Video Shots Comparison using the Mallows Distance

Costantino Grana, Daniele Borghesani, Rita Cucchiara
DII - Università degli Studi di Modena e Reggio Emilia
{surname.name}@unimore.it

Abstract

In this work, we focus on two aspects of the comparison of video shots. We present a new approach to extract a variable number of key frames from a shot, by the use of a hierarchical clustering with automatic level selection, in order to provide optimal allocation of features on different parts of the shot. We then employ the Mallows distance as an effective technique to compare the discrete distributions of features, independently from the features selected for the specific application. Results and comparisons on a soccer documentary video are provided.

1. Introduction

The increasing spread of Video Digital Libraries calls for the design of efficient Video Data Management Systems to manage video access, provide summarization, similarity search, and support queries according with available annotations. General internet users are very demanding in search, so the media search technologies for the mass have to be very simple, intuitive, and easy to use, as text search is [1].

Examples of automatic semantic annotation systems have been presented recently, most of them in the application domain of news and sports video. Most of the proposals deal with a specific context making use of ad-hoc features. In [2] the playfield area, the number and the placement of players on the play field, and motion cues are used to distinguish soccer highlights into subclasses. Differently, a first approach trying to apply general features is described by [3]. Employing color, texture, motion, and shape visual queries by sketches are provided, supporting automatic object based indexing and spatiotemporal queries.

Different systems have been developed to compare shots, and many of these simply extend research results obtained on image retrieval to the analysis of a representative key frame. In this paper we propose a system to generalize this approach by automatically selecting a variable number of key frames, and weighting them based on their significance with respect to the shot. Moreover, as illustrated in a different context in [4], we suggest the use of the Mallows distance in order to provide a flexible distance which just needs the computation of a distance matrix between the selected key frames. This allows the use of features which do not fit in a Euclidean framework and avoid the introduction of weighting factors to be tuned in order to balance the different features contribution.

In the following we introduce the similarity of video clips, describe our key frame extraction approach, and the use of the Mallows distance to compare distributions. The automatic annotation system description and the results are provided in sections 6 and 7, followed by the conclusions.

2. Similarity of Video Clips

The problem of clip similarity can be seen as a generalization of the problem of image similarity: as for images, each clip may be described by a set of visual features, such as color, shape and motion. These are grouped in a feature vector:

$$\mathbf{V}_i = \left[F_i^1, F_i^2, \ldots, F_i^N \right] \quad (1)$$

where i is the frame number, N is number of features and F_i^j is the j-th feature computed at frame i. However, extracting a feature vector at each frame can lead to some problems during the similarity computation between clips, since they may have different lengths; at the same time keeping a single feature vector for the whole clip cannot be representative enough, because it does not take into account the features temporal variability. A simple and common solution can be represented by the use of a fixed number M of feature vectors for each clip, computed on M frames sampled at uniform intervals within the clip. In previous experiments [5], as a tradeoff between efficacy and computational load we used $M = 5$ for clips of averaging 100 frames.

Shot	Length	KeyFrames
3	53	
12	77	
63	123	
64	194	
65	109	
70	80	
105	56	
147	173	

Fig. 1. Example of key frame extraction on a few selected shots.

Here, we propose a new approach to the problem, which aims at providing an equally good description of the frame variability, but with a context dependent decision: we automatically extract a variable number of key frames, and then employ the Mallows distance to compare the discrete distribution obtained from these key frames.

3. Key frame Extraction and Shot Description

Since not all the frames are equally important to obtain the final classification, due to perceptual redundancy in specific domains, we employ a hierarchical clustering method, based on *Complete Link* [6], to aggregate the different frames in each shot, which capture the most significant aspects of the frames variability. This technique guarantees that each clip must be similar to every other in the cluster, while any other clip outside the cluster has dissimilarity greater than the maximum distance between cluster elements. For this clustering method a dissimilarity measure between two clusters W_i and W_j is defined as

$$D(W_i, W_j) = \max_{S_x \in W_i, S_y \in W_j} d\left(S_x, S_y\right). \quad (2)$$

where d is computed as in the following section. The algorithm iteratively searches for the two most similar clusters and then merges them at each step. For each shot, it creates a hierarchy of shot partitions with as much levels as the number of frames in the shot, where level 1 is the final step where everything is merged in a single cluster.

