In *International Workshop on Description Logics* (DL '98), Trento, 1998, Franconi, E., De Giacomo, G., MacGregor, R.M., Nutt, W., Welty, C.A Eds.

Using Description Logics for Indexing Audiovisual Documents

Jean Carrive[1], François Pachet[2], Rémi Ronfard[1]

[1]Institut National de l'Audiovisuel (INA)
[2]SONY CSL-Paris & Lip6 (Paris 6)

{jean, remi}@ina.fr, pachet@csl.sony.fr

Extended Abstract

We address the problem of indexing broadcast audiovisual documents (such as films, news). Starting from a collection of so-called shots, we aim at building automatically high level descriptions of subsets of this collection, that can be used for annotating, indexing and accessing the document. We propose to represent documents and high level descriptions with the framework of description logics, enriched with temporal relations. We first define the problem as a classification problem. We then propose an algorithm to automatically classify sub-sequences of shots, based on a bottom-up construction of descriptions using the rule mechanism of the CLASSIC system.

1. Introduction

This study takes place in the field of audiovisual documents indexing. By audiovisual documents, we mean essentially video or film programs. Indexing is understood here in a very general sense, as the operation which allows whole or part of a document to be the result of a request. In practice, that goes from simple methods such as associating a few keywords with the whole document to much more sophisticated ones, such as describing deeply a document, for example with conceptual graphs [1].

1.1 Temporal documents

A specific characteristic of all audiovisual documents is their temporal dimension. This temporal dimension has two sides: the multi-layered aspect of documents, and their structural aspect.

1.1.1 Documents are multi-layered

The various information concerning audiovisual documents may be organized in a multi-layered structure. Each layer contains temporal information concerning a particular aspect of the document. The most basic layer is the *shot layer*, which is basically the segmentation of the audiovisual data into a set of discrete temporal objects. Shots are usually considered as the smallest syntactic units of film language [2]. Shots may be defined as what is filmed during one run of the camera, without edit. However, a most interesting information for indexing and understanding documents is the transition between shots. A *cut* means a brutal transition between two shots: the last image of the first shot is immediately followed by the first image of the second one. It can therefore be represented as a temporal objet with no duration (an event). Gradual transitions, such as a *fade* (*in* or *out*), *dissolve*, etc, are represented as standard temporal objects, intertwined between two shot objects.

Other typical layers are: the dialog layer for representing dialogs between characters. Yet another layer may be used for representing appearances of characters on screen, and so forth (see Figure 1).

The information contained in each layer is typically derived from analysis algorithms. It is important to notice that some extraction algorithms may be executed *a priori*, such as the detection of shot transitions [3]. Other algorithms need contextual information, such as face detection. In the first case, some algorithms may be too costly to be executed on the whole document. This is the case for example for text extraction, where the document as to be firstly segmented in time and space.

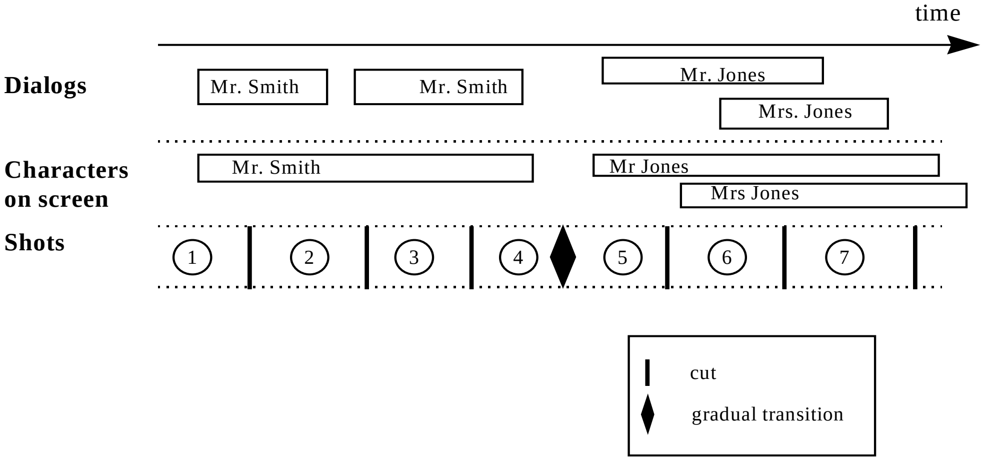

Figure 1: example of multi-layered description

1.1.2 Structured documents

The second aspect of audiovisual documents is their hierarchical nature. In most cases, a document may be split into *successive* sequences which are in turn split into *shots* (see Figure 2).

