INTEGRATING COLOR AND SPATIAL FEATURES FOR CONTENT-BASED VIDEO RETRIEVAL

Tong Lin[1], Chong-Wah Ngo[2], Hong-Jiang Zhang[3], Qing-Yun Shi[1]

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

[2]Department of Computer Science
The Hong Kong University of Science & Technology
Clear Water Bay kowloon, HK
cwngo@cs.ust.hk

[3]Microsoft Research, China
49 Zhichun Road
Beijing 100084, China
hjzhang@microsoft.com

ABSTRACT

In this paper, we present a novel scheme for content-based video retrieval by exploring the spatio-temporal information. A shot with significant content changes can be segmented into several subshots that are of coherent content, and shot similarity measure for video retrieval can be computed from the similarity between corresponding subshots. To characterize the temporal content variations in one shot, we developed two descriptors: Dominant Color Histograms (DCH) and Spatial Structure Histograms (SSH). By fusing temporal information into color content, DCH for a "group of frames"(GoF) are trying to capture the dominant colors with long durations, which would be the colors of the focused objects or background. SSH is a set of features extracted from color-blob maps to describe spatial information for one individual frame. Experimental results on real-world sports video prove that our proposed approach achieve the best performance on the average recall (AR) and average normalized modified retrieval rank (ANMRR) for video shot retrievals.

1. INTRODUCTION

To date, most works on content-based video retrieval (CBVR) [1] deal with problems like video partitioning and keyframe extraction. The compact representation of video content for shot similarity measure and shot retrieval remains one of the most challenging issues. In the current literatures, video shots are mostly represented by keyframes. Low-level features such as color, texture, and shape are extracted directly from keyframes for indexing and retrieval. For efficiency reason, video retrieval is usually tackled in a similar way as image retrieval. Such strategy, however, is ineffective since spatio-temporal information existing in videos is not fully exploited.

Compact video representation requires not only keyframe selection [6] but also keyframe construction. For instance, a zoom sequence can be represented by two selected keyframes before and after the zoom [4]; a panoramic image can be constructed to described panning sequence [2]. Besides keyframe extraction, the temporal variation among keyframes ought to be modeled to reflect the motion content of shots. Based on these intuitions, recently, Ngo proposed a motion-based video representation scheme [3] to temporally segment and describe the video content in a compact yet effective way for scene change detection. Zhao proposed a breakpoint detection algorithm [5] based on the nearest feature line (NFL) approach [10] to compactly select keyframes and effectively modeling the temporal relationships among those keyframes. Although those works have achieved reasonably good results, there has no work on how to utilizing the video objects (or pseudo-objects) to exploit the spatio-temporal events for video content representation and retrieval.

In this paper, we propose a novel approach to exploiting the spatio-temporal relationship of one shot by analyzing the color and spatial content of video color objects across time. Similar to [4], 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 the shot that has significant content changes, such as panning from indoor to out of window. Unlike [4] that employed 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. We define a "color object" as a color sphere in HSV color space and a color-blob in the frame image. We developed the following descriptors to characterize visual content variations for sub-shot extraction and representation:

- Dominant Color Histogram (DCH) for one "group of frames"(GoF) by fusing temporal information into the frame color histograms, in order to capture the most important colors according to temporal variations;
- Spatial Structure Histograms (SSH) to represent the spatial structural information for one individual frame, to provide complementary functionality to color histograms that lack information about spatial distribution of colors.

As sub-shots being extracted, shot similarity measure can be computed for video shot retrieval based on the similarities between corresponding subshots.

The rest of this paper is organized as follows. In Section 2, we first introduce the Dominant Color Histogram (DCH) and Spatial Structure Histograms (SSH). Then we describe subshot extraction and shot similarity measure based on DCH and SSH descriptors. Experimental results are given in Section 3 and the conclusion remarks are in Section 4.