Instead of a manual selection of the desired clustering level, or a threshold guided one, an automatic selection strategy is proposed. Let us define, following the terminology introduced in [7], the cluster diameter and the cluster distance as

$$\Delta\left(W_i\right) = D\left(W_i, W_i\right) \quad (3)$$

$$\delta(W_i, W_j) = \min_{S_x \in W_i, S_y \in W_j} d\left(S_x, S_y\right) \quad (4)$$

The *Clustering Score* at level n is defined as

$$CS_n = \max\left(\Delta_1 - \Delta_n, \delta_n\right) \quad (5)$$

where

$$\begin{aligned} \Delta_n &= \max_{W_i \in E_n} \Delta\left(W_i\right) \\ \delta_n &= \min_{W_i, W_j \in E_n, i \neq j} \delta\left(W_i, W_j\right) \end{aligned} \quad (6)$$

with E_n the set of clusters at level n. The selected level is the one which minimizes CS_n.

To avoid the creation of poorly significant clusters, we add an empirical threshold, which requires that the smallest cluster must be composed by at least 15 frames. From each cluster a key frame is generated as the median frame, computed as the frame which

minimizes the sum of distances from all other frames. An example of results is provided in Fig. 1

4. Mallows Distance

To describe a shot, we want to take into account the number of key frames obtained by the described process, the key frames' features and the cardinality of the clusters. Each shot is thus characterized by a discrete distribution of features:

$$\beta_i = \left\{\left(V_i^1, P_i^1\right), \ldots, \left(V_i^k, P_i^k\right)\right\} \tag{7}$$

where V_i^k is the vector of features extracted for shot *i* at key frame *k*, and P_i^k is the associated probability, defined as the percentage of frames, belonging to the cluster from which the key frame was drawn.

To compute the distance $D(\beta_1, \beta_2)$ between two distributions β_1, β_2, we use the Mallows distance [8,9] introduced in 1972. Consider two probability distributions *P* and *Q* on $\mathbb{R}^n$. Define

$$M = \left\{\text{probability distribution } \mu(\mathrm{x,y}) \text{ on } \mathbb{R}^n \times \mathbb{R}^n\right\} \mid \int_y d\mu(x,y) = P(x), \int_x d\mu(x,y) = Q(y) \tag{8}$$

Mallows proposed to measure the difference between two probability distributions as:

$$Mallows_p(P,Q) = \min_{\mu}\left(E_\mu \left\|x-y\right\|_p^p\right)^{1/p} \tag{9}$$

subject to the constraints of Eq. 8, where the $\|\cdot\|_p$ denotes the L_p norm, and $1 \le p \le +\infty$. For two discrete distributions $P = \left\{(x_1, p_1), \ldots, (x_n, p_n)\right\}$ and $Q = \left\{(y_1, q_1), \ldots, (y_n, q_n)\right\}$, with $\sum p_i = 1$ and $\sum q_i = 1$, minimizing the cost functional reduces to

$$\min_{\mu} \sum_{i=1}^{n}\sum_{j=1}^{m} \mu(i,j) C\left(x_i, y_j\right) \tag{10}$$

subject to

$$\mu(i,j) \ge 0; \sum\nolimits_{j=1}^{m} \mu(i,j) = p_i \quad \sum\nolimits_{i=1}^{n} \mu(i,j) = q_i; 1 \le i \le n; 1 \le j \le m \tag{11}$$

where $C(x_i, y_j)$ is the distance matrix between the elements of the distribution. Note that there is no constraint on *C*, i.e. we can use any formulation we have to compute the base distance between frame features. The dual of the linear programming problem of Eq. 10 is to find $u = [u_1, \ldots, u_n]$ and $v = [v_1, \ldots, v_m]$ in order to solve the problem:

$$\max \sum_{i=1}^{n} p_i u_i + \sum_{j=1}^{m} q_j v_j \tag{12}$$

subject to $u_i + v_j \le C\left(x_i, y_j\right), 1 \le i \le n; 1 \le j \le m$. By solving the dual problem, we achieve better computational efficiency. In our implementation, we used the simplex algorithm to solve the problem. Note that the above formulation gives a set of $n \times m$ constraints (the size of the distance matrix), which may prove difficult to solve, since the required simplex tableau will be $(n \times m) \times (1 + n + m + n \times m)$.

By employing the Mallows distance we can compare two shots, by comparing the associated discrete distribution of features. Since the two distribution may be of different lengths, no constraint is posed on the number of key frames we can choose. Moreover, the distance measure between two frames may be computed with whatever distance we need, and the only requirement is to provide a distance matrix between all frames of the distributions.

