MEAN SHIFT BASED VIDEO SEGMENT REPRESENTATION AND APPLICATIONS TO REPLAY DETECTION

Ling-Yu Duan, Min Xu, Qi Tian, Chang-Sheng Xu

Institute for Infocomm Research, 21 Heng Mui Keng Terrace, Singapore 119613
{lingyu, xumin, tian, xucs}@i2r.a-star.edu.sg

ABSTRACT

Effective and efficient representation of the low-level features of groups of frames or shots is an important yet challenging task for video analysis and retrieval. Key frame-based representation is limited by the difficulties in shot boundary detection of gradual transition and a variety of ways in key frame extraction. In this paper, we employ the mean shift-based mode seeking function to develop a new approach for compact representation of the video segment. The proposed video representation is motivated by recognizing that, on the global level, humans perceive images only as a combination of few most prominent colors. We exploit the spatiotemporal mode seeking in feature space to simulate "subjectivity" of human decisions to video segment retrieval and identification. The effectiveness of video representation and matching scheme is shown by initial experiments on replay detection in broadcast sports video.

1. INTRODUCTION

Content-based image and video browsing and retrieving has emerged as a challenging area in multimedia and computer vision. Various systems have been proposed such as QBIC [1], Photobook [2], and Virage [3]. These systems represent images via a set of low-level feature attributes (e.g. color, texture, shape, etc.). A retrieval is performed by matching the feature attributes of the query with those of the database images. Unlike an image, a video sequence consists of large amounts of frames. It becomes necessary to represent groups of frames or shots in a video with effective and efficient approaches. It is customary to describe the visual and color content of shots using key frames and key frame histograms, respectively [4]. While some researchers employ color-, motion-, and/or object-based criteria for appropriate key frame selection [5, 6], the underlying ambiguities in shot boundary detection and key frame extraction cause the variations and arbitrariness in the key-frame based representations.

Many studies have discovered that, when viewing the global color content, human visual system eliminates fine details and average colors within small areas [7]. Hence, on the global level, humans perceive images only as a combination of few most prominent colors. These findings motivate us to address the issue of video representation from the viewpoint of dominant features (e.g. color, motion, texture, etc.) in the context of groups of frames. There is a close relationship between dominant features detection and mode seeking in the feature space. In order to avoid the variations in the key-frame based representations, a more favorable approach is to consider the features of all the frames for seeking dominant modes to represent video content.

The histogram is the simplest and the most often used feature representations. It is often employed in combination with the Euclidean distance as a measure of dissimilarity, providing undemanding yet efficient retrieval method [8, 9]. However, the histogram representation does not match human perception very well and lacks discriminatory power in retrieval of large image and video databases. This weakness together with inefficient use of the data makes it necessary to use alternatives to histograms.

In this paper, we present a compact representation of the video content by employing the mode seeking function of mean shift procedure [10]. The comparison of two groups of frames is transformed into the distance measure between two sets of spatio-temporal feature modes. One of the main advantages of the proposed representation is that, unlike key frame histograms, it is uniquely defined by the kernel bandwidth parameter for a given set of frames, and thus provides a consistent standard feature set. Another advantage is the appropriate exploitation of spatio-temporal context information, which is taken into account by the multivariate kernel in the joint domain. Once the feature modes in a video sequence are identified, it is straightforward to employ the *Earth Mover's Distance* (EMD) [11] or the *Optimal Color Composition Distance* (OCCD) [7] to measure the video clip dissimilarity according to the resulting feature modes. We employ the proposed representation and matching scheme to detect replay scene in broadcast sports video.

This paper is organized as follows. In Section 2, we present a brief review of the mean shift procedure. In Section 3, we present the compact video representation scheme. To initially evaluate the effectiveness of this scheme, we propose the application of replay detection in broadcast sports video in Section 4. Finally, we conclude this paper in Section 5.

2. MEAN SHIFT PROCEDURE

Bayes' theorem is the basis of the statistical approach to pattern recognition. Its importance lies in the fact that it re-expresses the posterior probabilities in terms of quantities, i.e. the prior probability and the class-conditional probability, which are often much easier to calculate. For modeling the class-conditional probability, we have to consider the problem of estimating a probability density function $p(x)$, given a finite number of data points $X_n, n = 1,\ldots,N$ drawn from that density function.

