An Integrated Framework for the Management of Video Collection

Nicolas Moënne-Loccoz*, Bruno Janvier,
Stéphane Marchand-Maillet, and Eric Bruno

Viper group - CVMlab - University of Geneva,
Computer Science Department - 24 rue du General 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 developments 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 are providing a base for the solution to the problem of handling large amount of temporal data. MPEG-7 formalizes the definition of temporal points associated with any relevant information along the multimedia stream. This way, every local or global descriptor may be associated with a temporal reference point or interval within a multimedia document.

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

S. Bengio and H. Bourlard (Eds.): MLMI 2004, LNCS 3361, pp. 101–110, 2005.

In this paper, we present the framework we have constructed for the management of our video document collection. 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 permit rapid and objective evaluation of research findings. ¿From an extensible multimedia document model, we have built a database framework comprising all needed reference information of raw video documents. Efficient access to the original is ensured by a generic accessor 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 review the model we propose for a multimedia documents 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 Retrieval Evaluation challenge.

2 Temporal Document Modeling

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 metadata and a set of keyframes (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 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 Modeling the Temporal Dimension

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,

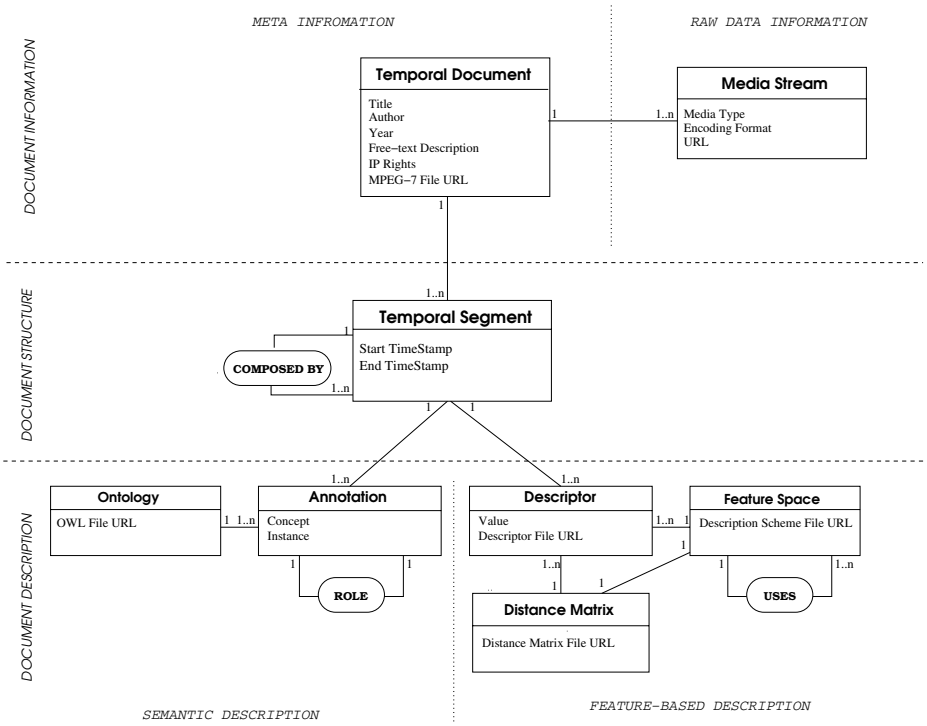

Fig. 1. Conceptual Model of a Video corpus representation

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 keyframe).

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:

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

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.

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

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.

2.2 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.

Semantic Description. We normalize external knowledge 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 the given temporal segment. This scheme allows us to use generic multimedia annotation frameworks such as that given by the Semantic Web (see [4] 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.

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. Descriptors may also be seen as pre-computations over several temporal segments. In this case, for fast access, distance matrices are also stored, representing the structure of the feature space for a given document collection.

Our framework therefore provides an efficient way to store 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 [2] and all data made available on the TRECVid data are also stored in the database. These descriptors, associated to pre-computed distance matrix, 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

The database is enriched with semantic descriptions of the documents. These annotations rely on an ontology based on the lexicon of the TRECVid Collaborative Annotation Forum [5] (figure 2). This ontology is centered around the concept of a video shot. It is widely acknowledge that shots form essential semantic elements of a video stream. However, within our data model, shots are a particular case of temporal segments. Thus, others ontologies may be used based for example on the concept of *scene* (set of visually correlated shots) or

