MPEG-7 Pictorially Enriched Ontologies for Video Annotation

C. Grana, R.Vezzani, D. Bulgarelli, R. Cucchiara
Dipartimento di Ingegneria dell'Informazione
Università degli Studi di Modena e Reggio Emilia

Abstract. A system for the automatic creation of *Pictorially Enriched Ontologies* is presented, that is ontologies for context-based video digital libraries, enriched by pictorial concepts for video annotation, summarization and similarity-based retrieval. Extraction of pictorial concepts with video clips clustering, ontology storing with MPEG-7, and the use of the ontology for stored video annotation are described. Results on sport videos and TRECVID2005 video material are reported.

1. Introduction

The emerging interest in Video Digital Libraries (DLs) requires efficient solutions for video annotation and access to its content. Video annotation associates to each video element a suitable description, stored in the Video DL together with the visual content. Often, MPEG-7 is adopted as a unified standard that includes visual parts and XML-based descriptions. Due to the inflexibility of the standard, non-compliant MPEG-7 extensions often are proposed [1]; as an alternative, RDF-based descriptions are adopted using OWL for representing the semantics associated with the video content [2].

However, in Video DLs, most of the content cannot be easily described by linguistic terms only, but the intrinsic visual content can be better expressed with perceptual cues. Since humans make use of their visual knowledge making mental images of the world, the video annotation should include a pictorially enriched knowledge of the video content. This work is aimed at the exploration of this new paradigm. Annotation of Video DLs needs an initial video partitioning into basic structural elements. This work is focused on *clips*, which are time segments that are manually or automatically extracted using certain perceptual cues or certain metadata embedded in the video [3,4]. Clips can correspond either to shots or to sub-shots in edited videos. A video is divided in hundreds of clips of different length, which must be annotated.

In context-based DLs, annotation is generally based on a defined ontology. 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. The traditional approach to video annotation consists of two steps:
1) manual creation of the (textual) ontology where concepts and relationships are formalized. In the simplest cases, the ontology is a taxonomy defining the semantic categories only;
2) annotation, by means of a manual or automatic matching of unknown clips against the concepts of the ontology. In case of automatic annotation, finite state automata or probabilistic reasoning are employed for a supervised classification of clips [5].

We propose to employ *pictorially-enriched (pict-en) ontologies* and perceptual similarity to provide automatic video annotation. The new framework consists in the following new steps:

1) manual creation of the textual ontology as described before, but enriched with pictorial concepts: some prototypal clips (called *pictorial concepts* or *prototypes*) are extrapolated from the video and associated with each category as a visual specialization of the textual concepts of the ontology. This can be done manually or automatically. Manual definition is not trivial since the prototypes should be selected after an analysis of a large manifold of video clips. Thus, we propose an automatic selection with an unsupervised hierarchical clustering, which extracts a variable number of pictorial concepts from large training sets of clips of each category;

2) annotation by means of automatic matching of unknown clips against the pictorial concepts of the ontology. This is based on perceptual-level similarity, such as motion, color or shape which are present in the clips.

Unlike other proposals [9], no semantic-level rules requiring a priori knowledge of the context are employed, nor are predefined reasoning automata. Instead, perceptual similarity alone guides the association of each clip to a pictorial concept, and consequently to the concept that corresponds to its generalization. Results are very promising, especially considering that the approach is very general and can be applied to any video DL whatsoever. The only context-based procedure is the initial textual ontology creation. The sole requirement is the availability of a sufficiently large training set of clips to select pictorial concepts. Therefore, since annotation is purely based on visual similarity, this approach is particularly suitable in context-based Video DLs, such as DL of a given sport where many repetitive and similar actions are collected.

In this paper, the framework is defined and results are provided both for specific sport video clips of Formula 1 races and for generic video clips taken form TRECVID 2005.

2. Similarity of video clips

The problem of clip similarity is a generalization of the problem of image similarity: as for images, each clip is described by a set of visual features. For every frame, some features about motion, color and shape can be exploited in order to create a representative vector:

$$V^{fr} = \left[V^{fr,1}, V^{fr,2}, .., V^{fr,fe} \right] \quad (1)$$

where *fr* is the index of frames and *fe* is index of features. However, in clip similarity, feature vectors cannot be extracted for each frame since clips can have different length, nor can a single feature vector for the whole clip be representative enough, because it does not take into account the temporal variability of the features. Here, a fixed number *C* of feature vectors are defined for each clip, computed on frames sampled at uniform intervals within the clip. For each *fr* frame, we exploit four types of features:

1. The color histogram, in 256 bins of HSV color space.
2. The 64 spatial color distributions: to account for the spatial distribution of the colors, an 8x8 grid is superimposed to the frame and the mean YCbCr color is computed for each area.
3. The 64 DCT coefficients for the frame texture: the coefficients of DCT transform are applied following the MPEG-7 specification for the Color Layout Descriptor.
4. The four main motion vectors: they are computed as the average of the MPEG motion vectors [7], extracted in each quarter of frame. The median value has been adopted since MPEG motion vector are not always reliable and are often affected by noise .