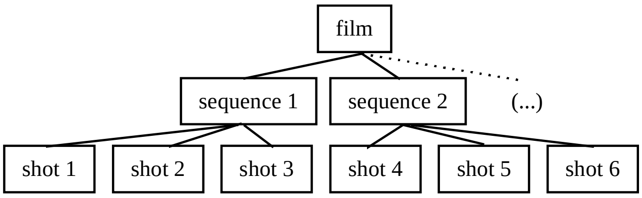

Figure 2: hierarchical structure

Usually, TV researchers know more about documents, and can classify them into *document types*: for instance, the newscast of CNN at 8pm, specific sitcoms, variety shows, western movies, etc. Within one specific document type, documents share several characteristics, such as film sets, news readers, or the organization of shots or sequences over time. For example, the temporal structure of some particular news programs could be described *in general* as an alternation of *in sets* sequences and *report* sequences, where in sets sequences are composed for example of still shots (no camera motion) of the news reader (say Mr. Smith) separated by cuts, with the logo of the channel in the top right corner of the screen.

1.1.3 The taxonomy of film events

We claim that there exists a taxonomy for some elements of film, and that some (partial) formalization of this taxonomy may be given. A simplified taxonomy of traditional transition (punctuation) effects between two shots (from [4]) is shown Figure 3.

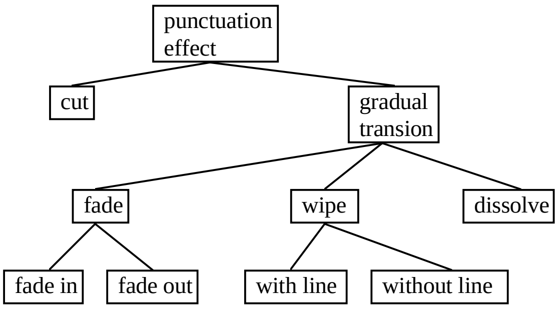

Figure 3: taxonomy of punctuation effects

In *International Workshop on Description Logics* (DL '98), Trento, 1998, Franconi, E., De Giacomo, G., MacGregor, R.M., Nutt, W., Welty, C.A Eds.

If we consider only subsets of the taxonomy where role fillers can be automatically extracted from the signal (as color histograms) it is important to notice that some parts of this taxonomy are entirely made up with *primitive* concepts, i.e. that no classification process can assign to an instance a position in the hierarchy. In the example above, an algorithm may detect a shot transition as a *gradual transition*, and then refine the description to *fade in*. However, concerning for example camera motions, an algorithm is unlikely to classify a camera motion as *pan*, and then refine the description to *pan right*.

1.1.4 Classification of temporal segments

The relative disposition of elements within a layer may convey some signification. For instance, in some contexts, a gradual transition between two shots may signify a transition between two sequences.

[5] proposes general rules to group shots into sequences. One of these rules specifies that a gradual transition surrounded on each side by at least three cuts is likely to be a sequence transition. In the example of Figure 1, according to this rule, there is a sequence limit between shots 4 and 5. This shot transition may be classified as « sequence limit ».

In order to describe meaningfully the document, elements of different layers are to be taken account. For example, in the example of the Figure 1, a TV researcher might be interested by a shot where Mr. Smith is speaking during the whole shot, say, in sight of using this shot for a documentary about Mr. Smith. Shot number 3 meets these requirements. Thus, this shot may be classified under a concept « shot of Mr. Smith talking». It is important to note that these concepts may be considered as *specializations* of the basic concepts of the taxonomy of film events (section 1.1.3).

2. Using DL for analysis

Structural analysis is the process that yield the temporal structure of the document, from the initial audiovisual document and the various layers containing additional information on the document. This process is of course made easier when the document type is known *a priori* (which is most often the case), since the document type is associated with generic temporal structures, as seen in the preceding section.