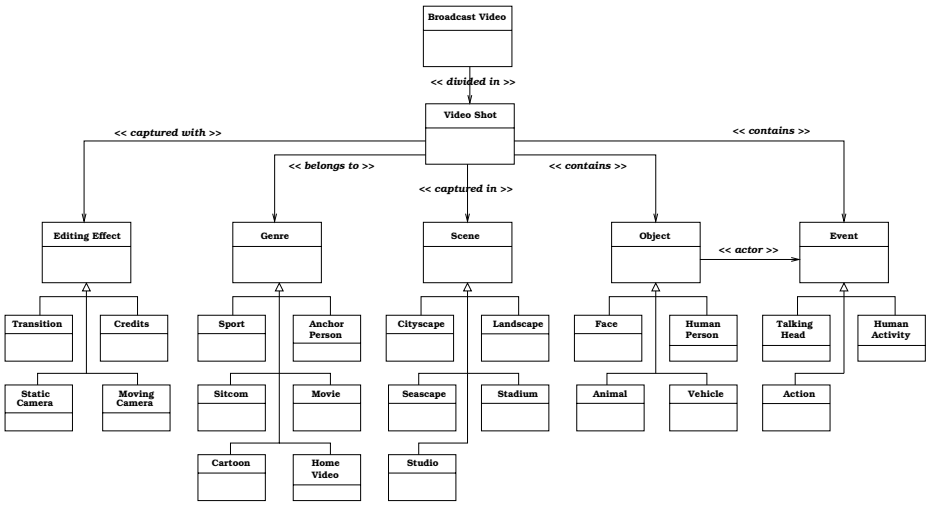

Fig. 2. Ontology for semantic annotation of video documents

story (set of semantically correlated shots). This simple 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 characterize 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 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.

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 [3]. 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 classic color histogram and the Jeffrey divergence. 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

Table 1. Performances of the shot boundaries detection algorithm

Performances	Our algorithm	Hardcut detection alone
Recall	86.2	67.6
Precision	77.2	78.8

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.

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 [7]. We used 35 hours of news programs and the ground truth provided by the TRECVid community.

¿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.

Event-Based Feature Space. We present a final example of processing exploiting our framework and illustrating its capabilities to facilitate event-based collection-wide access to multimedia information [1].

In an unsupervised context, we apply nonlinear temporal modeling of wavelet-based motion features directly estimated from the image sequence within a shot. Based on 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 our DBMS. The underlying model opens access to data using a document reference and a given temporal point within it. From there, any information related to that temporal point may be extracted. Subsequent links with other temporal points may be defined, based on the various notions of similarity we have created.

4.1 Assessing 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 [6]) 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 a MPEG-2 stream). One advantage of OVAL on 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 keyframe 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.

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.

Similarity queries are also available. Similarity derives from ongoing studies described in sections 3 and 5. Currently, we may combine several access techniques including textual matching of descriptions (directly supported by the DBMS), content-based visual similarity to similarity based on motion pattern, handled by external prototypes of indexing engines.

5 Example Applications

Finally, we see, as an appliclation, a situation where having an efficient data access and storage framework is crucial for objective systematic evaluation of our algorithms.

We have mentioned in section 3.2 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:

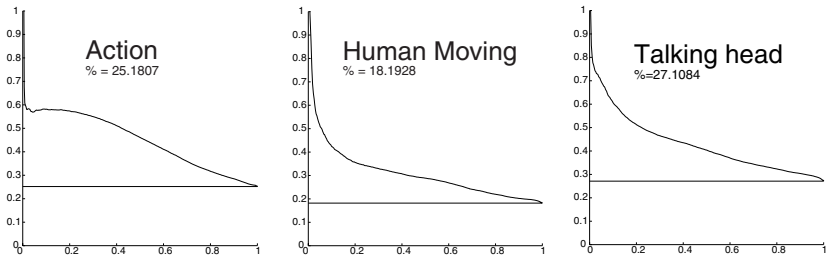

Fig. 3. Precision-Recall graph for three events. Horizontal lines represent the percentage of each label in the database (numerical values are given in titles)

- **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. Resulting Precision-Recall graphs averaged over each of the three above classes are presented in figure 3.

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 an application 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.

References

1. Eric Bruno and Stéphane 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.
2. J.L. Gauvain, L. Lamel, and G. Adda. The limsi broadcast news transcription system. *Speech Communication*, 37(1-2):89–108, 2002.
3. Bruno Janvier, Eric Bruno, Stéphane Marchand-Maillet, and Thierry 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.
4. Carlo Jelmini and Stéphane Marchand-Maillet. DEVA: an extensible ontology-based annotation model for visual document collections. In R. Schettini and S. Santini Eds, editors, *Proceedings of SPIE Photonics West, Electronic Imaging 2002, Internet Imaging IV*, Santa Clara, CA, USA, 2003.
5. Ching-Yung Lin, Belle L. Tseng, and John R. Smith. Video collaborative annotation forum: Establishing ground-truth labels on large multimedia datasets. In *Proceedings of the TRECVID 2003 Workshop*, 2003.
6. Nicolas 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.
7. R. Ruiloba, P. Joly, S. Marchand-Maillet, and Georges Quenot. Towards a standard protocol for the evaluation of video-to-shots segmentation algorithms. In *International Workshop in Content-Based Multimedia Indexing (CBMI)*, 1999.