Soccer Video Shot Classification Based on Color Characterization Using Dominant Sets Clustering

Li Li[1], Xiaoqing Zhang[1], Weiming Hu[1], Wanqing Li[2], and Pengfei Zhu[1]

[1] Institute of Automation, Chinese Academy of Sciences, Beijing, China
{lli,wmhu,xqzhang,pfZhu}@nlpr.ia.ac.cn
[2] University of Wollongong, Sydney, Australia
wanqing@uow.edu.au

Abstract. In this paper, we propose a novel approach for dominant color region detection using dominant sets clustering and apply it to soccer video shot classification. Compared with the widely used histogram based dominant color extraction methods which require appropriate thresholds and sufficient training samples, the proposed method can automatically extract dominant color region without any threshold setting. Moreover, the dominant color distribution can be sufficiently characterized by the use of dominant sets clustering which naturally provides a principled measure of a cluster's cohesiveness as well as a measure of a vertex participation to each group. The Earth Mover's Distance (EMD) is employed to measure the similarity between dominant color regions of two frames, which is incorporated into the kernel function of SVM. Experimental results have shown the proposed method is much more effective.

1 Introduction

In this paper, we focus on classifying soccer video shots into long, median and close-up or out of field view, as shown in Fig. 1. The definition of each shot type can be found in [1]. In the same way, we assign close-up and out of field shots into the same category due to their similar semantic meaning.

Soccer is one of the most popular games in the world. The semantic analysis of soccer video has attracted great attention due to its clear semantic information and commercial potentials. Xu *et al.* [2] used web-casting text to combine low-level features with high level semantics for semantic annotation, indexing and retrieval of sports games. Xie *et al.* [3] employed HMM and dynamic programming to detect concepts of "play" and "no play" in soccer video. Wang *et al.* [4] presented an analysis of soccer video for detecting the semantic notion of "offense". Ekin *et al.* [1] proposed a heuristic approach to classify soccer video shots into far, medium and close-up views, and annotate the shot as "in play"or "break". Duan *et al.* [5]introduced a visual feature representation model for sports videos, this model is combined with supervised learning to perform

P. Muneesawang et al. (Eds.): PCM 2009, LNCS 5879, pp. 925–931, 2009.

(a)Long (c)Medium (e)Closeup

(b)Long (d)Medium (f)Out of field

Fig. 1. Examples of predefined shot types

top-down semantic shot classifications. Those semantic shot classes are further used as a mid-level representation for high-level semantic analysis.

Among various features used in shot classification, color is a very useful low-level feature and color histogram is a general and flexible tool. The grass-ratio and non-field area distribution extracted from the color histogram are often used as important features [1][4][6] to classify soccer video shots into long, medium, and close-up view. In this scheme, firstly, the peak index, i_{peak}, which is used to compute the dominant field color for each histogram has to be determined, then, an interval $[i_{min}, i_{max}]$ with $[i_{min} \leq i_{peak} \leq i_{max}]$ about each peak is defined, where i_{max} and i_{min} satisfy the conditions: $H[i_{min}] \leq kH[i_{peak}]$ and $H[i_{max}] < kH[i_{peak}]$, and H refers to the color histogram. Each pixel is measured by the distance to the peak color (i.e., $d_{cylinder}$) by the cylindrical metric. If the pixel satisfies the constraint $d_{cylinder} < T_{color}$, then this pixel will be assigned to the dominant color region, where T_{color} is a pre-defined threshold. In the classification stage, the low grass pixel ratio value T_{close}, indicates that the frame is a close-up view or out-of-field view, while high grass ratio T_{medium} corresponds to a long view, and between, is a medium view. In this method, several thresholds have to be set, such as k which defines the dominant color interval, T_{color} which decides whether a pixel is the field color, T_{close} and T_{medium} in shot classification. The major drawback of this method is the difficulty of determining appropriate thresholds for various soccer video due to the illumination changes. Furthermore, it requires sufficient samples. Different from color histogram, Duan [7] proposed a nonparametric color characterization method based on mean shift procedure to seek modes for representing the colors of semantic importance for content analysis. Nevertheless, the computational cost of mean shift based clustering is relatively high.