2. THE PROPOSED APPROACH

2.1 Dominant Color Histogram

Color histogram is popularly used in content-based image retrieval (CBIR) for its simplicity and effectiveness. So it is natural to extend the idea of color histogram to CBVR, such as color histogram of key-frames [6]. In [7] Ferman introduced a set of color histograms called alpha-trimmed average histograms for one GoF, including average histogram and median histogram. Here we will present dominant color histogram for one GoF by

dominant color extraction and tracking. In [8], dominant color histogram for one image not only reduces the number of histogram bins but also enhances the performance of histogram matching, because it tends to capture the majority of color content and be robust to noise introduced by small bins. In this paper, dominant color histogram for one GoF depends not only on dominant colors of individual frames, but also their temporal variations. So this representation meets the nature of video as a temporal media. Dominant color histogram is distinctive from previous work with incorporating temporal information and some semantic considerations. GoF is a general concept that may be shots, subshots, and group of shots. In the following, we will deal shot as GoF by assuming that one shot has one single theme. If not, we can segment the shot into several subshots such that each subshot contains coherent content.

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. The focusing background or foreground objects should have longer duration in one shot. Color is an effective yet computational inexpensive feature used in content-based retrieval. Not all the colors presented in one shot, but the dominant colors of focused objects and background, will prevail the measure of shot similarity, with considering human perception. Dominant colors should be not only dominant in one frame, but also dominant across the entire shot. We want to capture the colors of focused objects in one shot by temporal variations and weigh them according to the temporal duration. Therefore, the shot representation by dominant color histogram is emphasizing the dominant objects or background, which is very different from previous color histograms. As details described in [12], we will give some key points about dominant color histogram.

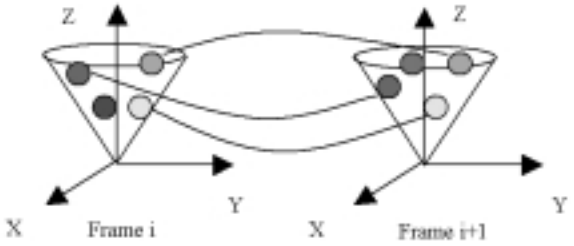

Figure 1: Dominant color extraction and tracking.

Firstly, we calculate the color histogram for each frame, from which dominant colors of the frame are identified. We use HSV color cone [8] that is quantized by a 3D Cartesian coordinate system with 20 values for X and Y, 10 values for Z (the lightness), respectively, as shown in Figure 1. Pixels of each frame, or DC blocks in I frames when MPEG1/2 video are used, are projected into the quantized HSV color space. The normalized distribution of these pixels in the 3D color space thus forms normalized 3D color histograms of the frame. All the dominant local maxima points in the 3D color histogram are identified; and a sphere surrounding each local maximum with a small neighborhood (with diameter of 3 quantization units) in the color space is defined as a color object. These color objects with the largest numbers of pixels (top 20 in our implementation) are identified as dominant color objects, which may (often) not represent a spatial object in a frame.

Then, dominant color objects defined as above in consecutive frames are tracked in the HSV space to identify dominant colors of a shot. If the positions of two dominant colors in two consecutive frames are sufficient close, these two colors are recognized as the same color. Such a color tracking process continues until all frames in the shot are tracked. After tracking, only the colors with longer durations are retained as dominant colors of an entire shot. In other 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 colors that are not only dominant in a frame, but also dominant across the entire shot. To give more weight to colors with longer durations in a shot since they are more dominant in perception, the histogram bins, corresponding to each dominant color are weighted by its relative duration as,

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

where d_0 is the duration of the shot, and d_1 is duration of the dominant color bin (x, y, z). Also, $hist_d^A(x,y,z)$ is normalized by the mean size of each dominant color within the shot. Therefore, the dominant color histogram of a shot represents both structural content in a frame and temporal content in a shot.

2.2 Spatial Structure Histogram

Spatial information would be very important to describe the global and local spatial configuration for one image. In [9], Zhou introduced a new concept called image structural feature, which is a feature in-between texture and shape but more general. The structural features are extracted by Water-Filling algorithm from edge maps [9]. 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

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

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

$$H_{area}(i) = \sum_{Rj \in \Omega i} Area(R_j), i = 0,1,...,7$$
$$\Omega_i = \{R_j \mid Area(R_j) \in [A_i, A_{i+1})\}, i = 0,1,...,7 \quad (3)$$
$$A_i = 1/2^{8-i}, i = 1,2,...,8; A_0 = 0$$

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

$$H_{pos}(i) = \sum_{Rj \in \Omega i} Area(R_j), i = 0,1,2,...,15$$
$$\Omega_i = \{R_j \mid Center(R_j) \in Block(i)\}, i = 0,1,2,...,15 \quad (4)$$

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

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

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

$$H_{sx}(i) = \sum_{Rj \in \Omega i} Area(R_j), i = 0, 1, 2,...,7$$
$$\Omega_i = \{R_j \mid Width(R_j) \in [B_i, B_{i+1})\}, i = 0, 1, 2,...,7 \quad (6)$$
$$B_i = 1/2^{8-i}, i = 1,...,7, 8; \quad B_0 = 0$$

where *Width(R_j)* is the width of minimum bounding rectangle (MBR) 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.

