$H_{vx}(i) = \sum_{Rj \in \Omega i} Area(R_j), i = 0,1,2,...,7 \quad (4)$$
$$\Omega_i = \{R_j \mid \sigma_x(R_j) \in [B_i, B_{i+1})\}, i = 0, 1, 2,...,7$$
$$B_i = 1/2^{8-i}, i = 1,...,7, 8; \quad B_0 = 0$$

where $\sigma_x(R_j)$ is the standard deviation of color-blob R_j in x direction, normalized by the image width. H_{vy} is similarly defined except that $\sigma_y(R_j)$ is normalized by image height. Span histogram in X direction is defined as

$$H_{sx}(i) = \sum_{Rj \in \Omega i} Area(R_j), i = 0, 1, 2,...,7 \quad (5)$$
$$\Omega_i = \{R_j \mid Width(R_j) \in [B_i, B_{i+1})\}, i = 0, 1, 2,...,7$$
$$B_i = 1/2^{8-i}, i = 1,...,7, 8; \quad B_0 = 0$$

where $Width(R_j)$ is the width of minimum bounding rectangle of color-blob R_j, normalized by the image width. H_{sy} is similarly defined except that $Height(R_j)$ is normalized by image height.

Area histogram describes the spatial complexity of the image. Position histogram is desired to identify similar spatial configuration such as a close-up shot on the head and shoulder of one player. Deviation histograms and span histograms represent the shape distributions of the color-blob map. We found that the span histograms are not rotation-invariant and not as robust as others, so span histograms are not used in our experiments. The spatial similarity between two images, *a* and *b*, is computed as

$$SshSim(a,b) = W0 \times Sim0 + W1 \times Sim1 + W2 \times Sim2 + W3 \times Sim3 \quad (6)$$

where *Sim0*, *Sim1*, *Sim2*, and *Sim3* are histogram similarity on H_{area}, H_{pos}, H_{vx}, and H_{vy}, respectively, by using histogram intersection. *W0*, *W1*, *W2*, and *W3* are corresponding weights that are equally set in our experiments.

2.3 Subshots Extraction and Shot Similarity Measure

It would be better to segment one shot with significant content variations into several subshots, because the aggregated representation is unpredictable if we compose all the variations into one feature vector, such as for one shot panning from indoor to outdoor. We propose one simple subshot extraction algorithm based on color and spatial structure changes, because subshot should be of coherent visual content for compact representation. Suppose the percentage of newly emerged dominant color bins is *p1*, the difference of spatial structural histogram between the current and previous I frame is *p2*. A new subshot is identified if all the following conditions are true:

$$P1 > T1, \text{ and } P2 > T2, \text{ and } P1 + P2 > T3 \quad (7)$$

where *T1*, *T2*, and *T3* are predefined threshold (in our experiment they are empirically set as 0.2, 0.2, 0.6). The similarity measure of two shots, *a* and *b*, is defined as

$$Sim(a,b) = \max_{i,j}(Sim(a_i, b_j)) \quad (8)$$

where $Sim(a_i,b_j)$ is the similarity of the subshot *i* in shot *a* and the subshot *j* in shot *b*, which can be computed in two ways:

$$Sim(a_i, b_j) = Wc \times DchSim(a_i, b_j) + Ws \times SshSim(a_i, b_j) \quad (9)$$

DchSim() is the similarity on DCH of two subshots, and *SshSim()* is the similarity on average SSH for two subshots. *Wc* and *Ws* are the corresponding weights that are equally set in our experiment.

3. SCENE BOUNDARY DETECTION

With the similarity measure between two shots defined, we have developed a video scene boundary extraction approach based on a competition analysis of splitting and merging forces. That is, each shot is subject to two kinds of forces from its neighborhood. One is the splitting force in two directions, from the previous shots and the following shots, respectively. If the ratio of splitting forces in two directions is large, there could be a potential scene boundary. Another is the merging force between the previous and following shots, preventing the current shot being attracted by one side and trying to merge the previous and following shots into the same scene. A simple rule-based algorithm is used to detect the signal jumps on the two forces.

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

$$w = 1/(1+d/C) \quad (10)$$

where *d* is the time span 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 (20 seconds in our implementation).

We define an splitting force to shot *i* from the previous shots as

$$Fs(i) = left(i) / right(i) \quad (11)$$

where

$$left(i) = max\{sim(i,i-1), sim(i,i-2), sim(i,i-3)\}$$

$$right(i) = max\{sim(i,i+1), sim(i,i+2), sim(i,i+3)\}$$

The splitting force for current scene boundary candidate $(i|i+1)$ is defined as

$$Fs(i|i+1) = (Fs(i)+1/Fs(i+1))/2 \quad (12)$$

The physical meaning of splitting force $Fs(i|i+1)$ is the average of splitting force to shot *i* from previous shots and splitting force to shot *i+1* from following shots.

The merging force for shot *i* is defined as

$$\mathrm{Fm(i)} = \frac{1}{3}\sum_{j=i-3}^{i-1} \max(sim(j, i+1), sim(j, i+2), sim(j, i+3)) \quad (13)$$

Its physical meaning is the average attraction between previous shots and following shots to shot *i*. The merging force for current scene boundary candidate $(i|i+1)$ is defined as

$$Fm(i|i+1) = (Fm(i)+Fm(i+1))/2 \quad (14)$$

Then we normalize $Fs(i|i+1)$ and $Fm(i|i+1)$ to $Fs'(i|i+1)$ and $Fm'(i|i+1)$ with following linear scaling to unit variance:

$$x' = \frac{\frac{x-\mu}{3\sigma}+1}{2} \quad (15)$$

It is guaranteed that 99% of $Fs'(i|i+1)$ and $Fm'(i|i+1)$ is in [0,1].

