Managing Video Collections at Large

Nicolas Moënne-Loccoz Bruno Janvier Stéphane Marchand-Maillet Eric Bruno
Viper Group, Computer Vision and Multimedia Lab
24, Rue du général Dufour,
1211 Geneva 4 – Switzerland
Nicolas.Moenne-Loccoz@cui.unige.ch

ABSTRACT

Video document retrieval is now an active part of the domain of multimedia retrieval. However, unlike for other media, the management of a collection of video documents adds the problem of efficiently handling an overwhelming volume of temporal data. Challenges include balancing efficient content modeling and storage against fast access at various levels. In this paper, we detail the framework we have built to accommodate our developments in content-based multimedia retrieval. We show that not only our framework facilitates the developments of processing and indexing algorithms but it also opens the way to several other possibilities such as rapid interface prototyping or retrieval algorithms benchmarking. In this respect, we discuss our developments in relation to wider contexts such as MPEG-7 and the TREC Video Track.

1. MOTIVATIONS

Video data processing has for long been of high interest for the development of compression and efficient transmission algorithms. In parallel, the domain of content-based multimedia retrieval has developed, initially from text retrieval, then for images and now addressing video content retrieval. Whereas in text and image retrieval the volume of data and associated access techniques are well under control, this is largely not the case for video collection management. Not only video data volume may rapidly grow complex and huge but it also requires efficient access techniques associated to the temporal aspect of the data.

Efforts in video content modeling such as MPEG-7 [9] are providing a base for the solution to the problem of handling large amount of multimedia data. While such a model is very well suited to represent a single multimedia document, it cannot be used efficiently for accessing, querying and managing a large collection of such documents due to its inherent complexity. Unfortunately, most video retrieval systems presented in state of the art literature [1, 4] do not explicitely discuss the way they address such management issues.

CVDB '04 Paris, France

In this paper, we detail the framework we have constructed for the management of video document collections in the context of our research in video content analysis. Rather than presenting a temporal document model alone, our ultimate goal is to develop content characterization and indexing algorithms for the management of large video collections. When addressing such problems, one rapidly faces the need for a favorable context on which to base these developments and also that permits rapid and objective evaluation of research findings. From an extensible multimedia document model, we have built a database framework comprising all needed reference information to raw video documents. Efficient access to the original document is ensured by a generic accessors called OVAL that we have embedded within several prototyping platforms. This way, we are combining the benefits of a classical DBMS for rapid access to indexed description data with the efficient random access capabilities of our platform.

In section 2, we are reviewing the model we propose for a multimedia document and associated description data. In section 3, we detail how we may create the required data associated with each video document. Section 4 presents access techniques that we have created to and from this data repository. In section 5, we show how our framework has been used to develop and evaluate novel video content characterization and indexing algorithms. Throughout the paper, we briefly discuss the relation between our developments and common efforts with in particular the TRECVid [16] Retrieval Evaluation challenge.

2. MODELING TEMPORAL DOCUMENT

The design of our framework is centered around the concept of temporal information. We consider that any part of our data store can be associated with a temporal stamp. The data itself may be located within either of the three layers depicted in figure 2. Namely, we follow a hierarchical scheme able to embed heterogeneous data such as an audio-visual (AV) stream (video) associated with meta-data and a set of key-frames (still pictures), themselves described by textual annotations. More formally, our scheme comprises:

- **Document Information** : global information about each document including meta-information and raw-data information. (Subsets of the *creation information*, *media information* and *usage information* of the MPEG-7 standard)