In this paper, in order to overcome the influence of thresholds we detect dominant color using dominant sets clustering. Since color histograms are hard to

characterize the color distribution of a homogeneous region, we utilize signature to model the color distribution by using dominant sets clustering [8]. An elegant property of the dominant sets clustering is that it naturally provides a principled measure of a cluster's cohesiveness as well as a measure of a vertex participation to each group. Note that the first dominant set is the biggest one, the others become smaller and smaller, therefore, the first dominant set corresponds to the dominant color in an image, which motivates us to choose the first dominant set as the dominant color. In this way, the dominant color distribution can be characterized without any threshold setting, then we apply the Earth Mover's Distance(EMD) to measure frame similarity and combine it with SVM for shot classification.

The rest of this paper is organized as follows. Section 2 introduces the clustering algorithm based on the concept of dominant set. Section 3 describes the color characterization, the classifier we used is presented in Section 4. Experimental results are presented in Section 5. Section 6 concludes this paper.

2 Dominant Sets Clustering

2.1 Concept of Dominant Set

Dominant set, defined by Pavan *et al.* [8], is a combinatorial concept in graph theory that generalizes the notion of a maximal complete subgraph to edge-weighted graphs. It simultaneously emphasizes on internal homogeneity and external inhomogeneity, and thus is considered as a general definition of "cluster". Pavan *et al.* [8] established an intriguing connection between the dominant set and a quadratic program as follows:

$$\begin{aligned} max \quad & f(\mathbf{x}) = \mathbf{x}^T \mathbf{W} \mathbf{x} \\ s.t. \quad & \mathbf{x} \in \Delta \end{aligned} \tag{1}$$

where

$$\Delta = \{\mathbf{x} \in \mathbb{R}^n : \mathbf{x} \geq 0 \,\text{and} \sum_{i=1}^{n} x_i = 1\}$$

and $\mathbf{W}$ is the similarity matrix. Let $\boldsymbol{u}$ denote a strict local solution of the above program. It has been proved by [8] that $\sigma(\boldsymbol{u}) = \{i | u_i > 0\}$ is equivalent to a dominant set of the graph represented by $\mathbf{W}$. In addition, the local maximum $f(\boldsymbol{u})$ indicates the "cohesiveness" of the corresponding cluster. *Replicator equation* can be used to solve the program (1):

$$x_i(t+1) = x_i(t) \frac{(\mathbf{W}\mathbf{x}(t))_i}{\mathbf{x}(t)^T \mathbf{W} \mathbf{x}(t)} \tag{2}$$

2.2 Dominant Sets Clustering Algorithm

The concept of dominant set provides an effective framework for iterative pairwise clustering. Considering a set of samples, an undirected edge-weighted graph

Table 1. Dominant-set clustering algorithm

Input: the similarity matrix $\mathbf{W}$
1. Initialize $\mathbf{W}^k, k = 1$ with $\mathbf{W}$
2. Calculate the local solution of (1) by (2): $\boldsymbol{u}^k$ *and* $f(\boldsymbol{u}^k)$
3. Get the dominant set: $\boldsymbol{S}^k = \sigma(\boldsymbol{u}^k)$
4. Split out $\boldsymbol{S}^k$ from $\mathbf{W}^k$ and get a new affinity matrix $\mathbf{W}^{k+1}$
5. If $\mathbf{W}^{k+1}$ is not empty, $\mathbf{W}^k = \mathbf{W}^{k+1}$ and $k = k + 1$, then go to step 2; else exit
Output: $\cup_{l=1}^{k}\{\boldsymbol{S}^l, \boldsymbol{u}^l, f(\boldsymbol{u}^l)\}$

with no self-loops is built in which each vertex represents a sample and two vertices are linked by an edge whose weight represents the similarity of the two vertices. To cluster the samples into coherent groups, a dominant set of the weighted graph is iteratively found and then removed from the graph until the graph is empty. Table 1 shows the clustering process. Different from traditional clustering algorithms, the dominant-set clustering automatically determines the number of the clusters and has low computational cost.