Figure 2: Examples of segmented color-blob maps.

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

$$SshSim(a,b) = W0 \times Sim0 + W1 \times Sim1 + W2 \times Sim2 + W3 \times Sim3 \quad (7)$$

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:

$$P1 > T1, \text{ and } P2 > T2, \text{ and } P1 + P2 > T3 \quad (8)$$

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

$$Sim(a,b) = \max_{i,j}(Sim(a_i, b_j)) \quad (9)$$

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:

$$Sim(a_i, b_j) = Wc \times DchSim(a_i, b_j) + Ws \times SshSim(a_i, b_j) \quad (10)$$

$$Sim(a_i, b_j) = DchSim(a_i, b_j) \times SshSim(a_i, b_j) \quad (11)$$

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

3. EXPERIMENTAL RESULTS

The experiment was conducted on 815 shots from a dozen of TV sports programs, with totally 163880 frames (110 minutes). The video database is very challenging because it contains a diversity of sports programs, including ball games, track and field events, diving, boxing, etc. Figure 3 shows the user interface of our experiments. We chose 8 semantic classes of querying shots as basketball-court, tennis-court, diving, swimming, close-up on head and shoulder, audience, underwater, and stadium, with examples shown in Figure 4. For each class we randomly picked out 20 shots as query examples. Eight retrieval methods were tested for compare:

I. Average Color Histogram (ACH);
II. Median Color Histogram (MCH);
III. Dominant Color Histogram (DCH);
IV. Spatial Structure Histograms (SSH);
V. Weighted sum of DCH and SSH similarities;
VI. Product of DCH and SSH similarities;
VII. Weighted sum of DCH and SSH similarities for subshots;
VIII. Product of DCH and SSH similarities for subshots.

We adopted the average normalized modified retrieval rank (ANMRR) in MPEG-7 standardization activities [11][7] and the average recall (AR) as performance measures. A low value of ANMRR means the relevant shots ranked at the top, and a high value of AR denotes more relevant shots found in the top K (K is the cut-off number for retrieval). Experimental results on AR and ANMRR for different shot classes (from 0 to 7) with different methods (from I to VIII) are shown in Table 1 and 2. The overall performances are on par, except that of SSH because

sports video has a wide range of spatial structure variations than color components. Within color histogram methods I, II, and III, the results of DCH is slightly better than ACH and MCH with considering that DCH achieves the best AR but almost the same ANMRR with ACH and MCH. The performance difference between DCH and the other two is supposed to be greater for the object-tracking shots with significant content variations. For methods V, VI, VII and VIII, the equally weighted sum of DCH and SSH similarities is inferior to the product versions, showing that weighted sum is not a good fusing approach. Overall, method VIII outperforms all the others to show that subshot extraction and representation by DCH and SSH could achieve the best performance on ANMRR and AR simultaneously.

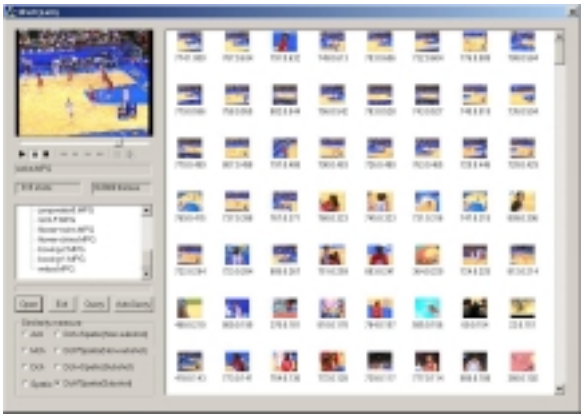

Figure 3: Interface for the shot retrieval experiments.

Figure 4: Shot examples of the semantic classes.