At an ideal scene boundary, $(i|i+1)$, $Fs'(i|i+1)$ should reach its maximum and $Fm'(i|i+1)$ should reach its minimum. However, it is not always this case. Therefore, these two situations are treated differently as following:

- Condition 1: $Fs'(i|i+1)$ reaches its maximum and $Fm'(i|i+1)$ reaches its minimum simultaneously. A scene boundary is identified if $Fs'(i|i+1) > T1$;

- Condition 2: *Fs'(i|i+1)* reaches its maximum, or *Fm'(i|i+1)* reaches its minimum. If *Fs'(i|i+1)>T2* and *Fs'(i|i+1)-Fm'(i|i+1)>T3*, a scene boundary is declared.

where *T1*, *T2*, and *T3* are predefined thresholds (set as 0.6, 0.7, 0.2, respectively, in our experiments) and the neighborhood is set as [-2,2] to detect maximum and minimum.

4. EXPERIMENTAL RESULTS

Two MPEG-7 test videos are used to evaluate our algorithm on scene boundary detection: *lgerca_lisa_1.mpg* and *lgerca_lisa_2.mpg*. They are both home videos and each has approximately 32,000 frames. Experimental results are listed in Table 1 and 2. Figure 3 and 4 show the splitting and merging forces for each video. In most cases, splitting force reaches its maxima and merging force reaches its minima simultaneously at one ground-truth scene boundary. Three false scene boundaries in *lgerca_lisa_1.mpg* are caused by background differences and lighting condition variations. In *lgerca_lisa_2.mpg* five scene boundaries are missed, but the results are reasonable when considering that scene is a subjective concept and different human subject has different rules to extract scenes. The original ground-truth scene boundaries defined by source provider are arguable because scene 1 to 5 are taken in gym, scene 9 to 11 are on stage, and scene 13 to 14 are in swimming pool. In summary, the results are very promising compared with other approaches.

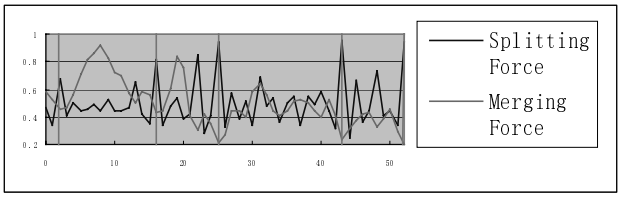

Figure 3: Force competition for lgerca_lisa_1.mpg, with ground-truth scene boundaries marked by a vertical line.

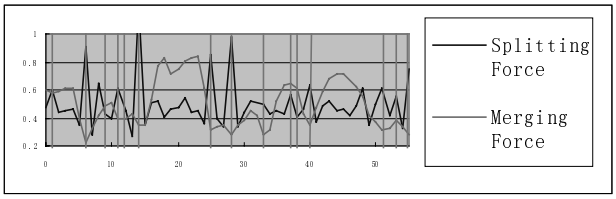

Figure 4: Force competition for *lgerca_lisa_2.mpg*, with ground-truth scene boundaries marked by a vertical line.

5. CONCLUSIONS

In this paper, we have presented an approach to extracting scene structure by using a new shot content representation and a new scene boundary detection algorithm. Our future work will be focused on integrating new motion descriptors into color-object-based shot representation to improve performance of the scene extraction algorithm.

REFERENCES

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] C. W. Ngo, T. C. Pong, H. J. Zhang, and R. T. Chin, "Motion-based video representation for scene change detection", *ICPR'00*, 2000.

[6] A. M. Ferman, et al, "Group-of-frames/pictures color histogram descriptors for multimedia applications", *ICIP'00*, 2000.

[7] T. Lin, H. J. Zhang, "Automatic video scene extraction by shot grouping", *ICPR'00*, 2000.

[8] X. S. Zhou, Y. Rui, and T. S. Huang, "Water-Filling: a novel way for image structural feature extraction", *ICIP'99*.

Scene	Scene Description	Shots	Correct	Miss	False
0	Kids learning roller-skater	0-1	1|2		
1	Kids playing in gym	2-15	15|16		12|13
2	Kid playing with water with parents	16-24	24|25		21|22
3	Hot balloon event	25-42	42|43		
4	Kids playing with parent on lawn	43-51	51|52		47|48
5	Indoor exercise	52	\	\	\

Table 1: Scene boundaries for *lgerca_lisa_1.mpg*.

Scene	Scene Description	Shots	Correct	Miss	False
0	Kids at home with cat	0	0|1		
1	Kids in gym	1-5	5|6		
2	Two kids playing high-bar (Over-illuminated)	6-8		8|9	7|8
3	Kid + teacher	9-10	10|11		
4	Kids jumping	11		11|12	
5	Kids in Gym	12-13	13|14		
6	Kids playing games in Gym (Over-illuminated	14-24	24|25		
7	Kids playing at home (Dim lighting)	25-27	27|28		
8	Kids driving outside home	28-32	32|33		
9	Kids dancing on stage (Part I of Play)	33-36		36|37	
10	(Part II of Play)	37		37|38	
11	(Part III of Play)	38-39	39|40		
12	After Play	40-50	50|51		
13	Swimming Pool	51-52		52|53	
14	Crowded Swim Pool	53-54	54|55		
15	Kid's party	55	\	\	\

Table 2: Scene boundaries for *lgerca_lisa_2.mpg*.