A dissimilarity index between two clips S_i and S_j is defined as

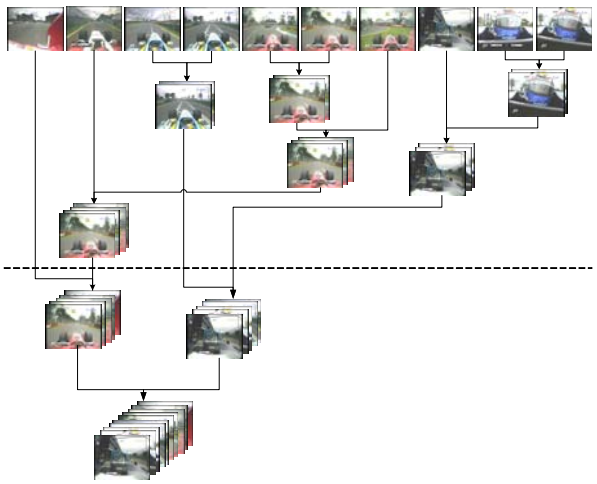

Fig. 1. Example of complete link clustering on 10 sub-shots. The Dunn's Index selected level is shown with a dotted line.

$$d\left(S_i, S_j\right) = \sum_{fr=1}^{C} \sum_{fe=1}^{N} k_{fe} \left\| V^{S_i, fr, fe} - V^{S_j, fr, fe} \right\|_1 \tag{2}$$

where, V is the feature vector, *fe* is the feature-type index, *fr* is the frame index of clip *S*. The L^1-norm between the feature vectors is multiplied by an appropriate constant k_{fe}. These empirical constants are tuned to optimize the classification results on a second training video, different from the videos employed during the ontology creation.

3. Pict-En ontology creation

After the definition of the textual domain ontology, a pict-en ontology requires the selection of the prototypal clips that can constitute pictorial concepts as specialization of each ontology category. Large training sets of clips for each category must be defined and an automatic process extracts some visual prototypes for every category.

Using the previously defined features and dissimilarity function, we employ a hierarchical clustering method, based on *Complete Link* [6]. This technique guarantees that each clip must be similar to every other in the cluster and any other clip outside the cluster has dissimilarity greater than the maximum distance between cluster elements. For this clustering method we defined the dissimilarity between two clusters C_i and C_j as

$$d(C_i, C_j) = \max_{x \in C_i, y \in C_j} d\left(x, y\right) \tag{3}$$

The algorithm proceeds as follows:

1. Initially we have *M* clusters $\{S_1\}, \{S_2\}, \ldots, \{S_M\}$. Let us call *E* the set of clusters. Each cluster contains a single clip.
2. The least dissimilar pair of clusters, R and S, is found according to Eq. 3, i.e. R and S are found such that
$$d\left(S, R\right) \le d\left(A, B\right) \ \forall A, B \in E. \tag{4}$$
3. R and S are merged into a new cluster.
4. If everything is merged in a single cluster then the algorithm stops, else it resumes from to step 2.

