Video Clip Clustering for Assisted Creation of MPEG-7 Pictorially Enriched Ontologies

Costantino Grana, *Member, IEEE*, Daniele Bulgarelli, Rita Cucchiara, *Member, IEEE*

***Abstract*—In this paper, we present a system for the assisted creation of *Pictorially Enriched Ontologies*, that is ontologies for context-based digital libraries enriched by pictorial concepts for video annotation, summarization and similarity based retrieval. Here we detail the approach for video clips clustering and pictorial concepts extraction together with the approach for storing the ontology within the MPEG-7 framework. The clustering is performed by Complete Link hierarchical clustering on color histograms and motion features. Results on Formula 1 TV material are reported.**

I. Introduction

The wide circulation of camera equipped cellular phones, consumer video cameras and Internet broad band connections, has nowadays delivered a huge amount of audio-visual data which needs to be stored and organized before its use. Efficient Video Data Management Systems must be designed to index the video according to its content and allow effective browsing and access. Automating large part of the exasperating process of annotation at the low level feature analysis stages to aid human operators would be an advantageous and achievable goal. Video summaries, for example, result in very compressed representation of the video without losing crucial contents and allow efficient browsing as well as a fast overview of the original contents by dropping the time spent on tedious operations such as fast forward and rewind. In this aspect, video summarization is popularly regarded as a good approach to the content-based representation of videos [1].

Video summarization, video abstraction and, more in general, efficient access to video content should require a semantic-level interpretation of the video clips. In this manner the search, indexing and the summaries creation with key frames or representative clips could be improved with content-based similarity paradigms. What we need is some manual and user friendly tool to ease this first classification process to extract and represent visual knowledge of the video clips' content. Visual knowledge can be described with textual ontologies that are normally adopted to structure the concepts and their relationships in context-based Digital Libraries (DL). We define "Pictorially Enriched Ontology" a new paradigm that improves the expressive power of the ontology by adding "pictorial concepts", i.e. concepts represented by archetypal images, key-frames, or visual objects. In this manner suitable specializations of textual concepts can be described synthetically with pictorial concepts [2]. However, the creation of Pictorially Enriched Ontologies is extremely challenging both for the pictorial concepts extraction and selection and for the choice of a common language to communicate the ontology structure.

The purpose of this work is the definition of an approach to analyze video clips and provide an assisted segmentation into semantic classes with similar visual contents. This kind of systems should be computationally manageable, content based, unsupervised, and flexible. This paper presents a framework for clustering shots or sub shots in given semantic areas. A standard (i.e. textual) ontology is assumed available for the Video DL, composed by a manifold of edited videos whose content can be described at high level with the ontology. Moreover, we assume to have a manual o automatic procedure to segment the video into meaningful clips [3].

After a first manual classification of the clips, based on the textual taxonomy, we provide an automatic grouping into classes which present similar contents. The unsupervised grouping technique used is the Complete Link Hierarchical Clustering [4], with a new criterion for suggesting the best level of the hierarchy, based on intra-cluster dispersion and inter-cluster distances ratios. Each video clip is represented in a compact way by means of the median color histogram in the quantized HSV color space and by the use of a set of motion features extracted by the MPEG motion vectors. These are extracted as detailed in [5]. In order to store and deliver the information on the Pictorially Enriched Ontology, we employ MPEG-7 descriptors and their associated Multimedia Descriptors Schemes [6].

II. Video Clips Similarity

A *shot* is a continuous camera run from the time the recording starts to the time the recording stops. A shot may last a few seconds, several minutes, or the entire program. For this reason in this work the focus is on *clips*, which could be shots, but also shorter elements which allow to better represent the content structure. Clips can be manually or automatically extracted segmenting shots with some

Manuscript received October 10, 2005. The work is supported by the DELOS Network of Excellence on Digital Libraries, as part of the IST Program of the European Commission (Contract G038-507618).

C. Grana, D. Bulgarelli and R. Cucchiara are with the Information Engineering Department, University of Modena and Reggio Emilia, Italy (phone: +390592056142; fax: +390592056129; e-mail: surname.name@unimo.it).

perceptual cues. Especially in sport videos, color and motion are the most representative features for visual knowledge.