- **Document structure** : the temporal decomposition of

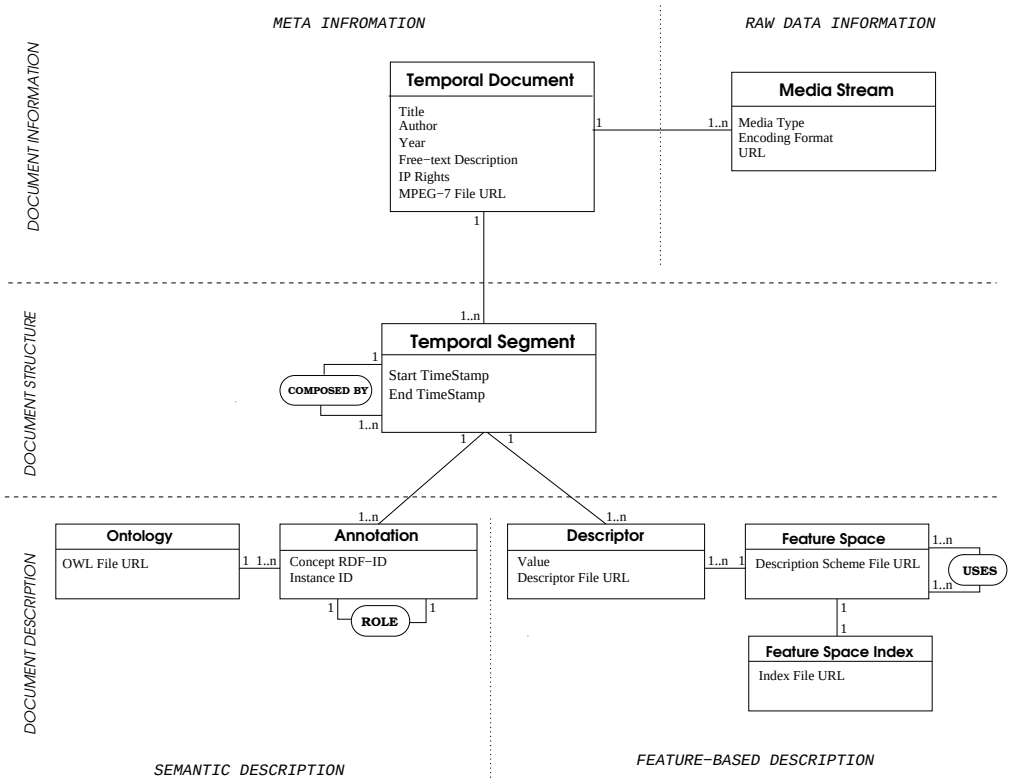

Figure 1: Conceptual Model of a Video corpus representation.

video documents that comes from the temporal segments covered by the description data.

- Document description : the set of description data that is either automatically extracted (*feature-based*) or entered manually by human operators (*semantic annotation*).

2.1 Temporal Structure

The key part of our model is the temporal decomposition of each document. We take the temporal dimension as a feature common to all modalities (visual, audio, textual) and exploit this property to create relations between pieces of information. By contrast, any other possible decomposition such as that proposed by the MPEG-7 standard would become an extra information attached to a particular information stream (eg, the spatial decomposition of a key-frame).

2.1.1 Temporal Segments

The notion of a *temporal segment* is therefore the central building block for our model. It is initially defined as a continuous temporal interval over the multimedia stream S:

$$I_a^b(S) = [a, b], \forall\ a, b \text{ s.t } \ 1 \leq \mathrm{a} \leq \mathrm{b} \leq \mathrm{T_S} \qquad (1)$$

where T_S is the total length of the stream. In the most general case, a temporal segment may also be an arbitrary composition of such intervals.

$$I_{(a_k)}^{(b_k)}(S) = \bigcup_k I_{a_k}^{b_k}, k = 1, \cdots, n \qquad (2)$$

Any temporal pattern may therefore be defined within our scheme. The aim is to create logical temporal entities with which to associate combined multimedia information. Since no absolute temporal reference may be used, the definition makes sense only in association to a particular document (as identified by its *document information*). The converse is also true. To be valid, any piece of information should come with a temporal reference. In particular, a complete document S is associated with $I_1^{T_S}(S)$ and any partition of S with a partition of that interval. Thus, our model readily copes concurrent temporal segmentations of a given document.

2.1.2 Temporal Synchronization

Multimedia documents are formed out of several synchronized information streams (corresponding to respective modalities). In our model, the central aspect relating all streams of a multimedia document is the notion of a temporal segment. In all cases, an information query should result at the very least in a pair [document ID, temporal segment]. This poses the question of how to keep streams synchronized for playback.

Since our framework allows a rapid and efficient access from raw data (see section 4.1), we will use original synchronized storages methods and playback whenever possible. In the simplest case of an audio visual document (video), we will simply store and access the original document and rely on multiplexing to preserve synchronization. In the case of a composite document, virtual documents will be composed and played using classical strategies. One such strategy is the use of SMIL, the Synchronized Multimedia Integration Language to compose virtual documents, playable with any

compatible player.

2.2 Description spaces

Temporal segments organize the data along the temporal dimension. We define a further classification of the information contained in the *document description* layer (the temporal information) into main categories. We define the *asserted description* as the description that is given from an external knowledge source and the *deduced description* as being a description inferred or computed from the multimedia stream itself. Typically, the asserted description may be provided by a human operator annotating the document in question and therefore be located at a rather high semantic level. The deduced description is computed automatically and corresponds to the document features extracted from the data itself. This distinction places us in a favorable context for the development and test of multimedia information processing algorithms. For example, deduced description will form an automated characterization that the asserted description may help in evaluating (see section 5 for an example).