5. Frame Features

To provide a general purpose system, we avoid selecting context dependent features, relaying on broad range properties of the clips. To allow easier interoperability and feature reuse, we tried to select features which comply with the MPEG-7 standard [10]. In particular, similarly to what was done in a previous work [5], the following three features are employed:

1. Scalable color: a color histogram, with 16 values in H and 4 values in each S and V (256 bins in total).
2. Color layout: to account for the spatial distribution of the colors, an 8x8 grid is superimposed to the picture and the average YCbCr values are computed for each area.
3. Parametric motion: making use of the MPEG motion vectors, the translational motion of each quarter of frame is estimated.

The distance between two frames is the average between the distances between the corresponding features. The distance between histograms is computed as the histogram intersection, the one between the two color layouts is the average of the color distance in the YCbCr space for all the 64 values in the grid, and the parametric motion is the average of the vector distances of the four quadrants.

6. Automatic annotation

Given a domain-specific video digital library, we assume that it is possible to partition the clips of videos of that context into a set of *L* classes $\mathbf{C} = (C_1, \ldots, C_L)$, which describe different contents or camera views. Given a large set of training clips, we implemented an interactive user-friendly interface to quickly assign each of them to a specific class C_k and then employ it for automatic annotation purposes (more details are given in the following). An unknown clip can be classified using a nearest neighbor approach and the similarity measure defined above.

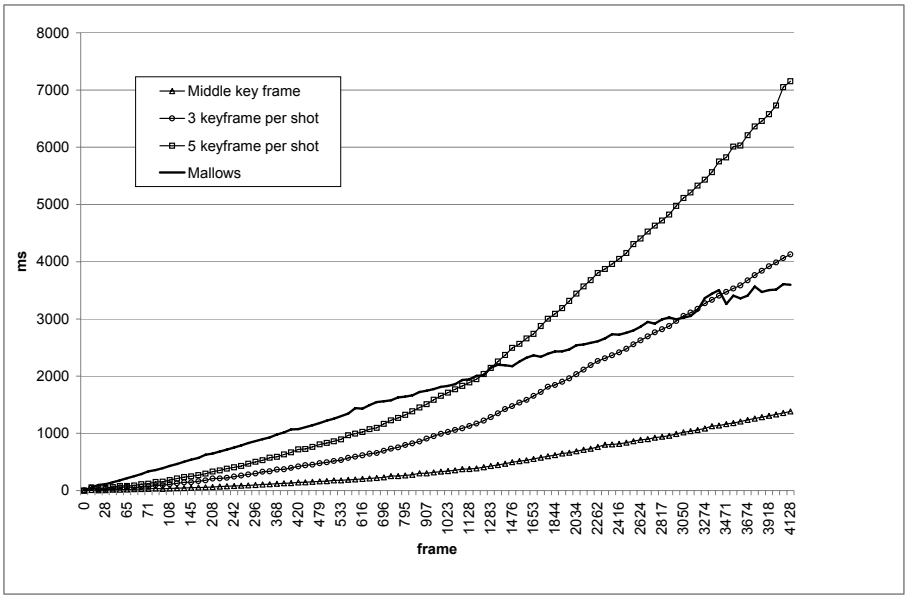

Fig. 2. Processing times (ms)

We developed a program for Microsoft Windows which allows opening videos, provides fast browsing capabilities, allows moving with single frame steps and contains all the described modules. In particular, shot detection may be performed also in batch mode, in order to apply the system to a large set of videos (we used it on TRECVID 2005 and 2006 dataset). The system has a classification scheme manager which allows to define a taxonomy, its classes and to manually associate a shot to a defined concept, describing its contents.

7. Results

The system was tested on a DVD source, a documentary video about the Italian victory of last world cup, so it contains different kinds of generic soccer scenes (classical game framing like short or long views, half body views, faces, little or large groups of fans, interviews with players or coaches, graphical animations). The video follows the PAL format, so it is a 720x576 interlaced video. Prior to elaboration, it was deinterlaced by discarding field 2, horizontally reduced by a factor of 2, and cropped by 40 pixels from top and bottom to remove the useless black stripes. A reference manual annotation was employed: the most interesting categories for our tests are called *Subject*, *Number*, *Framing*, and *View*. Subject describes the character/object depicted in the scene (players, stadium, fans, coach, etc…), Number is the number of people depicted (when significant), Framing describes the direction and distance from which the subject is imaged (long views, half body views, etc…), and View gives a rough description of the position of the subject in the scene. More specific details were omitted, since our aim is focused on general purpose systems, avoiding the use of domain specific feature.