3 Color Characterization

In soccer videos, we tend to regard green regions as playing field and brown regions as skin region because the color information implies semantic concept with help of domain knowledge. Color histogram has been widely used to characterize color information. However, the histogram representation does not coincide with human perception very well. Many studies have discovered that, 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 in the context of groups of frames. Clustering is an effective way to learn the structure of multidimensional patterns from a set of unlabeled samples. Among considerable clustering methods, we employ dominant sets clustering because the dominant set is directly mapped to the selection of dominant color. Since viewers usually focus on the middle-image, instead of preforming on the entire image, we partition the entire image into nine regions. The partitioning principle is the Golden Section spatial composition rule[1], which suggests dividing up the screen in 3:5:3 along width and height direction as shown in Fig.2(a). It is observed that, backgrounds such as audiences ,advertising boards, and logos are typically at the top of video frames as shown in Fig. 2(a), hence the most often used regions are R_1, R_2 and R_3. We add another region R_5 based on the fact that a whole human body is usually visible in the medium view, in other words, foot of players is usually visible, while together with foot is the field view. If R_2 is close-up view and R_5 is long view, then this frame is regarded as medium view. Similar to image segmentation, each region to be clustered is represented as an edge-weighted undirected graph $G = (V, E)$, where each pixel corresponds to a node. The similarity between pixels i and j is measured by:

(a) (b)

Fig. 2. Examples of Golden Section spatial composition in (a), (b) The clustering result of selected regions

$$w_{ij} = e^{\frac{-\|\mathbf{F(i)}-\mathbf{F(j)}\|_2^2}{\sigma_I}} * \begin{cases} e^{\frac{-\|\mathbf{X(i)}-\mathbf{X(j)}\|_2^2}{\sigma_X}}, & \text{if } \|\mathbf{X(i)}-\mathbf{X(j)}\|_2 < r \\ 0, & \text{otherwise} \end{cases}$$

where σ_I and σ_X are positive real numbers which affect the decreasing rate of intensity and spatial proximity of similarity, and $\mathbf{F(i)} = [v, vs\sin(h), vs\cos(h)](i)$, where h, s, v are the HSV values of pixel i. As described in Table 1, the dominant sets clustering algorithm begins with the above similarity matrix and iteratively bipartitions the pixels into dominant set and non-dominant set. Hence, this algorithm produces the clusters progressively and hierarchically. The clustering process usually stops when all pixels are grouped into one of the clusters or when certain criteria are satisfied. Fig.2(b) illustrates an example of clustering result of the selected regions.

4 Classifiers

According to clustering results, each region of image frame can be represented by a signature $P = \{(p_i, w_{p_i}), 1 \le i \le m\}$, where p_i denotes the average color of cluster P_i, w_{p_i} is the normalized cluster size of i, and satisfies $\sum_{i=1}^{n} w_{p_i} = 1, 0 < w_{p_i} \le 1$. The earth mover's distance (EMD) [9] has been proved to have promising performance in image retrieval and visual tracking because it can find optimal signature alignment and thereby can measure the similarity accurately. For arbitrary two signatures P and Q, $P = \{(p_i, w_{p_i}), 1 \le i \le m\}$, $Q = \{(q_i, w_{q_i}), 1 \le i \le n\}$, where m and n are the number of clusters in P and Q, respectively. The EMD between P and Q is computed by

$$D(P,Q) = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} d_{ij} f_{ij}}{\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij}} \tag{3}$$

Where d_{ij} is the Euclidean distance between p_i and q_j, and f_{ij} is the optimal match between two signatures P and Q that can be computed by solving the Linear Programming problem.

$$\textbf{min} \quad WORK(P,Q,F) = \sum_{i=1}^{m}\sum_{j=1}^{n} d_{ij} f_{ij}$$

$$\begin{aligned} \textbf{s.t.} \qquad & f_{ij} \geq 0 \\ & \sum_{j=1}^{n} f_{ij} \leq w_{p_i} \\ & \sum_{i=1}^{m} f_{ij} \leq w_{q_i} \\ & \sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij} = min(\sum_{i=1}^{m} w_{p_i}, \sum_{j=1}^{n} w_{q_i}) \end{aligned}$$

For classification, we employ SVM not only because of its solid theoretical basis but also its empirical successes. The EMD distance between frame regions is incorporated into the kernel function of the SVM classification framework by Gaussian function:

$$K(P,Q) = exp(-\frac{1}{\lambda M} D(P,Q)) \tag{4}$$

where M is a normalization factor which is mean value of the EMD distances between all training samples. λ is a scaling factor which is empirically decided by cross validation.