The histogram is the oldest and most widely used density estimator. The histogram representation is to estimate the class-

conditional probability with the histogram method. However, the discontinuity of histograms requires the selection of the boundaries of the bins in advance of observing the data. It is unlikely that they represent true structure in the distribution. Another serious problem is the huge number of data points to obtain a density estimate in high dimensions. Most of the bins would be empty, and thus it leads to inefficient use of the data.

Apart from the histogram, the kernel estimator is quite elegant and of wide applicability. The mean shift procedure is derived by the density gradient estimation. Note that the discontinuity of histograms causes extreme difficulty if derivates of the estimates are required.

Given n data points $X_i, i = 1,\ldots,n$ in the d-dimensional space R^d, the multivariate kernel density estimator with kernel $K(x)$ and window width h is given by

$$\hat{f}(x) = \frac{1}{nh^d}\sum_{i=1}^{n} K\left\{\frac{1}{n}(x - X_i)\right\} \tag{1}$$

For the Epanechnikov kernel, the density gradient estimate becomes

$$\hat{\nabla} f(x) = \frac{n_x}{n(h^d c_d)}\frac{d+2}{h^2}\left(\frac{1}{n_x}\sum_{X_i \in S_h(x)}[X_i - x]\right) \tag{2}$$

where the region $S_h(x)$ is a hypersphere of radius h having the volume $h^d c_d$, centered on x, and containing n_x data points. The last term

$$M_h(x) \equiv \frac{1}{n_x}\sum_{X_i \in S_h(x)} X_i - x \tag{3}$$

is called the sample mean at $x \in X$. The repeated movement of data points to the sample means is called the *mean shift procedure* [12]. The mean shift vector always points towards the direction of the maximum increase in the density. In [12], Cheng has shown that the mean shift is a mode-seeking process on a surface constructed with a "shadow" kernel and studied the convergence for mean shift iteration. Since efficient mean shift computation requires efficient range searching, Comaniciu et al. [10] proposed a computational module of the mean shift procedure, and successfully applied it to two low-level vision tasks: discontinuity preserving filtering and image segmentation.

3. VIDEO REPRESENTATION SCHEME

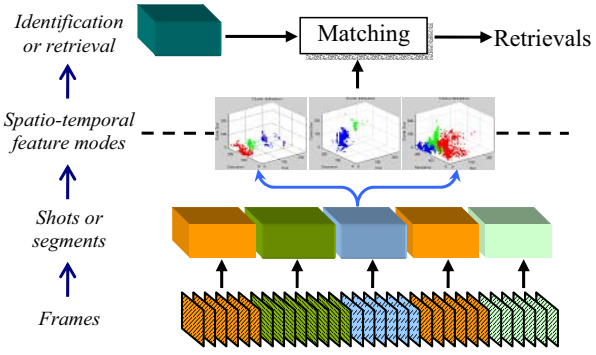

Figure 1. Mean-shift based video representation scheme.

Figure 1 illustrates the proposed video representation scheme.

As shown in Figure 1, this scheme includes three main stages: video segmentation, spatio-temporal feature mode seeking, and matching. The proposed scheme produces unique features in three aspects:

- A generalized approach to abstract groups of video frames using dominant feature modes is proposed. Unlike key frame histograms, it is quite compact, and avoids the weakness (as discussed in Section 2) inherent to the histogram methods.
- The spatio-temporal mode seeking using the multivariate kernel in the joint spatial-range domain elegantly takes into account the video context information.
- The dominant modes-based representation transforms the comparison between two groups of frames to the distance measure between sets of elements. The EMD and OCCD metrics were employed to solve the optimal mapping problem.

We next describe these stages.

3.1 Video Segmentation

A generic mechanism for segmenting video into groups of frames is through the popular shot-based representation model. Various shot-boundary detection algorithms were proposed [13]. Once the shot boundaries are identified, we can perform the mode seeking within a shot followed by the matching at the shot level as shown in Figure 1.

However, robust algorithms for detecting various types of shot boundaries have not been found yet [13]. An alternative is to perform the uniform segmenting of video. The length of a segment can be controlled by the user. This is similar in spirit to the uniform video sub-sampling for key frames extraction in [8]. The statistical modeling may be used to represent the context of a set of consecutive segments [14]. The mode seeking can reduce the computational complexity in statistical learning.