In order to implement our data model, the distinction to consider is between *semantic description* and *feature-based description*, which corresponds to distinct and complementary storage modes.

2.2.1 Semantic description

Semantic description is integrated in the model through manual annotations. As free text annotation may provide a noisy description due to the lexical and cultural differences among annotator, the external knowledge is normalized by the use of an **ontology**. The semantic description therefore lists the set of instances of concepts (as defined by the ontology) that occur within a temporal segment. This scheme allows us to use generic multimedia annotation frameworks such as that given by the Semantic Web (see [7] for a more detailed proposition). As a complement, associations between instances may be created, according to their possible *roles*, as defined by the ontology. Note that our proposed model is directly able to represent different semantic descriptions, using different ontologies.

Clearly, tradeoffs are to be determined between the complexity of the ontology used and the level of description needed. An important factor to take into account is also the complexity of the annotation, strongly related to the size of the ontology at hand. In our research-oriented scheme however, the semantic description plays a crucial role. It provides a semantic organization of the content that may be used for high-level querying and browsing the collection, and for training or evaluation of classification or recognition algorithms.

2.2.2 Feature-based description

The main goal of our framework is to store, organize and create relations between automatically computed features. These are seen as a description deduced on a particular temporal segment. A feature-based description (or simply, a descriptor) of a multimedia content is defined in relation to a feature space. In the general case, a descriptor attached to a temporal segment corresponds to a set of points or a trajectory within that feature space. Further, as some descriptors may be computed from other descriptors (e.g. shape descriptor computed on a spatial segmentation), feature spaces may be related through a *uses* relationship. Here again, our model closely matches the underlying architecture of the feature extraction procedures used.

For the sake of simplicity, simple descriptors are represented by their values. In the most complex case, we use external files storing these values. In order to access such descriptors, an index may be constructed for the corresponding feature space. A feature space index is a file storing the accessing methods as long as the index data. For now, we have used complete distance matrices to index feature spaces, but for obvious computing reasons others indexing structures should be used. For example tree based index may be used, such as VP-Tree or M-Tree (see [3] for a review on indexing structures in metric spaces).

Our framework therefore provides an efficient way to store the output of multimedia stream content analysis algorithms for evaluation or comparison purposes. The co-existence of both levels of description within an unified repository makes it easy to define evaluation or supervised training procedures. Further, as a complement to the semantic description, the feature-based representation of the temporal segments opens the way to construct query and browsing mechanisms.

3. ENTERING THE DATA

We have mapped our model onto a database schema. Our database currently handles more than 60GB of video data coming from the two corpora gathered by the MPEG-7 and the TREC Video Retrieval Evaluation (2003) communities. This heterogeneous set of videos contains many genres, including sport, sitcom series, variety program, TV news and documentaries. It illustrates typical TV broadcast by the variety of its content and is widely used as a benchmark for video analysis and indexing tasks.

We have processed the raw documents, in order to extract low-level information about their temporal structure (shot detection), their activity content (camera displacement, regions of activity, event) and their global color distribution. The speech transcripts extracted by Automatic Speech Recognition (ASR) at LIMSI laboratory [5] and all data made available on the TRECVid data are also stored in the database.

These descriptors provide us various viewpoints on the raw documents according to their intrinsic audio-visual properties. In parallel with these automatic processes, we have manually annotated part of the collection, not only to enrich the description, but also to efficiently evaluate our algorithms.

3.1 Semantic annotations

Manual annotation of the documents relies on an ontology based on both the taxonomy presented in [14] and the lexicon of the TRECVid Collaborative Annotation Forum [8] (figure 2). This ontology is centered around the concept of a video shot. It is widely acknowledged that shots form essential semantic elements of a video stream. However, within our data model, shots are just a particular case of temporal segments. Thus, other ontologies may be used, based for example on the concept of *scene* (set of visually correlated shots) or *story* (set of semantically correlated shots). This ontology creates annotations that provide us sufficient information for easy access to our database content and corresponds well to the documents features we wish to charac-

