Shot-based Retrieval by Integrating Color and Motion Features

Yuxin Peng[1] and Chong-Wah Ngo[2]

[1]Institute of Computer Science and Technology, Peking University
pengyuxin@icst.pku.edu.cn

[2]Department of Computer Science, City University of Hong Kong
cwngo@cs.cityu.edu.hk

Abstract

In this paper, a novel approach is proposed for shot-based retrieval by integrating color and motion features in video. To effectively measure shot similarity, the color and motion similarity between two shots are jointly measured. In color similarity measure, a weighted bipartite graph is constructed to model the similarity between two shots. Then optimal matching based on Kuhn-Munkres algorithm is employed to compute the maximum weight of a constructed bipartite graph as the similarity value between two shots by guaranteeing the one-to-one mapping among keyframes. In motion similarity measure, a motion histogram is constructed to represent every shot, the motion similarity between two shots is measured by the intersection between their motion histograms. The final similarity is based on the weighted sum of color and motion similarity value. Experimental results indicate that the proposed approach achieves superior performance than some existing methods in sport videos with 3,392 shots.

Keywords: Shot-based Retrieval, Color and Motion Similarity; Optimal Matching.

1. Introduction

Due to the drastic advances in multimedia and internet applications, the effective techniques for video retrieval are increasingly demanded. One critical component in these techniques is the similarity measure of video information. Shot-based retrieval, as the basis for video retrieval, clustering and summarization, remains a challenging problem. In this paper, a novel approach is proposed for shot-based retrieval by integrating color and motion features in videos.

In shot-based retrieval, representative works include [1-6]. In [1][2], one keyframe is extracted to represent the content of a shot using unsupervised clustering method. In [3], nearest feature line (NFL) is employed to extract the keyframes. After keyframes are extracted, shot-based similarity measure is equivalent to image-based similarity measure. As a result, shot-based retrieval can be tackled in a similar way as image retrieval. However, in addition to image information, video also contains spatio-temporal and motion information. The approaches in [1][2][3] did not exploit the special information existing in videos. In [4][5], subshot is proposed for shot-based similarity measure. A shot with significant content changes is represented by several coherent subshots. The shot similarity is measured based on their corresponding subshots. In [4], subshots are segmented based on its motion content, and keyframes are extracted and constructed to represent the different motion subshot. For example, a static subshot is represented by one frame, a pan subshot is represented by constructing a panoramic image, and a zoom subshot is represented by two selected keyframes before and after zoom. The shot similarity is equal to the average of maximum similarity and the second largest similarity value in all pair of keyframes. In [5], dominant color histograms (DCH) and spatial structure histograms (SSH) are proposed to extract and represent subshot, the similarity between two shots is equal to the maximum similarity of their subshots. The methods in [4][5] exploit the motion and spatio-temporal information existing in shots, however, the methods using the maximum and the second largest similarity value can not fully and objectively measure the shot similarity. The method in [6] assumes the frames in two shots are similar in temporal order, dynamic programming is employed to measure the shot similarity. But the assumption is not always right. Besides, the retrieval speed is slow because the similarity measure is based on every pairs of frames between two shots.

In addition, the existing methods [1-6] only employ color feature for shot similarity measure. Motion feature, however, is not yet exploited for effective similarity measure between two shots. In fact, the same classes of videos, such as swimming, diving, football, basketball, volleyball, etc. in sport videos, have the nearly same motion patterns. For example, the turn motion from up to down is the common features in diving classes. The common motion

features can be fully employed for shot-based similarity measure. In this paper, a novel approach is proposed for shot-based retrieval. The similarity measure between two shots is the focus in this paper by integrating color and motion features in videos. The major contributions of the proposed approach are as follows:

- Color similarity measure. A graph matching algorithm, namely optimal matching (OM), is adopted for color similarity measure between two shots. A weighted bipartite graph is constructed to model the similarity between two shots: every vertex in a bipartite graph represents one keyframe in a shot, and the weight of an edge represents the color similarity for a pair of keyframes between two shots. Then OM based on Kuhn-Munkres algorithm is employed to compute the maximum weight of a constructed bipartite graph as the similarity value between two shots by guaranteeing the one-to-one mapping among keyframes.
- Motion similarity measure. A motion histogram is constructed to represent a shot, the motion similarity is measured by the intersection between two motion histogram. The final similarity is based on the weighted sum of color and motion similarity value.