A common way to describe the color contents of a video segment S is to use the histograms of its frames. A shorter representation is the *Median Histogram* H^S, which is obtained by sorting the array of frame histogram values for every bin in ascending order, and taking only the central member of the ordered array. The other feature used is the mean of the motion features, based on sub-image average MPEG motion vectors, as defined in [5].

We define a dissimilarity index between S_i and S_j as

$$d\left(S_i, S_j\right) = \frac{1}{2}\left(D_H\left(H^{S_i}, H^{S_j}\right) + D_M\left(M^{S_i}, M^{S_j}\right)\right) \quad (1)$$

where $D_H(\cdot,\cdot)$ is the dissimilarity between two histograms and $D_M(\cdot,\cdot)$ is the one between two motion features. These values are normalized between 0 and 1. The distance between histograms has to cope with different totals (the median doesn't in fact preserve the sum over all bins), so we compute this distance as described in [7]:

$$D_H\left(H^{S_i}, H^{S_j}\right) = \frac{1}{2\sqrt{N_{H^{S_i}} N_{H^{S_j}}}} \sum_{k=1}^{B} \left| C_1 H^{S_i}(k) - C_2 H^{S_j}(k) \right| \quad (2)$$

where H^{S_i}, H^{S_j} denote the median histograms of clips *i* and *j*, *B* is the number of histogram bins, and $C_1 = \sqrt{N_{H^{S_j}} / N_{H^{S_i}}}$, $C_2 = \frac{1}{C_1}$, $N_{H^{S_i}} = \sum_{k=1}^{B} H^{S_i}(k)$.

The distance between motion features $D_M(\cdot,\cdot)$ is given by the Euclideandistance between the feature vectors, after normalization with respect to image size.

III. Clustering of Video Clips

After identifying the features of each clip and defining the equation which determines their distance, denoting by C_i the i^{th} cluster, the following criterion is used:

$$x \in C_i \Leftrightarrow d(x,y) > \varepsilon_i, \ \forall y \in C_j, \ j \neq i \quad (3)$$

$$\varepsilon_i = \max_{x,y \in C_i} d(x,y) \quad (4)$$

This criterion implies a strong condition on the similarity of clips in a cluster, because each clip must be similar to every other in the cluster. In this case, with each cluster C_i is an associated maximum dissimilarity ε_i defined by Eq. 4, which measures the maximum dissimilarity between any two clips in the cluster. Any other clip outside the cluster must have dissimilarity greater than ε_i. Hierarchical clustering methods based on *Complete Link* generate clusters which satisfy the previous condition. For this clustering method we defined the dissimilarity between two clusters C_i and C_j as