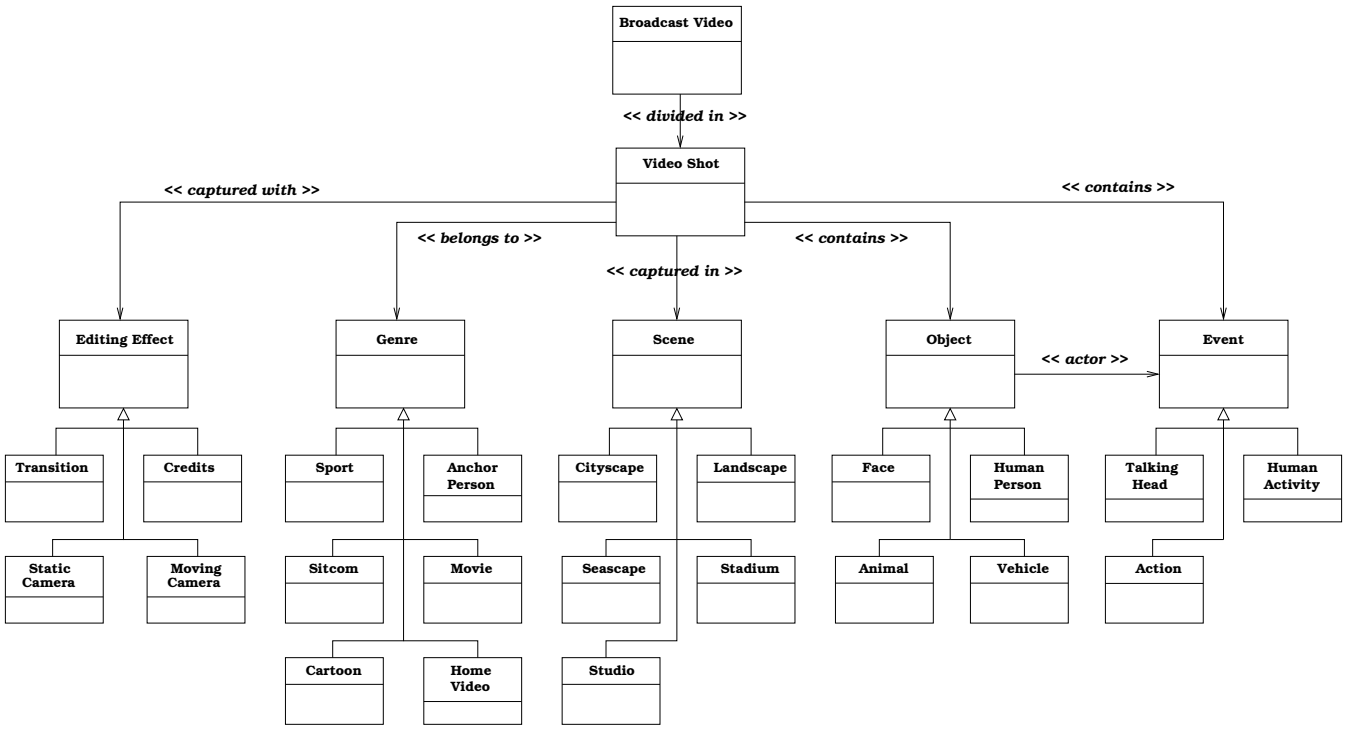

Figure 2: Ontology for semantic annotation of video documents.

terize automatically (eg, editing effects, human face, *etc*).

3.2 Data Processing

As already mentioned, our goal is to facilitate multimedia data processing and organize the storage and access of the data and its associated description. Here, we detail the raw data processing that leads to storing data descriptions within our framework. In section 5, we give the example of a complete application based on our framework.

3.2.1 Temporal partitioning

Since the temporal structure of multimedia documents is central to our framework, the first step we take is to achieve temporal segmentation of multimedia streams. This approach is compatible with the fact of considering a video shot as a temporal unit for subsequent processing.

An automatic algorithm for video temporal segmentation based on the minimization of an information-based criterion has been developed [6]. It offers very good detection performance for abrupt as well as smooth transitions between shots. The algorithm proceeds according to the following steps.

The video content is first abstracted by a color dissimilarity profile using the classic color histogram and the Jeffrey divergence as similarity measure. The complexity of further processing is then reduced by easily detecting non-ambiguous events such as hard transitions and sequences of still frames. An information-based segmentation is performed using a minimum message length criterion and a Dynamic Programming algorithm. This parameter-free algorithm uses information theoretic arguments to find the partitioning which agrees with the Occam's razor principle : the simplest model that explains data is the one to be preferred. The minimization process is fast by using the characteristics of video data like the presence of hard-cuts and redundancies to reduce the search for the solution. The computational complexity will depend on the video data but it is typically running in linear time.

At this stage, we obtain temporal segments whose definition is not guaranteed to match that of a shot. Since we see this level of decomposition as containing useful information, it is stored within our database. However, to remain compatible with other studies, a final merging algorithm uses statistical hypothesis testing to group together segments that are unlikely to form different shots.