2. Color Similarity Measure

In color similarity measure, the weighted bipartite graph of two shots is constructed as follows:

- Let $X = \{x_1, x_2, \mathsf{K}, x_p\}$ as a query shot with p frames, and x_i represents a frame in X.
- Let $Y_k = \{y_1, y_2, \mathsf{K}, y_q\}$ as a shot in the video database with q frames, and y_j is a frame in Y_k.
- Let $G_k = \{X, Y_k, E_k\}$ as a weighted bipartite graph, where $V_k = X \cup Y_k$ is the vertex set, $E_k = \{\omega_{ij}\}$ is the edge set, and ω_{ij} represents the color similarity value between x_i and y_j.

In the proposed approach, the color similarity value ω_{ij} is computed as the follows:

$$\omega_{ij} = \frac{1}{A(x_i, y_j)} \sum_h \sum_s \sum_v \min\{H_i(h,s,v), H_j(h,s,v)\}$$

$$A(x_i, y_j) = \min\left\{ \sum_h \sum_s \sum_v H_i(h,s,v), \sum_h \sum_s \sum_v H_j(h,s,v) \right\}$$

The histogram is in HSV color space. Hue is quantized into 18 bins while saturation and intensity are quantized into 3 bins respectively. The quantization provides $162(18\times3\times3)$ distinct color sets. After $G_k = \{X, Y_k, E_k\}$ is constructed, OM based on Kuhn-Munkres algorithm [7] is employed to measure similarity between X and Y_k, the algorithm is given in Figure 1.

1. Start with the initial label of $l(x_i) = \max_j(\omega_{ij})$ and $l(y_j) = 0$, where $i, j = 1,2,...,t$ and $t = \max(p,q)$.
2. Compute $E_l = \{(x_i, y_j) \mid l(x_i) + l(y_j) = \omega_{ij}\}$, $G_l = (X, Y_k, E_l)$ and one matching M in G_l.
3. If M contains all the vertices in X, M is the optimal matching of G_k and the algorithm ends. Otherwise, goto step 4.
4. Find a vertex $x_i \in X$ and x_i is not inside M. Set $A \leftarrow \{x_i\}$ and $B \leftarrow \phi$, where A and B are two different sets.
5. Let $N_{G_l}(A) \subseteq Y_k$ as the set of vertices that matches the vertices in set A. If $N_{G_l}(A) = B$, then goto step 9, otherwise goto step 6.
6. Find a vertex $y_j \in N_{G_l}(A) - B$.
7. If $(z, y_j) \in M$, set $A \leftarrow A \cup \{z\}, B \leftarrow B \cup \{y_j\}$ and goto step 5. Otherwise goto step 8.
8. There exists an augmenting path P from x_i to y_j. Set $M \leftarrow M \oplus E(P)$ and goto step 3.
9. Compute $a = \min_{\substack{x_i \in A \\ y_j \notin N_{G_l}(A)}} \{l(x_i) + l(y_j) - \omega_{ij}\}$, then construct a new label $l'(v)$ by

$$l'(v) = \begin{cases} l(v) - a & v \in A \\ l(v) + a & v \in B \\ l(v) & otherwise \end{cases}$$

Compute $E_{l'}, G_{l'}$ based on l'.

10. Set $l \leftarrow l', G_l \leftarrow G_{l'}$, goto step 6.

Fig. 1: Kuhn-Munkres Algorithm for OM.

The computational complexity of Kuhn-Munkres algorithm is $O(n^4)$, where $n = p + q$, is the number of vertex of G_k. The color similarity $Similarity_{color}(X, Y_k)$ between shot X and shot Y_k is defined as follows:

$$Similarity_{color}(X, Y_k) = \frac{\omega_{OM}(X, Y_k)}{\min(p,q)}$$

where $\omega_{OM}(X, Y_k)$ is the total weight after OM.

The above approach can measure effectively the shot similarity. However, some shots often have thousands of frames, it is time consuming for Kuhn-Munkres algorithm to compute a bipartite graph composed of hundreds of vertices. Considering the content redundancy in a shot, for example, a static shot may include thousands of frames, however, one frame

can be selected in the shot to represent the shot content. To speed up matching time, the two improved approaches are proposed as follows:

(1) *Bipartite graph construction based on subshots.*

A shot with significant content changes is represented by several coherent subshots. The method in [4] is utilized to segment a shot into several coherent subshots based on its motion content. Then keyframes are extracted and constructed to represent the different motion subshot. The detail is presented in table 1:

Subshot	Keyframe
Static	Select one frame
Pan or tilt	Form a new panoramic image
Zoom	Select first and last frames
Multiple motion	Reconstruct background
Indeterministic	Select one frame

Table 1: Subshot selection and construction [4].