$$d(C_i, C_j) = \max_{x \in C_i, y \in C_j} d(x,y) \quad (5)$$

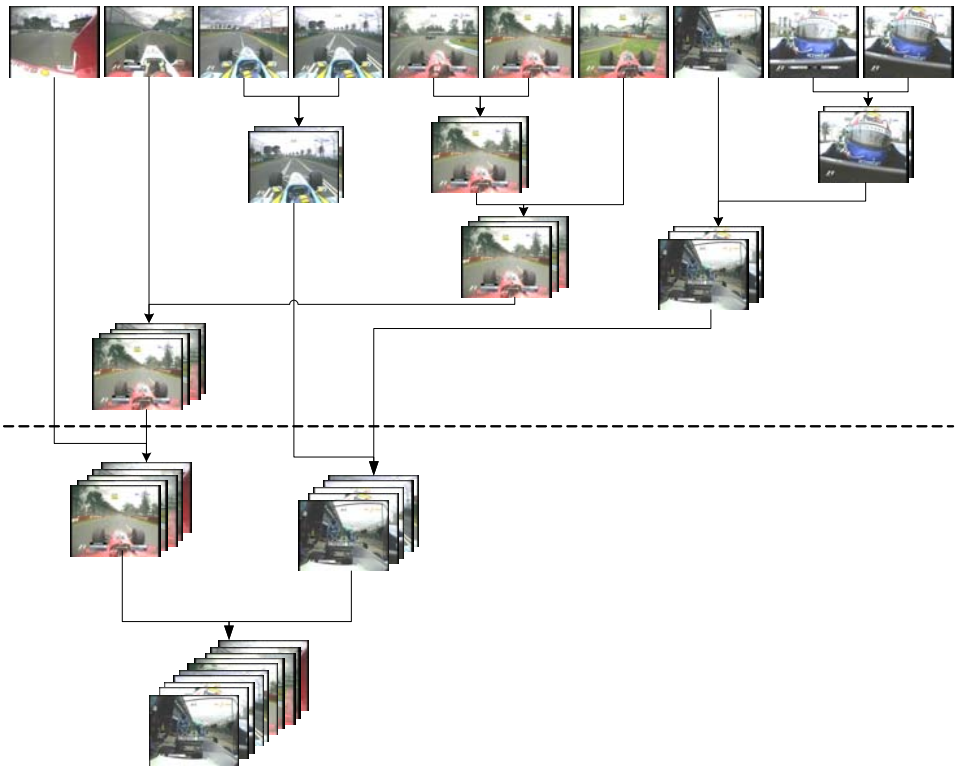

Fig. 1 Example of complete link clustering on 10 sub-shots. The Dunn's Index selected level is shown with a dotted line.

The algorithm proceeds as follows:

1. Initially we have N clusters $\{S_1\},\{S_2\},\ldots,\{S_N\}$. Let's call E the set of clusters. Each cluster contains a single clip.
2. Find the least dissimilar pair of clusters, R and S, according to (5), i.e., find R and S such that
$$d(S,R) \le d(A,B) \;\; \forall A,B \in E.$$
3. Merge R and S into a new cluster.
4. If evereything is merged in a single cluster then stop, else go to step 2.

This algorithm produces a hierarchy of clips partitions with N levels and i clusters at level i (the initial level is N). To implement the algorithm a proximity matrix D was used. An $N \times N$ proximity matrix $D = \left[d(S_i, S_j) \right]$ has as entries the dissimilarity between two shots. At each step, the matrix is updated by deleting rows and columns corresponding to cluster R and S and adding a new row and column corresponding to the newly formed cluster. The values in the new row/column are the maximum of the values in the previous ones. Initial generation of matrix D requires $\frac{1}{2}N(N-1)$ computations of $d(\cdot,\cdot)$.

To provide the user with a first selection of the number of clusters, thus an automatic solution, we chose to avoid using fixed thresholds, but employed Dunn's *Separation Index* [8]. Let us define

$$\Delta(C_i) = d(C_i, C_i) \tag{6}$$

and

$$\delta(C_i, C_j) = \min_{x \in C_i, y \in C_j} d(x,y) \tag{7}$$

called respectively *Diameter* and *Set Distance*. The Separation Index at level n is:

$$SI_n = \frac{\min_{1 \le i,j \le n, i \ne j} \delta(C_i, C_j)}{\max_{1 \le k \le n} \Delta(C_k)} \tag{8}$$

The proposed level is the one which maximizes SI_n.

After interaction with the user, who can accept or modify the number of clusters, or individually include or exclude them from the ontology,

IV. Representing Pictorially Enriched Ontologies with MPEG-7

An ontology defines the terms used to describe and represent an area of knowledge. Ontologies are used by people, databases, and applications that need to share domain information. Ontologies include computer-usable definitions of basic concepts in the domain and the relationships among them [9]. These can be effectively defined with OWL, but the language does not embody any construct for including a pictorial representation. This is instead included in the MPEG-7 standard [6]. MPEG-7 has much less sophisticated tool for knowledge representation, since its purpose of standardization limits de-facto the definition of new data types, concepts and complex structures. Nevertheless the MPEG-7 standard can naturally include pictorial elements such as objects, key-frames, clips and visual descriptors that can be used in the ontology description.