This algorithm produces a hierarchy of clips partitions with *M* levels and *i* clusters at level *i* (the initial level is *M*). To implement the algorithm, a proximity matrix *D* was used: initially, it contains the distances between each pair of clips. At each step, the matrix is up-

```
<?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="CameraCar"/>              <!-- This is the ontology -->
      <Term termID="Spectators"/>
      <Term termID="ExternalCarView"/>
      <Term termID="People"/>
    </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.

dated by deleting rows and columns corresponding to clusters 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. The generation of the proximity matrix requires $\frac{1}{2}M\left(M-1\right)$ computations of $d\left(\cdot,\cdot\right)$.

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

$$\Delta\left(C_i\right)=d\left(C_i,C_i\right) \tag{5}$$

$$\delta(C_i,C_j)=\min_{x\in C_i,y\in C_j} d\left(x,y\right) \tag{6}$$

The Separation Index at level *n* is:

$$SI_n=\frac{\min\limits_{1\le i,j\le n,i\neq j}\delta\left(C_i,C_j\right)}{\max\limits_{1\le k\le n}\Delta\left(C_k\right)} \tag{7}$$

The proposed level is the one which maximizes SI_n.

This approach was tested on different training sets of video clips: the automatic solution has been visually evaluated and was judged satisfying enough. An assisted approach may be employed to allow the user to enlarge or restrict the number of clusters by browsing among the cluster hierarchy. To characterize every visual prototype, the average feature set, at each of the *C* selected frame is used, and the clip which minimizes the distance from the average is employed.

	Classification							
	Camera car		**E**xternal car view		**S**pectators		**P**eople	
C	40	45%	20	22%	0	0%	30	33%
E	10	1%	1050	95%	30	2%	20	2%
S	0	0%	0	0%	20	100%	0	0%
P	10	8%	10	8%	0	0%	100	84%

Table 1. Confusion matrix for the test on Formula 1 videos.

4. Ontologies in MPEG-7

Ontologies may be effectively defined with OWL, but this language does not contain any construct for including a pictorial representation. On the other hand, such feature is present in the MPEG-7 standard. MPEG-7 has much less sophisticated tools for knowledge representation, since its purpose of standardization limits 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 in the ontology description.

Therefore, our system stores the pict-en ontology following the directions of the MPEG-7 standard, and in particular 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 (see Fig. 2 for the list of terms of the Formula 1 taxonomy). The collection model is instead an *AnalyticModel,* and therefore it describes the association of labels or semantics with collections of multimedia content. The collection model contains a *Content CollectionType* DS with a set of visual elements which refer to the model being described. In particular, we linked the selected clips and a representation of their features by means of the *Scalable Color* D, *Color Layout* D and the *Mixture Camera Motion Segment Type* D. In Fig. 2, an example of an MPEG-7 description of a pict-en ontology is provided, wherein the two different parts, i.e. the ontology and the pictorial concepts, can be seen.

5. Automatic annotation

The pict-en ontology, could be directly employed to give a visual index of the elements of the videos. Indeed, its significance is related with its usefulness in the annotation of new clips of the same domain. Annotation is provided by the nearest neighbor rule: from the new clip, *C* feature vectors are extracted and instead of using all the elements of the training set, the distance is computed only against the prototypes that are stored with their features in the pict-en ontology. In this manner we decouple the phase of ontology creation, which can be very time consuming, and the phase of the annotation that should be done in a fast way. Using MPEG-7 for feature description, allows the pict-en ontology to be used without problems by different systems, without the need of using the same software library. Moreover, the match may be performed on a subset of the features included into the ontology, for example for time constraints. Moreover the procedure is very general , and after the initial textual taxonomy creation, all the other steps are automatic and can be applied in different video contexts. In our experiments, a good tradeoff between efficacy and computational load suggest the use of $C = 5$.

	Classification					
	Group of people		**P**erson in foreground		**C**omputer graphics	
G	132	(77%)	28	(16%)	12	(7%)
P	135	(52%)	119	(46%)	4	(2%)
C	2	(3%)	29	(40%)	41	(57%)

Table 2. Classification performance on a subset of TRECVID2005 videos.

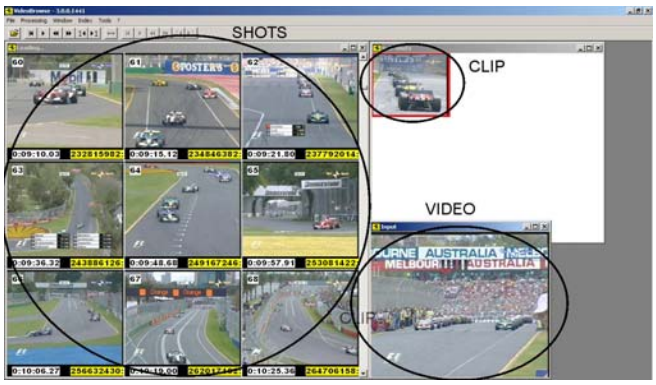

Fig. 3. Example interface output for a query of `<external car view>`.