As a first example of evaluation facilitated by our framework, table 1 presents the results of an experience using 70 videos of the TRECVid corpus and the evaluation framework of [15]. We used 35 hours of news programs and the ground truth provided by the TRECVid community. The

Performances	Our algorithm	Hardcut detection alone
Recall	92.4	67.6
Precision	80.2	78.8

Table 1: Performances of the shot boundaries detection algorithm

performances of the algorithm are comparable to the best results obtained by the participants of TRECVID 2003. The main advantage of our shot boundary detection algorithm is that we make a minimum number of assumptions about

the definition of a video transition. The algorithm will detect any kind of special effect without any particular modeling. From these results, we have built confidence in our algorithm and used its results for the processing of streams where ground-truth was not available.

3.2.2 Activity-based video decomposition

We now briefly review an example data processing that produces content descriptors stored in our database. The aim [12] is to decompose a given video shot (as defined above) into several spaces characterizing meaningful parts of its content

- **Capturing Effects** : trajectories of the affine parameters of the camera displacement
- **Capturing Environment** : descriptors of the background
- **Moving Objects** : salient regions of activity
- **Events** : trajectories of salient regions w.r.t background

Spatial salient points are extracted from each frame and matched between two successive frames. The global affine motion model (*Camera Displacement*) is estimated from the set of points trajectories. Salient regions of activity are extracted and tracked along the stream using the background model and the feature distribution of the points. As an example of extra information created from raw data and stored within our database in relation to the original data, figure 3 illustrates how a video shot is represented by the plots of the affine parameters of the trajectories of the camera displacements, the mosaic of the scene (represented in the system by the MPEG-7 *Scalable Color Descriptors (SCD)* and *Non-Homogenous Texture* descriptors) and the set of salient regions of activity (represented in the system by the MPEG-7 *SCD* and *Motion Trajectories* descriptors). The

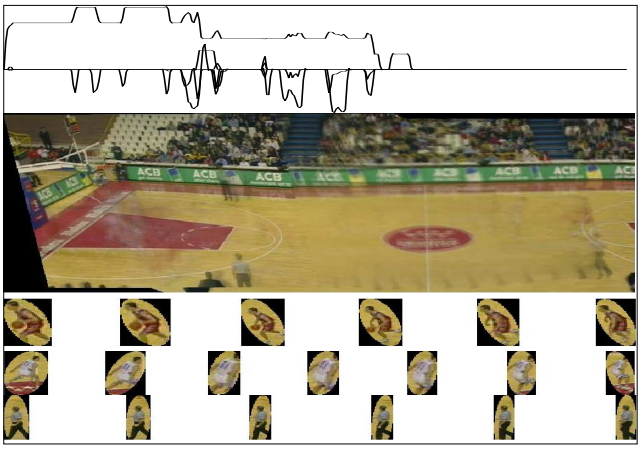

Figure 3: Representation of the visual content of part of a basketball game using salient features analysis: trajectories of the 6 affine parameters of the camera displacement, mosaic of the scene, 3 samples of salient regions of activity.

temporal granularity of such a description is not homogenous. Salient regions of activity could be defined on temporal segments that are subparts of a shot. Here again, the temporal structure of our data model allows to describe documents at different temporal granularity and to potentially combine description of temporal segments to a coarser temporal granularity.

3.2.3 Event-based feature space

We present a final example of processing made possible by our framework and illustrating its capabilities to facilitate event-based collection-wide access to multimedia information [2].

In an un-supervised context, we apply nonlinear temporal modeling of wavelet-based motion features directly estimated from the image sequence within a shot. Using SVM-regression, this nonlinear model is able to learn the behavior of the motion descriptors along the temporal dimension and to capture useful information about the dynamic content of the shot. An inter-shot similarity measure based on this activity characterization is then applied to documents within our repository. The similarity measure is defined as a quadratic error between models. We are therefore able to compute similarity matrix at the collection level that we store within our repository. In section 5, we show how to construct and evaluate a complete application based on these computations.

4. EXPLORING THE DATABASE

We now have a data repository that stores structured temporal audio-visual data enriched with low-level and semantic data. Basic access is given by the DBMS[1]. The underlying model opens access to data using a document reference and a given temporal segment within it. From there, any information related to that temporal segment may be extracted. Subsequent links with other temporal segments may be defined, based on the various notions of similarity we have created.