3.2 Spatio-temporal Feature Mode Seeking

We restrict our analysis to visual cues, i.e. color, motion, texture. The goal of this stage is to obtain representative modes by the spatial-temporal feature clustering. After clustering, only a small number of dominant features remain. Each representative feature mode and its corresponding percentage form a pair of attributes that describe the prominent characteristics within a video shot or segment. The dominant mode descriptor is defined to be

$$Mode = \{(c_i, p_i),\ i = 1,\ldots,N\} \tag{4}$$

where N is the total number of dominant modes, c_i is a feature vector, p_i is its percentage, and $\sum_i P_i = 1$. N can vary from shot to shot. We can call the descriptor *Mode* as *video signature*.

In [15], we introduced a cone-shaped motion vector space (MVS) to represent motion vector fields (MVF). The MVS space provides a visualized representation of the MVF, and transforms the integrated analysis of motion and texture cues to the problem of feature space analysis. Thus, a video frame is typically represented as a two-dimensional lattice of p-dimensional

vectors. The space of the lattice is known as the *spatial* domain, while the color, motion, or texture information is represented in the *range* domain.

3.2.1. Spatial feature clustering

We employ the joint domain kernel

$$K_{h_s,h_r} = \frac{C}{h_s^2 h_r^3} k\left(\left\|\frac{\mathrm{x}^s}{h_s}\right\|^2\right) k\left(\left\|\frac{\mathrm{x}^r}{h_r}\right\|^2\right) \quad (5)$$

to perform mean shift clustering of color pixels and motion vectors within each image frame, where X^s and X^r are the spatial part and range part respectively, h_s and h_r are the kernel bandwidths, C is the normalization constant. According to the clustering results, each image frame F_j can be represented by

$$F_j = \{A_i\}_{i=1,..,m},\ A_i =< o_i, \bar{c}_i, r_i > \quad (6)$$

, where o_i denotes the pixels or motion vectors associated with the cluster A_i, $\bar{c}_i$ is the average color or motion vector of o_i, r_i is the normalized cluster size of A_i, $\sum_{i=1}^{m} r_i = 1,\ 0 \le r_i \le 1$.

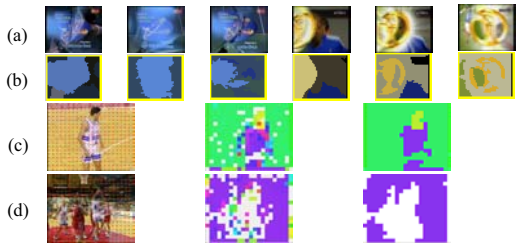

Figure 2. Examples of spatial feature clustering. (a) Frames with intruding graphics at the beginning and end of a replay scene; (b) frames after color-based spatial clustering; (c) (d) motion-based spatial clustering in the MVS space [15] (From left to right: frames overlapped with motion vectors, MVS representation, and the resulting MVF after spatial clustering).

3.2.2. Temporal feature clustering

We employ the joint domain kernel in Equation (5) to perform temporal mean shift clustering of spatial modes

$$S = \{V_{ij}\}_{j=1,\ldots,k;i=1,\ldots,m(j)}, V_{ij} =< \bar{c}_{ij}, r_{ij} > \quad (7)$$

obtained from the spatial feature clustering on a series of image frames $\{F_j\}_{j=1,\ldots,k}$, where $m(j)$ denotes the cluster number in F_j, V_{ij} denotes the spatial cluster i in the frame j. Different from the spatial clustering, the range part is the mode feature and the spatial part is the mode percentage. Their different nature has to be compensated by proper normalization.

Suppose the results of temporal feature clustering contain M clusters $\{C_\ell | \ell = 1,\ldots,M\}$, cluster C_ℓ contains N_ℓ feature points. We have $\sum_{\ell=1}^{M} N_\ell = N$, N is the total number of feature points. $N = \sum_{j=1,\ldots,k} m(j)$. Finally we obtain spatio-temporal modes

$$MODE = \{Mode_\ell\}_{\ell=1,\ldots,M},\ Mode_\ell =< O_\ell, Y_\ell, N_\ell > \quad (8)$$

where O_ℓ are the associated feature points of C_ℓ, Y_ℓ is the cluster centre of C_ℓ, N_ℓ is the number of feature points in C_ℓ. It is easy to represent the dominant mode descriptor in Equation (4) using the spatio-temporal modes $MODE$ in Equation (8).

3.2.3. An example of color characterization

As we have discussed above, the goal of mode seeking is to learn an effective representation of training data itself. The modes are expected to capture the prominent features which achieve the invariance properties from large amounts of data. Various tasks can be based on this mode-based representation, such as semantic shot classification, color characterization, etc.