We claim that the structural analysis of an audiovisual document may be seen as a classification process. This involves 1) representing generic temporal structures for documents types (see section 1.1.2), and 2) devising an algorithm to aggregate primitive film events and classify them according to these generic temporal structures.

2.1 Description Logics and temporal classification

Description logics are knowledge representation languages; they allow to represent knowledge in a structured way by separating definitions of *concepts* (terminological representation system or *Tbox*) from description of *individuals* (assertional representation system or *Abox*). Concepts are sets of individuals and *roles* represent binary relations between individuals. Concepts and roles descriptions are organized in hierarchies with the *subsumption* relation [6]. Several description logics systems are available; the ideas proposed in this paper are implemented with the CLASSIC system [7].

Various works have been conducted to classify temporal structures, mainly in the field of plan recognition. [8] and [9] propose to extend the notion of subsumption to plans, while [10] propose a formal language for reasoning about time and action. However, these results are difficult to apply in our context because we have only partial descriptions of the general temporal structure of type documents. This temporal structure is made of with several « plans » which only apply locally on a portion of the document. Thus, it is not known *a priori* what segments may intervene in what plan. Moreover, some segments may participate in several plans at once.

2.2 Film events as concepts

It is natural to represent film events in our taxonomy as concepts in the sense of description logics. For instance, the concepts of a shot of the news reader in the example illustrated by the Figure 1 may represented by a CLASSIC concept as follows:

```
(define-concept ‘READER-SHOT ‘(and
    SHOT
    (exactly 1 character-on-screen)
    (fills camera-motion still)
    (at-most 1 character-speaking)
    (all transition-to-next-shot CUT)
    (all transition-to-previous-shot CUT)
    ...))
```

As we can see, the CLASSIC concept represents only a part of the information : the temporal structure is not expressed. For instance, the concept definition above doesn’t specify temporal relations between *character-on-screen* and *character-speaking*. This is essentially due to the limitations of the description logics formalism. We propose to represent this structure using the rule-based inference mechanism of CLASSIC.

2.3 Grouping temporal units

In order to have some temporal segment classified, as a shot of the news reader, illustrated in section 1.1.4, one must first express this segment as a combination of some other segments. We express this combination as a grouping rule.

2.3.1 Structure expressed as grouping rules

The structure of the document is expressed as grouping rules which aggregate temporal forms of low level into temporal forms of higher levels. We have identified two main categories of grouping rules. In the first category, rules aggregate two instances of two distinct concepts into one instance of a concept of a higher level. In the second category, rules aggregate N instances of the same concept into one instance of a concept of a higher level.

In order to define these rules, we need to define the concept TEMPORAL, which represent temporal intervals. This concept is defined as follows :

```
(define-concept ‘TEMPORAL ‘(and
    (exactly 1 begin)
    (all begin integer)
    (exactly 1 end)
    (all end integer)
    (< begin end)))
```

The general form of rules of the first category is:

C1 R C2 à G *(1)*

with:

C1, C2 temporal concepts (inheriting from TEMPORAL)

R : temporal relation

G : concept inheriting from TWO-TEMP-GRP, group of two temporal instances, defined by:

```
(define-concept ‘TWO-TEMP-GRP ‘(and
    TEMPORAL
    (exactly 1 first-temporal)
    (all first-temporal TEMPORAL)
    (exactly 1 second-temporal)
    (all first-temporal TEMPORAL)))
```

The general form of rules of the second category is:

C R à G *(2)*

with:

C : temporal concept

R : temporal relation

G : concept inheriting from TEMPORAL-SEQ, sequences of temporal instances, defined by:

```
(define-concept ‘TEMPORAL-SEQ ‘(and
    TEMPORAL
```

In *International Workshop on Description Logics* (DL '98), Trento, 1998, Franconi, E., De Giacomo, G., MacGregor, R.M., Nutt, W., Welty, C.A Eds.

```
            (at-least 2 element)
            (all element TEMPORAL)))
```

Some sub-categories have to be defined for each of the two main categories, in order to specify how to instantiate the resulting concept *G*. There are several ways to precise the role fillers of the resulting concept: that can be the common values of one very role of the premise concepts, the value of one particular role of one particular premise concept, the most specific generalization of values of one particular role, etc.