4.1 Accessing data

Our storage scheme enables easy access to any part of a document from a reference to that document, along with a temporal segment. We have developed a suitable software framework, called OVAL (Object-based Video Access Library [11]) that permits random access of data on AV streams. Typically, OVAL offers a common API on AV streams so as to emancipate from the actual type of storage used for that particular stream (advantages of particular storage modes may however still be accessed, such as motion vector within an MPEG-2 stream). One advantage of OVAL over other data access libraries is that its abstraction enables generic VCR-like operations and also adds random access facility to data streams. For example, using OVAL, a key-frame in a video stream is retrieved online by the sequence of `open`, `goto` and `extract` operations, thus avoiding duplication of data that may become obsolete. OVAL includes index pre-computation and buffering facilities so as to make the use of these operations as efficient as possible.

OVAL is written in C++ and wrapped into a MATLAB MEX mechanism to allow for easy video and audio frame access within MATLAB. A Java JNI extension of OVAL is also proposed.

[1]The presented model is implemented using the open source *MaxDB* Database Management System : *http://www.mysql.com/products/maxdb/*

4.2 Querying documents

OVAL and our DBMS now form our base for query audio-visual data. From this setup, we have constructed a global access framework that makes transparent data access at various levels and from different modes.

Documents may first be queried explicitly by attributes known to be present within their associated description (eg comprised within the ontology used in the case of annotations). Attributes here may either be textual or by values of features. In that sense, a document fully matches or not the query. Such queries correspond to *key* searches and *range* searches in a description space.

Proximity search are also available. Such queries are based on similarity measures that derive from ongoing studies described in sections 3 and 5. Similarity queries are K-Nearest Neighbor queries or Top-K ranked queries that are performed using the indexing structure of the corresponding feature space. Such queries are very efficient in term of their response time, but at the cost of the computation of the index (recall we use distance matrices). As pointed out in section 2 other indexing structures may be used to limit the complexity of the index creation, but it will increase the response time of the queries.

Proximity searches are performed on a single feature space. However, we may compose complex queries by combining several feature spaces. There is essentially two possibilities for combining modalities for a query. Either the combination is done at the query level and one (unified) multi-modal query is sent to the server. This process refers to information *fusion* [13]. As we do not address yet the problem of designing an online similarity server able to process complex queries (e.g. embedding information fusion), this is not handled at the level of what we describe in this paper. One route to follow for such an embedding is to build on the experience gained with the development of the GIFT [18], where a vector model for measuring feature similarity enables their combination in a transparent way.

In the current implementation of our system, multi-modality is shifted at the client level. It is to the client to manage information fusion and to decompose a given query into unit queries whose results are the recombined. This strategy has already been successfully used in several query interfaces (see eg [17]).

5. EXAMPLE APPLICATIONS

We see at least two situations where having an efficient data access and storage framework is crucial for further developments. These are collection exploration and feature evaluation.

5.1 Collection-wide indexing and exploration

By definition, collection-level problems such as collection-wide document retrieval or browsing involve large collections of documents. In the case of video documents, the volume is multiplied by the temporal dimension of each document. There is therefore no simple way to handle this massive volume of data. Further, since we wish to enable access to these documents at various temporal levels (eg shot level or scene level), we need a unified way of accessing multimedia data.

Using our framework, we were able to easily construct a basic video retrieval application by simply creating a query interface to our database. Semantic descriptions are queried using basic text search facilities. Visual similarity is handled by classical content-based image retrieval [18] and temporal document similarity uses pre-computed similarity matrices [2]. One important feature to highlight here is our capacity of retaining long-term information. Whenever accessing the data stored, one may visually detect description errors and correct them so as to form a better base for subsequent processing. To this end, we have further developed an interface that enables the modification of annotations while browsing, thus facilitating the input of asserted data.

While the above access technique relies on queries and thus results into a set of documents answering that query, we introduce the need for a management made at the collection level. This involves the consideration of inter-document relationships and their mapping onto a visual space. We refer to this mapping as *Collection guiding* [10] since this analysis process will facilitate the comprehension of the collection as a whole and possibly highlight its structure such as density and disparity. Typically, each item in a multimedia collection is considered as a point in its corresponding feature space. A graph is then created, based on given inter-point relationships (the most intuitive being similarity). Using discrete optimization techniques, we characterize several underlying structures of this data set within this space such as Minimum Spanning Tree (MST) and Minimum Covering Set (MCS). By its optimality at various levels, the MST readily provides a global structure whose aspect and composition carries information on the global properties of the collection. Further, it may serve as a marker structure to preserve when mapped onto a reduced space in view of visualization. The MCS (or k-median set) may be used to assess collection disparity and perform adaptive sampling to show the underlying essentials of the collection (see [10] for details). Figure 4 shows a view of a collection obtained by mapping feature points onto the 2D space.