Here is an example of color characterization in tennis video. We choose a series of continuous image frames as training data. Through the spatio-temporal color mode seeking, we capture the dominant color modes as shown in Figure 3(a) (b). For each incoming frame, we employ the K-nearest-neighbor rule to detect selected color modes. As shown in Figure 3(c), we can determine the court view shot according to the ratio of colors belonging to court color modes. The duration of testing video is about 120 minutes, and 386 court view shots are included. We get the results of Recall 98.7% and Precision 97.2 %.

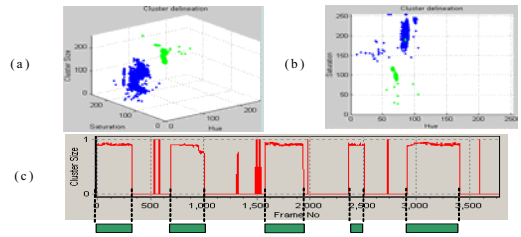

Figure 3. An example of color characterization in tennis video. (a)(b) 3D and 2D representations of dominant color modes corresponding to tennis court color (the training data comprises 5000 continues image frames), (c) the percentage curve of tennis court color, green bars indicating the court view shot boundary.

3.3 Matching

Video similarity matching can be considered at two levels: the shot and sequence levels. The sequence-level matching relies on the shot-level matching and sliding window algorithms such as text tiling [16]. Currently, we focus on the shot-level matching.

We transform the shot-level matching into the distance measure between sets of elements. As video content has been represented using feature modes in Equation (4) and (8), it is straightforward to apply the EMD or OCCD to the matching between two sets of modes. The EMD and OCCD metrics were originally used to measure the difference between two images in terms of color components [7, 11]. They provide a metric for distributions. The EMD uses linear programming to compute the optimal mapping, while the OCCD uses weighted graph matching to compute the optimal mapping. The advantage of the weighted graph matching is its guaranteed algorithm complexity of $O(n^3)$.

The distance measures from multiple visual attributes can be integrated to improve sufficient discriminatory information. A simple approach is taking a weighted average. Different weight assignments may reflect different user requirements.

4. APPLICATION: REPLAY DETECTION

We employ the video representation scheme to identify replay scenes in broadcasted sports video. In general, it might be of great difficulty to classify scenes into either live or replay by means of image analysis. An alternative is to represent and identify the special digital video effects (SDVE) inserted at the beginning and end of a replay scene.

As shown in Figure 2(a), the overlapped 'flying graphics' is a typical kind of SDVE. To obtain robust and distinguishable representation of the SDVE, we choose a set of SDVE video segments as training data, and employ the spatio-temporal mode seeking to capture dominant colors that best describe the overall color appearance. And then we employ a sliding window to perform EMD based similarity matching between the resulting color modes and the segments within a window over the whole video data.

Our proposed replay detection approach has been tested on four matches of soccer video from the 2002 FIFA World Cup. Table 1 gives the performance in terms of precision and recall. The representation of 'flying graphics' are trained by the 54 replay scenes in SEN-FRA match as shown in Figure 4. The duration of each selected SDVE segment is about 0.49s on average. The similarity matching in four matches are all based on the trained SDVE model in Figure 4(b).

Table 1. Performance on replay detection

Match	Total	Correct	False Alarm	Recall	Precision
GER-KOR (25/06/02)	67	65	4	97.0%	94.2%
GER-BRA (30/06/02)	33	30	5	90.9%	85.7%
SEN-TUR (22/06/02)	48	46	4	95.8%	92.0%
SEN-FRA (31/05/02)	54	52	6	96.3%	89.7 %

The accurate detection of replay scenes might provide an efficient way to sports highlights extraction. For example, for broadcast soccer video, we can summarize the replay scenes into three main classes: *Shot*, *Foul*, and *Out of bounds*. These replay shots are inserted within the following *Out of Play* segment [15] in the event of an exciting shot, a serious/disputable/inapparent foul, and out of bounds within the *In Play* segment [15]. Thus, the replay scene classification is to elegantly incorporate production knowledge into highlights extraction. Currently we are working towards soccer highlights extraction by classifying replay scenes with our proposed mid-level representations [15]. Here is some statistics about replay scenes from the four matches listed in Table 1: Shot 47.5%, Foul 34.7%, Out of bounds 9.4%, and Others 8.4%.