The major result obtained by this system is the ability to get the same (or sometimes better) efficiency, in terms of recall and precision, using a lower amount

Table 1. Results of nearest neighbor classification of shots.

		first	**second**	**third**
middle key frame	*subject*	0.64	0.79	0.84
	number	0.20	0.33	0.39
	framing	0.42	0.73	0.83
	field view	0.59	0.72	0.79
3 key frames	*subject*	0.66	0.78	0.84
	number	0.24	0.33	0.41
	framing	0.44	0.75	0.82
	field view	0.57	0.74	0.80
5 key frames	*subject*	0.69	0.78	0.84
	number	0.26	0.35	0.42
	framing	0.45	0.76	0.83
	field view	0.61	0.73	0.79
variable # of key frames	*subject*	0.69	0.79	0.84
	number	0.27	0.37	0.40
	framing	0.46	0.75	0.83
	field view	0.62	0.73	0.79

of CPU work, This leads to a significant reduction in the required processing time. This peculiarity is much more evident when large amount of data has to be processed.

In Table 1 some experimental result are reported. The first column contains the percentage of cases in which the nearest neighbor was found in the same category. The second and the third columns contain the percentage of cases in which a category match was found at least in one of the first two or three nearest neighbors, respectively. It is possible to observe that we obtain at least the same performance as we expect increasing the number of key frames taken (that indicatively provides a more precise characterization of the shot). Fig. 2 shows the time (in ms) needed to perform the processing over shots. In these tests we are computing the complete distance matrix, which describes all the distances between shots. Because of the increased work needed to compare all shot to all others (the matrix is upper triangular), this time increases dramatically when using a fixed number of key frames per shot, while the increase is much slower when employing the Mallows distance. In addiction to this, the Mallows distance approach shows a linear progression, instead of the exponential progression showed by the fixed number of key frames methods.

8. Conclusions

We presented a new approach to extract a variable number of key frames from a shot and to compare shots based on the features computed on this set of frames. By means of an automatic level selection, the clustering approach allows for an optimal allocation of features on different parts of the shot, while the Mallows distance provides an effective technique to compare the obtained distributions, independently from the features selected for the specific application.

9. Acknowledgments

This work is supported by the DELOS NoE on Digital Libraries, as part of the IST Program of the European Commission (Contract G038-507618).

10. References

[1] A. Jaimes, M. Christel, S. Gilles, R. Sarukkai, W. Ma, “Multimedia information retrieval: what is it, and why isn't anyone using it?” in Proceedings of the 7th ACM SIGMM international Workshop on Multimedia information Retrieval, Hilton, Singapore, November 10-11, 2005.

[2] M. Bertini, R. Cucchiara, A. Del Bimbo, C. Torniai, “Video Annotation with Pictorially Enriched Ontologies,” in Proceedings of IEEE International Conference on Multimedia and Expo, Amsterdam, The Netherlands, 1428-1431, 2005.

[3] S. Chang, W. Chen, H.J. Meng, H. Sundaram, D. Zhong, “A Fully Automated Content-Based Video Search Engine Supporting Spatiotemporal Queries,” in IEEE Transactions on Circuits and System for Video Technology, vol. 8 num.5, 602-615, 1998.

[4] J. Li, J.Z. Wang “Real-time computerized annotation of pictures,” in Proceedings of the ACM Multimedia Conference, ACM, Santa Barbara, CA, pp. 911-920, 2006.

[5] R. Vezzani, C. Grana, D. Bulgarelli, R. Cucchiara, “A semi-automatic video annotation tool with MPEG-7 content collections,” in Proceedings of IEEE International Symposium on Multimedia (ISM2006), San Diego (CA), USA, pp. 742-745, Dec. 11-13, 2006.

[6] A.K. Jain, R.C. Dubes, “Algorithms for clustering data,” Prentice-Hall, Englewood Cliffs, NJ, 1988.

[7] J. Dunn, “A fuzzy relative of the Isodata process and its use in detecting compact, well-separated clusters”, in Journal of Cybernetics, vol. 3, n.3, pp. 32–57, 1973.

[8] C. L. Mallows, “A not on asymptotic joint normality,” Annals of Mathematical Statistics, vol. 43, no. 2, pp. 508–515,1972.

[9] D. Zhou, J. Li, H. Zha, “A New Mallows Distance Based Metric For Comparing Clusterings,” in Proceedings of the 22nd International Conference on Machine Learning, Bonn, Germany, 2005.

[10] Information technology - Multimedia content description interface - Part 3: Visual, ISO/IEC Std. 15 938-3:2003, 2003.