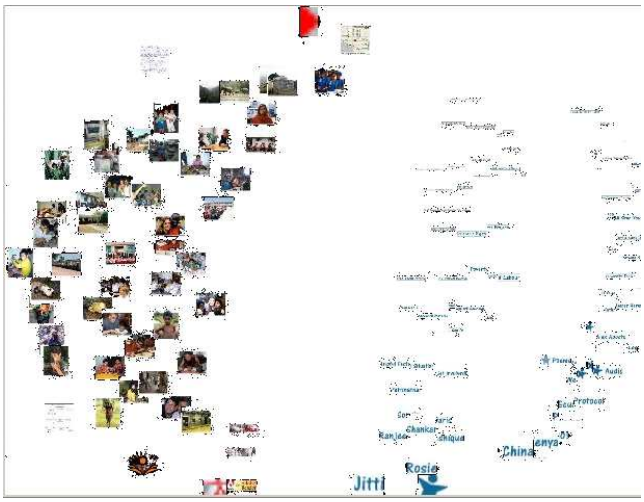

Figure 4: Mapping of a multimedia collection created by the Collection Guide. Here, we are interested in the collection structure emerging from this representation

5.2 Feature evaluation

In the specific case of video, the collection guiding strategy may apply at various levels, from complete documents to frames via the definitions of shots and scenes. Here again,

we benefit from the abstract storage of temporal segments that may transparently handle any of these specifications. Finally, we have proved our framework to be essential for objective systematic evaluation of our algorithms.

We have mentioned in section 3.2.1 the evaluation of our temporal partitioning procedure against external ground-truth. We have also evaluated the above video retrieval application by testing its ability to discriminate events within videos. Using the ontology defined in 3.1, we have defined three types of generic event classes:

- **Action** corresponds to high activity events, such as sport and dance sequences.

- **Human activity** corresponds to events representing human or crowd walking or doing large gestures.

- **Talking head** corresponds to close-up view on talking people, such as anchor scenes in news, dialog scenes in sitcom.

More than 800 video shots have been manually annotated by one of these three labels, or the label *null* when shots do not contain any of these three events (30% of the documents). Each document was then used as base for a query by similarity over the whole set of documents. The quantitative evaluation of the method is given by Precision-Recall graphs. Figure 5 displays average Precision and Recall computed on all video shots for each event label. Horizontal lines in the graphs represent the statistical mean value of Precision when documents are randomly selected (which is equal to the percentage of labels in the database). The fact that P-R curves are above these lines means that the retrieval operation performs better than a random selection. We can observe that for the three events, P-R curves are largely above the "random case" which validate the ability of the similarity measure to sort documents according to their dynamic content.

6. CONCLUSION

We are advocating the use of an advanced data storage and retrieval framework for the development and evaluation of multimedia processing algorithms. We have based the development of our framework around the temporal properties of the data to be stored. Within our data model, raw data, annotations and extracted features coexist and may even overlap along the temporal dimension. Although not explicitly using any standard, we remain fully compatible with alternative description schemes such as MPEG-7 while not being constrained by their syntax or structure.

We have presented several applications that we have based on our framework. We believe that the use of such a framework is unavoidable for the development of video indexing and retrieval applications. We further showed that the very same framework may also serve for the evaluation. Duality between development and evaluation is made evident using an incremental annotation scheme whereby ground-truth is incrementally built for subsequent processing or objective systematic evaluation. Further developments will address the test and extension of our models to handle richer multimedia data. We are also ready to accommodate and process new data coming from forthcoming TRECVid 2004.

7. ACKNOWLEDGMENTS

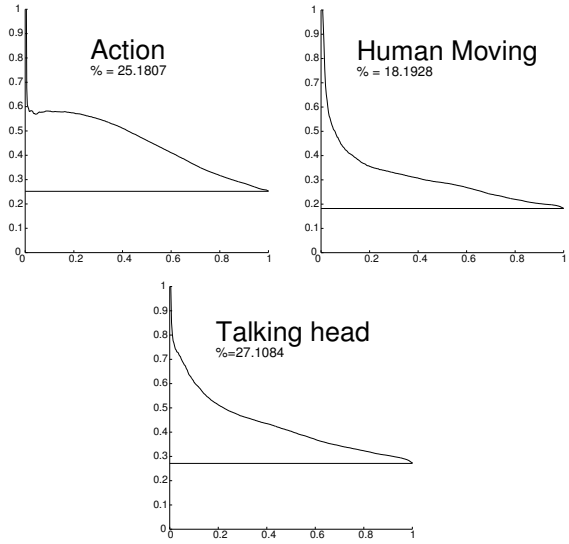

Figure 5: Precision-Recall graph for a) Action events, b) Human moving events and c) Talking head events. Horizontal lines represent the percentage of each label in the database (numerical values are given in titles). The sum is not equal to 100% because of the non-annotated shots.Precision-Recall graph for three events.