The first extensive test of the system has been done on available video DLs of Formula 1 races of the last three years. The classes constituting the ontology are very simple and are: `<camera car>`, `<spectators>`, `<external car view>`, and `<people>` (such as mechanics, journalists, or pilots). 340 clips were employed to create the ontology, with a total duration of 18 minutes (27800 frames). The tests were conducted on a 90-minutes video, with 1340 clips (135289 frames). The achieved results are shown in Table 1, while Fig. 3 shows an example of the software interface in a query operation.

Each new clip has been classified against the prototype and the correspondent textual concept has been annotated. The total percentage of correctly classified clips is 90.3%. Apart from the `<spectators>` clips, which are not really significant, in `<external car view>` a really good recall rate can be observed. On the other hand, the classification of "camera car" clips was not as effective. A possible reason for this is probably related to the low frequency of camera car clips on the ontology training set. In fact, clips belonging to `<external car view>` were 58% of the total, while `<camera car>` were only 10%. It is important to consider that a larger Digital Library would likely show a higher number of sub-classes, which would increase the computation time. In the current tests, a 10 minutes video requires 6 minutes for the automatic classification and annotation.

A second test was based on a subset of TRECVID2005 dataset: we used a *general broadcasting material* ontology, made of 280 clips. In this test the classes were `<group of people>`, `<person in foreground>`, `<computer graphics>` (e.g. at the end of a news program) (Fig. 4). As expected, the results are worse compared to a domain-specific video DL. Due to the high variability of perceptual aspects of the clips, an average 58% of correct recall has been reached (Table 2). It is worth noting that no threshold has been used in the annotation, and no category of "unknown" clips, which could have improved the precision, has been added . This result is satisfactory, since it means that more than one clip out of two can be automatically annotated, without semantic rules and only using perceptual similarity. Nevertheless, we think that the creation of a complete pict-en ontology with source material of such a broad and general domain is not feasible without obtaining an explosion of ontology size and computation requirements.

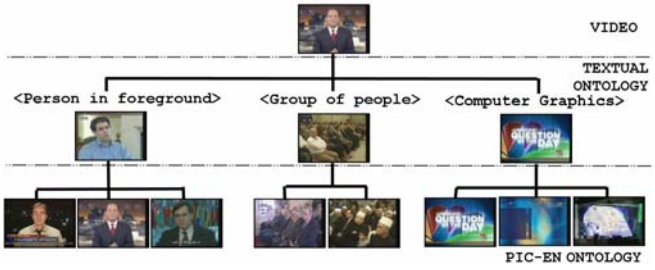

Fig. 4. Graphical representation of the TRECVID test ontology. The clips are grouped into three classes, each of which is described by a set of visual prototypes. Only very few prototypes are shown for space constraints.

6. Conclusions

We presented a system for the creation of a specific domain ontology, enriched with visual features and references to multimedia objects. The ontology is stored in MPEG-7 compliant format, and can be used to annotate new videos. The annotation has shown satisfactory results when the domain is narrow and well defined, and becomes less effective on contents from too broad a context. Nevertheless, this approach allows a system to behave differently by simply providing a different ontology, thus expanding its applicability to mixed sources Digital Libraries.

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). We would like to thank Filippo Barbieri for his work on clustering and code fixing.

References

[1] J. Ricard, D. Coeyrjolly, and A. Baskurt, Art extension for description, indexing and retrieval of 3D Objects, in Proceedings of the 17th International Conference on Pattern Recognition, Cambridge, UK, (2004) 79-82.

[2] A. Chebotko, Y. Deng, S. Lu, F. Fotouhi, A. Aristar, H. Brugman, A. Klassmann, H. Sloetjes, A. Russel, and P. Wittenburg, OntoELAN: an Ontology-based Linguistic Multimedia Annotator, Proceedings of the 6th International Symposium on Multimedia Software Engineering (ISMSE 2004), Miami (FL), USA, (2004) 329-336.

[3] C. Grana, G. Tardini, and R. Cucchiara, MPEG-7 Compliant Shot Detection in Sport Videos, in Proceedings of the IEEE International Symposium on Multimedia (ISM 2005), Irvine (CA), USA, (2005) 395-402.

[4] N. Lian, and Y. Tan, Probabilistic approach to *k*-nearest neighbor video retrieval, Proceedings of the IEEE International Symposium on Circuits and Systems (ISCAS 2004), Vancouver, Canada, (2004) 193-196.

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

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

[7] G. Tardini, C. Grana, R. Marchi, and R. Cucchiara, Shot Detection and Motion Analysis for Automatic MPEG-7 Annotation of Sports Videos, in Proceedings of the 13th International Conference on Image Analysis and Processing, Cagliari, Italy, (2005) 653-660.

[8] J.C. Dunn, A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters, Journal of Cybernetics, vol. 3 num. 3 (1973) 32-57.

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