According to the method in table 1, the matching time of Kuhn-Munkres algorithm can be speed up greatly. However, some shots only have one subshot and one keyframe, for example, one keyframe in the static shot, and one panoramic keyframe in the pan shot. In this situation, Kuhn-Munkres algorithm is equal to compute the maximum similarity value with the one keyframe. To solve this problem, the following approach is employed.

(2) *Bipartite graph construction based on the same number of keyframes in shots.*

The complete bipartite graph can be then constructed based on the same number of keyframe. In this situation, the problem in method(1) can be solved, the shot similarity can be efficiently measured by guaranteeing the one-to-one mapping among keyframes. Although the method(2) may exist keyframe redundant representation, this will not affect the final similarity measure.

3. Motion Similarity Measure

Motion histogram for shot is constructed based on motion vector field (MVF). For a given frame in a video, MVFs are extracted between the current and the next frame and motion characteristics are calculated. In the experiment, the sport video is stored as MPEG format, so MVFs are extracted from MPEG video directly.

In the proposed approach, motion histogram for shot is calculated based on two inductors: *angle inductor* and *intensity inductor*. The *angle inductor* induces the direction of motion vector, while *intensity inductor* induces motion energy or activity. They are calculated as follows:

$$angle(i,j) = arctg(\frac{dy_{i,j}}{dx_{i,j}})$$

$$\text{int}\, ensity(i,j) = \sqrt{dx_{i,j}^2 + dy_{i,j}^2}$$

where $(dx_{i,j}, dy_{i,j})$ denote two components of motion vector. The angle in 2π is quantized into n angle ranges. Then intensity in each angle range is accumulated over a shot to form a motion histogram with n bins, denoted by $H_i(angle)$, i is the shot, $angle \in [1, n]$. In this implementation, n is set to 8. In addition, only the MVFs in P-frame are considered in order to reduce computation complexity. Finally, the motion similarity between two shots X and Y_k is defined as

$$Similarity_{motion}(X, Y_k) = \frac{1}{A(H_X, H_{Y_k})} \sum_{angle} \min\{H_X(angle), H_{Y_k}(angle)\}$$

$$A(H_X, H_{Y_k}) = \max\left\{\sum_{angle} H_X(angle), \sum_{angle} H_{Y_k}(angle)\right\}$$

4. Experimental Results

The experimental database is composed of 3 hours sport videos with 3,392 shots. These videos are recorded from sport games. They include many different sport games and commercials. In the experiment, six classes of sport games are selected as the query shots, which are swimming, judo, volleyball, football, fence-play and field hockey, as shown in Figure 2. For each class, 10 shots are randomly picked as query examples. Five methods are tested for comparison purpose:

Ⅰ Weighted sum of color similarity with three keyframes and motion similarity. In this experiment, the weight of color is 0.7, and the weight of motion is 0.3.

Ⅱ Color similarity measure with three keyframes.

Ⅲ Color similarity measure with subshot representation.

Ⅳ Motion-based shot representation and similarity measure [4].

Ⅴ Color similarity measure with one keyframe.

In methods Ⅰ and Ⅱ, the first, middle and last frame in every shot are extracted as keyframes to construct the complete bipartite graph. In method Ⅴ, the first frame in every shot is extracted as the keyframe. In addition, all the five methods employ 162 bins in HSV color space to represent the color features, and the color similarity is measured by the histogram intersection, so the experimental results can prove the effectiveness of the proposed approach.

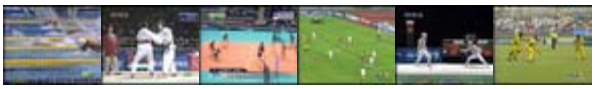

Fig. 2: Query examples of six semantic classes.

	Ⅰ	Ⅱ	Ⅲ	Ⅳ	Ⅴ
1	0.7551	0.7202	0.6191	0.6247	0.6663
2	0.5575	0.5263	0.5650	0.5650	0.5250
3	0.6112	0.5895	0.6473	0.6502	0.5725
4	0.7167	0.6725	0.6334	0.6206	0.6767
5	0.8918	0.8633	0.8020	0.7898	0.7408
6	0.6985	0.6940	0.6761	0.6746	0.6313
AR	0.7051	0.6776	0.6572	0.6542	0.6354

Table 2: AR for different shot classes (from 1 to 6) with different methods (from Ⅰ to Ⅴ).