4. CONCLUSIONS

In this paper we present a novel scheme on shot content representation and similarity measure by subshots extraction and representation. Two content descriptors are developed to measure video content variations and represent subshots: dominant color histogram (DCH) and spatial structure histograms (SSH). Eight similarity measures are tested and our proposed method (subshot representation by DCH and SSH) gives the best performance on ANMRR and AR. Our future work is to develop new motion descriptors that could be incorporated in the subshot representation scheme with DCH and SSH to improve the performance of shot retrieval applications.

5. REFERENCES

[1] Y. Rui and T. S. Huang, A Uniform Framework for Video Browsing and Retrieval, *The Image and Video Processing Handbook*, Alan Bovik, ed., Academic Press, 2000.

[2] M. Irani and P. Anandan, "Video indexing based on mosaic representations," *Proceedings of IEEE*, vol. 86, pp. 905-921, May 1998.

[3] C. W. Ngo, T. C. Pong, H. J. Zhang, and R. T. Chin, "Motion characterization by temporal slices analysis", *Proceedings of CVPR2000*, vol. 2, pp. 768-773, 2000.

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

[5] L. Zhao, et al, "Key-frame extraction and shot retrieval using Nearest Feature Line (NFL)", *International Workshop on Multimedia Information Retrieval*, in conjunction with *ACM Multimedia2000*, Los Angeles, USA, 2000.

[6] H. J. Zhang et al, "An integrated system for content-based video retrieval and browsing", *Pattern Recognition*, May 1997.

[7] A. M. Ferman, et al, "Group-of-frames/pictures color histogram descriptors for multimedia applications", *ICIP2000*.

[8] W. Y. Ma and H. J. Zhang, "Content-based image indexing and retrieval", in *Handbook of Multimedia Computing*, Borko Furht, ed. CRC Press, 1998.

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

[10] S. Z. Li, K. L. Chan, and C. L. Wang, "Performance evaluation of the Nearest Feature Line method in image classification and retrieval", *IEEE Trans. on PAMI*, Vol. 22, No. 11, Nov. 2000.

[11] MPEG Video Group, Description of core experiments for MPEG-7 color/texture descriptors, *ISO/MPEGJTC1/SC29/WG11 MPEG98/M2819*, July 1999.

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

	I	II	III	IV	V	VI	VII	VIII
0	0.3951	0.3996	0.3916	0.3637	0.3925	0.4018	0.3850	0.3965
1	0.4417	0.4514	0.4556	0.2694	0.4722	0.4889	0.4708	0.4792
2	0.4967	0.4967	0.5025	0.5475	0.5558	0.5125	0.5667	0.5150
3	0.8485	0.8515	0.8758	0.4727	0.7742	0.8894	0.7682	0.8939
4	0.4737	0.5158	0.5000	0.2553	0.4053	0.5000	0.4211	0.4947
5	0.8227	0.7818	0.7818	0.6409	0.7818	0.7909	0.8318	0.8636
6	0.7387	0.7403	0.8016	0.3177	0.7145	0.7919	0.7129	0.7823
7	1.0000	1.0000	1.0000	0.7000	1.0000	1.0000	1.0000	1.0000
AR	*0.7453*	*0.7482*	*0.7584*	*0.5096*	*0.7281*	*0.7679*	*0.7366*	*0.7750*

Table 1: AR for different shot classes (from 0 to 7) with different methods (from I to VIII).

	I	II	III	IV	V	VI	VII	VIII
0	0.6919	0.7046	0.7093	0.7405	0.7066	0.7031	0.7097	0.7075
1	0.6925	0.6780	0.6930	0.8145	0.6882	0.6810	0.6843	0.6801
2	0.4691	0.4724	0.4690	0.5687	0.4558	0.4711	0.4502	0.4696
3	0.3023	0.2970	0.2932	0.6853	0.3477	0.2629	0.3274	0.2516
4	0.6276	0.5903	0.6163	0.8230	0.6699	0.6214	0.6620	0.6212
5	0.2526	0.2514	0.2706	0.4552	0.2521	0.2539	0.2650	0.2620
6	0.3361	0.3401	0.2945	0.7630	0.3590	0.3000	0.3517	0.3029
7	0.0142	0.0108	0.0128	0.4357	0.0057	0.0037	0.0060	0.0045
ANMRR	*0.4838*	*0.4778*	*0.4800*	*0.7552*	*0.4978*	*0.4710*	*0.4937*	*0.4714*

Table 2: ANMRR for different shot classes (from 0 to 7) with different methods (from I to VIII).