Therefore our system stores the *Pictorially Enriched Ontology* following the directions of the MPEG-7 standard, and in particular it uses a double description provided by the *ClassificationSchemeDescriptionType* DS combined with a *ModelDescriptionType* DS which includes a *Collection ModelType* DS. The classification scheme allows the definition of a taxonomy or thesaurus of terms which can be organized by means of simple term relations. The collection model instead is an *AnalyticModel*, that is it describes the association of labels or semantics with collections of multimedia content. The collection model contains a *Content CollectionType* DS which contains a set of visual elements which refer to the model being described. In particular we

```
<?xml version="1.0" encoding="iso-8859-1"?>
<Mpeg7 xmlns="urn:mpeg:mpeg7:schema:2001"
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xmlns:mpeg7="urn:mpeg:mpeg7:schema:2001"
  xsi:schemaLocation="urn:mpeg:mpeg7:schema:2001 Mpeg7- 2001.xsd">
  <Description xsi:type="ClassificationSchemeDescriptionType">
    <ClassificationScheme uri="urn:mpeg:mpeg7:cs:Formula1">
      <Term termID="Accident"/>
      <Term termID="Box"/>
      <Term termID="CameraCar"/>
      <Term termID="CloseUp"/>              <!-- This is the ontology -->
      <Term termID="LargeView"/>
      <Term termID="PitStop"/>
      <Term termID="Other"/>
    </ClassificationScheme>
  </Description>
  <Description xsi:type="ModelDescriptionType">
    <Model xsi:type="CollectionModelType">
      <Label href="urn:mpeg:mpeg7:cs:Formula1:CameraCar"/>
      <Collection xsi:type="ContentCollectionType">
        <VisualFeature xsi:type="ScalableColorType" numOfCoeff="16"
          numOfBitplanesDiscarded="0" >
          <Coeff>  187 123 99 283 124 188 43 72 339 0 22 482
                   208 31 92 382</Coeff>
        </VisualFeature>              <!-- These are the pictorial concepts -->
        <Content xsi:type="VideoType">
          <Video>
            <MediaLocator xsi:type="TemporalSegmentLocatorType">
              <MediaUri>file://race1.mpeg</MediaUri>
              <MediaTime>
                <MediaTimePoint>T00:00:02:20080F30000
                  </MediaTimePoint>
                <MediaDuration>PT2S15075N30000F</MediaDuration>
              </MediaTime>
            </MediaLocator>
          </Video>
        </Content>
      </Collection>
    </Model>
  </Description>
</Mpeg7>
```

Fig. 2. Example of an MPEG-7 description of a Pictorially Enriched Ontology. In this extremely short example two descriptions are present: the taxonomy and a class' pictorial elements (*CollectionModel*). Only a single clip is included with an example Visual Descriptor.

link the selected clips and a representation of their features by means of the *Scalable Color* D and the *Mixture Camera Motion Segment Type* D. In Fig. 2, an example of an MPEG-7 description of a Pictorially Enriched Ontology is provided, in which it is possible to see the two different parts, i.e. the ontology and the pictorial concepts.

V. Example of Application

The system was applied on a set of Formula 1 videos, in order to test the feasibility of the proposed technique as an assisted ontology creation. After a first manual annotation of the video, we automatically produced a set of representatives for each class, and then let the user select which one he would like to include in the final ontology as a separate pictorial concept. The choice of the cluster representative was done by selecting the clip with the least distance from the median color histogram.

With this system, the number of elements which was presented to the user selection was reduced to less then 10%, with respect to the original number of clips. The clustering results were found satisfactory, and the automatic level selection needed to be manually adjusted only rarely. Fig. 3 shows the results of the clustering procedure on an unclassified video sequence in which the ability of our system to select significant representatives is shown.

VI. Conclusions

We presented a system for the assisted creation of Pictorially Enriched Ontologies and a complete MPEG-7 representation of the annotated taxonomy. This system fits in the framework of *assisted creation* since the first taxonomy and class selection for the video is done manually, and the software simply provides a visual tool which allows automatic video segmentation. The user is later aided by an automatic selection of a number of examples, chosen as representatives of the defined classes. The automatic selection of examples also allows a more effective selection, with respect to manual decisions, since this is based on the features used during the retrieval phase.

The use of the ontology for inference on other video material which refers to the same domain would allow a system to provide flexible and extensible annotations.

Acknowledgment

We would like to thank Filippo Barbieri for the work he did on clustering and code fixing.

References

[1] M. Yeung, B.L. Yeo, *Segmentation of video by clustering and graph analysis*, in Computer Vision and Image Understanding, Vol. 71 No. 1, July, pp 94-109, 1998.