5 Experimental Results

The total length of soccer video is about 95 minutes (670 shots) of "Udinese vs AC Milan" in the united football match of Italy in 640×480 resolution. The first half (311 shots)is used as the training samples and the other (359 shots)as test samples. The first step for shot classification is shot boundary detection. In this paper, we use twin-comparison approach[10] to detect shot boundary, then key frames are extracted for each shot. Due to the computational simplicity of our algorithm, each frame is downsampled, by a rate of four in both direction, that is, 80×60 is the actual frame resolution for shot classifier. Experimental results in Table 2 summarizes the accuracies for each type of events when $\lambda = 2^{-1}$with the best cross-validation rate. The close-up view type has the best performance, while the recall amounts to 1, and the next best is the long view. The performance of the Medium shots detection is the lowest. Some Medium views are mistakenly regarded as the Long views, however, the precision is over 0.9. Overall, the recall-precision rates are satisfactory. We also compared the overall performance of the

Table 2. Experimental results

Shot Class	Total	Our method		Ekin's method	
		precision	recall	precision	recall
Close-up	161	0.96	1	0.87	0.73
Medium view	58	0.95	0.62	0.72	0.67
Long view	140	0.88	0.96	0.71	0.87

proposed method with Ekin's [1] method. Grass colored pixel ratio were used in Ekin's method. Clearly, the proposed method outperforms Ekin's method because the detection grass colored pixel ratio highly depends on the threshold as well as the classification of shots.

6 Conclusion

This paper have proposed a novel method for characterizing dominant color information. The dominant color can be effectively extracted without any threshold setting by the use of dominant sets clustering. The Earth Mover's Distance (EMD), which is a robust similarity measurement, is incorporated into the kernel function of SVM. Experiments show that this model is more effective for shot classification in soccer video. Moreover, this color distribution model can be generalized to analyze other kinds of video and tasks.

Acknowledgment

This work is partly supported by NSFC (Grant No. 60825204 and 60672040) and the National 863 High-Tech R&D Program of China (Grant No. 2006AA01Z453).

References

1. Ekin, A., Tekalp, A.M., Mehrotra, R.: Automatic soccer video analysis and summarization. IEEE Transactions on Image Processing 12(7), 796–807 (2003)
2. Xu, C.S., Wang, J.J., Lu, H.Q., Zhang, Y.F.: A novel framework for semantic annotation and personalized retrieval of sports video. IEEE Transaction on Multimedia 10(3), 421–435 (2008)
3. Xie, L., Xu, P., Chang, S.-F., Dirakaran, A., Sun, H.: Structure analysis of soccer video with domain knowledge and hiddern markov models. Pattern Recognition Letters 25(7), 767–775 (2004)
4. Wang, L., Lew, M., Xu, G.: Offense based temporal segmentation for event detection in soccer video. In: Workshop on Multimedia Information Retrieval (MIR), New York, USA (October 2004)
5. Duan, L.-Y., Xu, M., Chua, T.S., Tian, Q., Xu, C.S.: A unified framework for semantic shot classification in sports video. IEEE Transaction on Multimedia 7(6), 1066–1083 (2005)
6. Tong, X.F., Liu, Q.S., Lu, H.Q.: Shot classification in broadcast soccer video. Vision and Image Analysis 7(1), 16–25 (2008)
7. Duan, L.-Y., Xu, M., Tian, Q., Xu, C.S.: Nonparametric color characterization using mean shift. In: Proceedings of the eleventh ACM international conference on Multimedia, November 2003, pp. 243–247 (2003)
8. Pavan, M., Pelillo, M.: Dominat sets and pairwise clustering. IEEE Trans. Pattern Analysis and Machine Intelligence 29(1), 167–172 (2007)
9. Rubner, Y., Tomasi, C., Guibas, L.J.: The Earth mover's distance as a metric for image retrieval. International Journal of Computer Vision 40(2), 99–121 (2000)
10. Zhang, H., Kankanhalli, S., Soliar, S.: Automatic partitioning of full-motion video. Multimedia Systems 1(1), 10–28 (1993)