2.3.2 The need for a temporal logic

The rules expressed above mention temporal constraints between temporal intervals: Mr. Smith talking *during* Mr. Smith on screen, for example. In order to represent these temporal constraints, we need a formalism to represent temporal relations. The choice of this formalism is important ; it must ensure a good compromise between expressiveness and tractability.

XXXDire pourquoi Allen de base ne marche pas…

In our case, we propose to choose the temporal model presented by [11] – *Pointizable Interval Algebra* – which is based on the interval algebra of Allen [12]. In this model, disjunctions of Allen basic relations are transformed into conjunctions of constraints on the bounds of these intervals. Only a subset of Allen interval algebra may be expressed in this way. For example, the temporal relation

$$A \{\textit{before} \vee \textit{meets} \vee \textit{overlaps}\} B$$

is transformed to:

$$\text{begin}(A) < \text{begin}(B)$$
$$\text{end}(A) < \text{end}(B)$$

but the relation:

$$A \{\textit{before} \vee \textit{after}\} B$$

has no equivalent.

2.3.3 A strategy for grouping segments

The principle of grouping presented here is based on the rule-based mechanism offered by CLASSIC. In CLASSIC, it is possible to associate rules to a concept definition. These rules are triggered each time an instance of this concept is created. XXXTu dis 2 fois la même phrase…

When an instance *a* of the concept A is created, the rule is triggered. In our case, for the rule A R B à G, the rule consist in searching all instances b_i of the concept B such that *bi* R *a*. *a* and b_i are then grouped together in a new instance of G.

3. Conclusion

We have defined the problem of indexing audiovisual documents, and have shown that it involves classifying temporal structures using multi-layered information. The classification problem is made even more complex by the fact that only subsets of the initial temporal structures must be classified, and that these subsets are not known a priori.

We have presented a method in the framework of description logics, to retrieve temporal structure of audiovisual documents by expressing temporal structure types as grouping rules of temporal concepts. The implementation with CLASSIC of the ideas expressed above is in progress, and shows the validity of our approach on small examples. Some work is left to do, particularly to identify useful sub-types of grouping rules, and ensuring that the approach may be used on real world, full size documents.

4. Bibliography

1. Simonnot, B., *Modélisation multi-agents d'un système de recherche d'information multimédia à forte composante vidéo*, . 1996, Université Henri Poincaré - Nancy I. p. 259.
2. Katz, S.D., *Film Directing Shot by Shot*. 1991: Michael Wiese Production.
3. Yeo, B.-L., Liu, B., *Rapid Scene Analysis on Compressed Video.* IEEE Transactions on Circuits and Systems for Video Technology, 1995. **5**(6): p. 533-544.

In *International Workshop on Description Logics* (DL ‘98), Trento, 1998, Franconi, E., De Giacomo, G., MacGregor, R.M., Nutt, W., Welty, C.A Eds.

4. Arijon, D., *Grammar of film language*. 1976: Focal Press, London & Boston.
5. Aigrain, P., Joly, P., Longueville, V., *Medium Knowledge-Based Macro-Segmentation of Video into Sequences*, in *Intelligent Multimedia Information Retrieval*, A.P.M. Press, Editor. 1997.
6. Nebel, B., *Reasoning and Revision in Hybrid Representation Systems.* LNAI, 1990. **422**.
7. Borgida, A., Brachman, R.J., McGuiness, D.L., Resnick, L.A. *CLASSIC: A Structural Data Model for Objects*. in *ACM SIGMOD Int. Conf. on Management of Data*. 1989.
8. Devanbu, P.T., Litman, D., *Taxonomic Plan Reasoning.* Artificial Intelligence, 1996. **84**: p. 1-35.
9. Weida, R., Litman, D. *Terminological Reasoning with Constraint Networks and an Application to Plan Recognition*. in *Proceedings of the Third International Conference on Principles f Knowledge Representation and Reasoning (KR'92)*. 1992. Cambridge, Massachussetts.
10. Artale, A., Franconi, E. *A Computational Account for Description Logic of Time and Action*. in *Proc of the 4th International Conference on Principles in Knowledge Representation and Reasoning (KR94).* 1994.
11. van Beek, P. Reasoning about qualitative temporal information. in Proceedings of AAAI'90. 1990.
12. Allen, J.F., *Towards a general theory of action and time.* Artificial Intelligence, 1984. **23**(2): p. 123-154.