	Ⅰ	Ⅱ	Ⅲ	Ⅳ	Ⅴ
1	0.3866	0.3842	0.4926	0.4876	0.4466
2	0.5271	0.5264	0.5113	0.5073	0.5476
3	0.4906	0.4958	0.4384	0.4363	0.5092
4	0.4028	0.4246	0.4710	0.4755	0.4183
5	0.2298	0.2343	0.2878	0.2934	0.3657
6	0.4094	0.3899	0.4019	0.4107	0.4571
ANMRR	0.4077	0.4092	0.4338	0.4351	0.4574

Table 3: ANMRR for different shot classes (from 1 to 6) with different methods (from Ⅰ to Ⅴ).

AR (average recall) and ANMRR (average normalized modified retrieval rank) are adopted for performance evaluation [8]. The values of AR and ANMRR range from [0, 1]. A *high* value of AR denotes the superior ability in retrieving relevant shots, while a *low* value of ANMRR indicates the high retrieval rate with relevant shots ranked at the top.

Experimental results on AR and ANMRR for six semantic classes with five methods are shown in Tables 2 and 3. In overall, the three methods using OM (methods Ⅰ, Ⅱ and Ⅲ) outperform the two existing methods (methods Ⅳ and Ⅴ) in term of AR and ANMRR. The main reasons are: OM provides a good mechanism for similarity measure and ranking through one-to-one mapping among keyframes. Comparing with the two proposed methods using color features (methods Ⅱ and Ⅲ). In method Ⅲ, some shots only include one subshot and one keyframe based on camera motion, then OM is employed to only compute the maximum similarity with one keyframe. While in method Ⅱ, three keyframes are extracted in every shot, then a complete bipartite graph is constructed, shot similarity can be efficiently measured by OM, the problem in method Ⅲ can be solved in method Ⅱ, so the method Ⅱ outperforms the method Ⅲ in term of AR and ANMRR. In addition, method Ⅰ achieves best AR and ANMRR in the five methods. Comparing with methods Ⅰ and Ⅱ, method Ⅰ adds the motion features based on method Ⅱ. The result indicates motion features is useful for shot-based similarity measure, especially in sport video.

5. Conclusions

In this paper, a novel approach has been proposed for shot-based similarity measure by integrating color and motion features in videos. In color similarity measure, optimal matching is employed to compute the maximum weight of a constructed bipartite graph as the similarity value between two shots. In motion similarity measure, a motion histogram is constructed to represent every shot, the motion similarity is measured by the intersection between two motion histogram. Experimental results have indicated that the proposed approach achieves superior AR and ANMRR than some existing methods.

Currently, the implementation of OM is based on Kuhn-Munkres algorithm which requires $O(n^4)$, where n is the number of keyframes. Faster versions of OM algorithms exist [9], for instance, OM can run in $O(n(m + n\log n))$, where m is the number of matching edges. In future, faster algorithm will be incorporated in the proposed approach for more efficient retrieval.

6. References

[1] X. Liu, Y. Zhuang , and Y. Pan, "A New Approach to Retrieve Video by Example Video Clip," *ACM Multimedia Conf.*, 1999.

[2] Y. Wu, Y. Zhuang, and Y. Pan, "Content-based Video Similarity Model," *ACM Multimedia.*, 2000.

[3] L. Zhao, W. Qi, S. Z. Li, *et al*, "Key-Frame Extraction and Shot Retrieval Using Nearest Feature Line (NFL), *ACM SIGMM International Workshop on Multimedia Information Retrieval (MIR00)*, 2000.

[4] C. W. Ngo, T. C. Pong, and H. J. Zhang, "Motion-based Video Representation for Scene Change Detection," *Int. Journal of Computer Vision*, 50(2), pp. 127-143, Nov 2002.

[5] T. Lin, C. W. Ngo, H. J. Zhang, *et al*, "Integrating Color and Spatial Features for Content-based Video Retrieval", *IEEE International Conference on Image Processing (ICIP 2001)*, pp. 592-595, 2001.

[6] L. Chen, and T. S. Chua, "A Match and Tiling Approach to Content-based Video Retrieval," *Int. Conf. on Multimedia and Expo*, 2001.

[7] W. S. Xiao, *Graph Theory and Its algorithms*, Beijing: Aviation Industrial Press, 1993.

[8] MPEG video group, "Description of Core Experiments for MPEG-7 Color/Texture Descriptions", *ISO/MPEGJTC1/SC29/WG11 MPEG98/M2819*, 1999.

[9] A. Schrijver, *Combinatorial Optimization: Polyhedra and Efficiency*, Berlin: Springer, Vol. A, 2003.