This work is funded by EU-IST project M4 (www.m4project.org) and the Swiss NCCR IM2 (Interactive Multimodal Information Management).

8. REFERENCES

[1] A. Amir, M. Berg, S.-F. Chang, W. Hsu, G. Iyengar, C.-Y. Lin, M. Naphade, A. Natsev, C. Neti, H. Nock, J. R. Smith, B. Tseng, Y. Wu, and D. Zhang. IBM Research TRECVID-2003 Video Retrieval System. In *Proceedings of the TRECVID 2003 Workshop*, 2003.

[2] E. Bruno and S. Marchand-Maillet. Nonlinear temporal modeling for motion-based video overviewing. In *Proceedings of the European Conference on Content-based Multimedia Indexing, CBMI'03*, September 2003.

[3] E. Chávez, G. Navarro, R. Baeza-Yates, and J. Marroquin. Searching in metric spaces. *ACM Computing Surveys*, 33(3):273–321, Sept. 2001.

[4] G. Gaughan, A. F. Smeaton, C. Gurrin, H. Lee, and K. McDonald. Design, implementation and testing of an interactive video retrieval system. In *Proceedings of the 5th ACM SIGMM international workshop on Multimedia information retrieval*, pages 23–30. ACM Press, 2003.

[5] J. Gauvain, L. Lamel, and G. Adda. The limsi broadcast news transcription system. *Speech Communication*, 37(1-2):89–108, 2002.

[6] B. Janvier, E. Bruno, S. Marchand-Maillet, and T. Pun. Information-theoretic framework for the joint temporal partioning and representation of video data. In *Proceedings of the European Conference on Content-based Multimedia Indexing, CBMI'03*,

September 2003.

[7] C. Jelmini and S. Marchand-Maillet. DEVA: an extensible ontology-based annotation model for visual document collections. In R. Schettini and S. S. Eds, editors, *Proceedings of SPIE Photonics West, Electronic Imaging 2002, Internet Imaging IV*, Santa Clara, CA, USA, 2003.

[8] C.-Y. Lin, B. L. Tseng, and J. R. Smith. Video collaborative annotation forum: Establishing ground-truth labels on large multimedia datasets. In *Proceedings of the TRECVID 2003 Workshop*, 2003.

[9] B. Manjunath, P. Salembier, and T. Sikora, editors. *Introduction to MPEG-7: Multimedia Content Description Language*. Wiley, 2001.

[10] S. Marchand-Maillet. Collection guiding. Technical Report 03.06, Viper Group – CVML – University of Geneva, 2004.

[11] N. Moënne-Loccoz. OVAL: an object-based video access library to facilitate the developement of content-based video retrieval systems. Technical report, Viper group - University of Geneva, 2004.

[12] N. Moënne-Loccoz, E. Bruno, and S. Marchand-Maillet. Video content representation as salient regions of activity. In *Proceedings of the International Conference on Image and Video Retrieval, CIVR'04*, July 2004.

[13] V. Pavlovic, G. Berry, and T. S. Huang. Fusion of audio/visual information for human-computer interaction. In *Workshop on Perceptual User Interfaces (PUI)*, pages 69–71, 1997.

[14] M. Roach, J. Mason, L.-Q. Xu, and F. Stentiford. Recent trends in video analysis : a taxonomy of video classification problems. In *Proceedings of the International Conference on Internet and Multimedia Systems and Applications, IASTED*, August 2002.

[15] R. Ruiloba, P. Joly, S. Marchand-Maillet, and G. Quenot. Towards a standard protocol for the evaluation of video-to-shots segmentation algorithms. In *International Workshop in Content-Based Multimedia Indexing (CBMI)*, 1999.

[16] A. F. Smeaton, W. Kraaij, and P. Over. TRECVID 2003 - An Introduction. In *Proceedings of the TRECVID 2003 Workshop*, 2003.

[17] J. R. Smith, A. Jaimes, C.-Y. Lin, M. Naphade, A. Natsev, and B. Tseng. Interactive search fusion methods for video database retrieval. In *IEEE International Conference on Image Processing (ICIP)*, 2003.

[18] D. M. Squire, H. Müller, W. Müller, S. Marchand-Maillet, and T. Pun. *Design and Evaluation of a Content-based Image Retrieval System*, chapter 7, pages 125–151. Idea Group Publishing, 2001.