[2] M. Bertini, R. Cucchiara, A. Del Bimbo, C. Torniai, "Domain Knowledge Extension with Pictorially Enriched Ontologies," in Proceedings of The 11th International Conference on Computer Analysis of Images and Patterns (CAIP), Sep 2005, pp. 652-660.

[3] C. Grana, G. Tardini, R. Cucchiara, MPEG-7 Compliant Shot Detection in Sport Videos, in Proc of IEEE Int Symposium on Multimedia (ISM2005), Dec 12-14, 2005, Irvine, California, USA.

[4] A.K. Jain and R.C. Dubes, Algorithms for clustering data. Englewood Cliffs, NJ: Prentice-Hall 1988.

[5] G. Tardini, C. Grana, R. Marchi, R. Cucchiara, Shot Detection and Motion Analysis for Automatic MPEG-7 Annotation of Sports Videos, in Proceedings of the 13th ICIAP, Sept. 2005, pp. 653-660.

[6] MPEG MDS Group, Text of ISO/IEC FDIS 15938-5 Information Technology Multimedia Content Description Interface Part 5 Multimedia Description Schemes, ISO/IEC JTC1/SC29/WG11 N4205, July 2001.

[7] A.M. Ferman, A.M. Tekalp, Two-stage hierarchical video summary extraction to match low-level user browsing preferences, IEEE Transactions on Multimedia, Vol.5, Iss.2, June 2003, 244- 256.

[8] J.C. Dunn, "A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters," J. Cybern., vol. 3, no. 3, pp. 32-57, 1973.

[9] OWL Web Ontology Language Use Cases and Requirements, Jeff Heflin, Editor, W3C Recommendation, 10 February 2004, http://www.w3.org/TR/2004/REC-webont-req-20040210/.

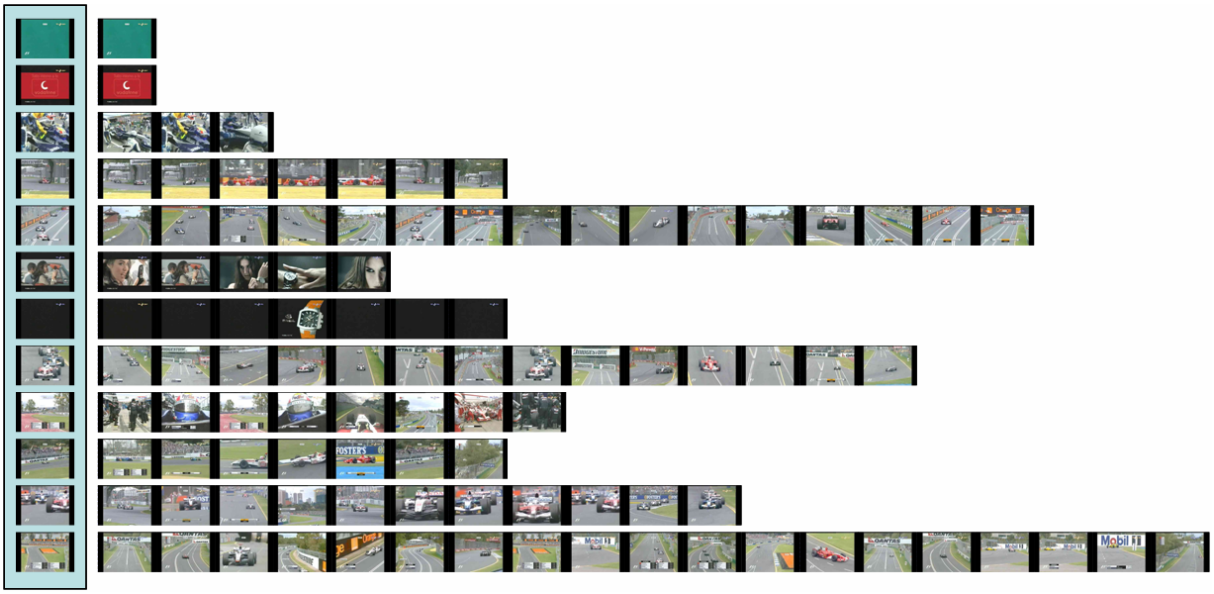

Fig. 3 Complete clustering of a short Formula 1 sequence. The column on the left shows the automatic extracted representatives for each cluster.