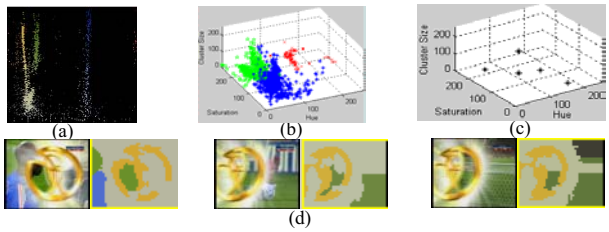

Figure 4. Mode-based representation of flying graphics in the 2002 FIFA World Cup. (a) 2D visualization of feature points from 54 replay scenes; (b) delineation of dominant color modes; (c) cluster centers after spatio-temporal mode seeking; (d) sample images along with spatial clustering results, jointly represented by the 'green' and 'blue' delineations in (b).

5. CONCLUSION

We have proposed a new approach for video representation. Its effectiveness has been initially shown by detecting replay scenes in broadcast sports video. The mean shift-based mode seeking provides a generalized approach to abstract groups of video frames using dominant feature modes. Unlike key frame histograms, it is quite compact, and avoids the weakness inherent to the histogram methods. Moreover, the spatio-temporal mode seeking elegantly incorporates the video context information. We believe that the proposed compact representation can be widely used in content-based video indexing and retrieval, such as shot classification, video trailer identification, and video query, etc. Currently, we are studying the roles of mean shift kernel bandwidth selection in the multi-scale video structure analysis.

6. REFERENCE

[1] M. Flickner et al., "Query by Image and Video Content: The QBIC System," *IEEE Computer* 28(9): 23-32, 1995.

[2] A. Pentland et al., "Photobook: Tools for Content-based Manipulation of Image Databases," In *Proc. of SPIE Storage and Retrieval for Image & Video Databases II*, pp. 34-47, 1994.

[3] A. Hampapur et al., "Virage Video Engine," In *Proc. of SPIE Storage and Retrieval for Image & Video Databases V*, pp. 188-197, 1997.

[4] P. Aigrain, H.J. Zhang, and D. Petkovic, "Content-based Representation and Retrieval of Visual Media: A State-of-the-art Review," *Multimedia Tools and Applications* 3(3): 3-26, 1996.

[5] H.J. Zhang et al., "An Integrated System for Content-based Video Retrieval and Browsing," *Pattern Recognition* 30(4): 643-658, 1997.

[6] C. Kim and J.N. Hwang, "An Integrated Scheme for Object-based Video Abstraction," In *Proc. of ACM Multimedia 2000*, pp. 303-311, 2000.

[7] A. Mojsilovic, J. Hu, and E. Soljanin, "Extraction of Perceptually Important Colors and Similarity Measurement for Image Matching, Retrieval, and Analysis," *IEEE Transactions on Image Processing* 11(11): 1238-1248, 2002.

[8] A.K. Jain, A. Vailaya, and X. Wei, "Query by Video Clip," *Multimedia Systems* 7: 369-384, 1999.

[9] A. M. Ferman, A.M. Tekalp, and R. Mehrotra, "Robust Color Histogram Descriptors for Video Segment Retrieval and Identification," *IEEE Transactions on Image Processing* 11(5): 497-508, 2002.

[10] Y. Cheng, "Mean Shift, Mode Seeking, and Clustering," *IEEE Pattern Analysis and Machine Intelligence* 17(8): 790-799, 1995.

[11] Y. Rubner, C. Tomasi, and L. J. Guibas, "A Metric for Distributions with Applications to Image Databases," In *Proc. of ICCV*, pp. 59-66, 1998.

[12] D. Comaniciu and P. Meer, "Mean Shift: A Robust Approach toward Feature Space Analysis," *IEEE Pattern Analysis and Machine Intelligence* 24(5): 1-18, 2002.

[13] A. Hanjalic, "Shot-Boundary Detection: Unraveled and Resolved," *IEEE Transactions on Circuits and Systems for Video Technology* 12(2): 90-105, 2002.

[14] M. Han, etc., "An Integrated Baseball Digest System Using Maximum Entropy Method," In *Proc. of ACM Multimedia 2002*, pp. 347-350, 2002.

[15] L.-Y. Duan, etc., "A Mid-level Representation Framework for Semantic Sports Video Analysis," In Proc. of *ACM Multimedia 2003*, pp.33-44, 2003.

[16] L. Chen and T.-S. Chua, "A Match and Tiling Approach to Content-based Video Retrieval," In *Proc. of ICME 2001*, pp. 417-420, 2001.