A Framework for Multi-Agent Multimedia Indexing

Bernard Merialdo
Multimedia Communications Department
Institut Eurecom
BP 193, 06904 Sophia-Antipolis, France
merialdo@eurecom.fr
March 31st, 1995

ABSTRACT

In this paper, we propose a framework for the usage of a multi-agent architecture to perform automatic indexing of multimedia documents. Because of the diversity of the data types that are involved in multimedia documents, multimedia indexing requires the combination of very different technical approaches: rule-based, syntax-based, statistical, probabilistic, computational etc... We propose a framework where we can combine simple agents (generally based on a single media) to build a system which is able to perform the automatic detection of complex multimedia events. We address three issues in this paper. First, we give a general overview of the organization of a multi-agent system, and define the various ways by which simple agents can be combined. Then we propose a probabilistic formulation that allows to combine the results of the computation of these agents, and in particular, is able to handle the uncertainties that arise in the recognition processes. Finally, we give some indications on how distributed search algorithms could be used to efficiently handle the computations that are involved by these agents.

Throughout the paper, we will consider a typical example of the kind of indexing that we would like to be able to achieve, to serve as an illustration for the various aspects that we are proposing.

1. INTRODUCTION

Because of the evolution of the technology, the usage of multimedia documents is getting more and more popular. Such documents, combining video, sound, graphics and text, are very interesting for the user interface because they convey information in a very natural way for the user. However, when considered as data objects, multimedia documents have the drawback that they support only a limited set of operations. The basic operation of a multimedia document is presentation, or play-back, where the document is read from storage and visualized using a number of output devices such as monitor, speaker, video display etc... To create a document, some editing functions are available, such as cut-and-

paste, copy, special effects, fast forward and reverse. To transmit, store or retrieve a document, compression or normalization functions are also used. These functions have in common that they only use the syntactic structure of the document, that is the way the information is stored, and not the semantic content, the kind of information that is contained in the document. This shows the need for more complex processing capabilities that allows to analyse a multimedia document to detect the occurrence of complex events, which can then be used for an intelligent processing such as filtering or retrieval.

We follow a typical example of the kind of indexing that we would like to achieve: consider the video recording of a two-hour meeting where you would like to locate a certain part of the discussion. You can play it, which will take you two hours. Using the fast forward and reverse functions, you can wander inside the document and visualize excerpts to go faster. But there currently is no way to directly access a position where, for example, "*Mister X was talking about topic Y*".

To perform the recognition of such events, we can use video, audio, and maybe text analysis techniques that have been studied for a long time to provide elementary pieces of information. For example word spotting on the audio component can be used to detect if a given word has been pronounced, face recognition on the video component can be used to detect if a given person is present etc... We then have to combine these elementary pieces to evaluate if they give consistent evidence for the occurrence of the complex event. In the following, we propose a framework where these tasks are performed by simple recognition agents, and these agents are combined to build a complex multi-agent recognition system. We will give some examples of simple agents and define the various ways by which they can be combined.

2. STATE OF THE ART

The problem of Multimedia Retrieval covers various fields of investigation, of which we feel that three are particularly important:

1. relevant information has to be extracted from individual flows of data, such as sound and video streams. While such information can be provided manually by the user, the most promising approach is to pattern recognition techniques, such as speech recognition [Wilcox et al., 1992], image analysis and vision [Smoliar and Zhang, 1994],
2. basic informations from various flows have to be combined to build Multimedia

indices [Gabbe et al., 1994]. This, in turn, covers several aspects:

- a conceptual modelization for the possible ways of combination, which relates to the definition of the architecture of multi-agent systems [Levis, 1993]
- an evaluation of the validity of the results that are provided, taking into account the uncertainty of pattern recognition procedures, which relates to the degree of belief in inference networks [Jensen et al., 92] [Pearl, 1988],
- an efficient implementation of the computations that are required by these combinations, which relates closely to distributed problem solving [Corkill and Lesser, 1987],

3. the interface should allow the user to search through these indices, and should be able to present the corresponding data efficiently [Teodioso and Bender, 1993] [Arman et al., 1994].

The focus of this paper is on point 2, with an emphasis on the first two aspects. The content of this paper reflects ongoing work.

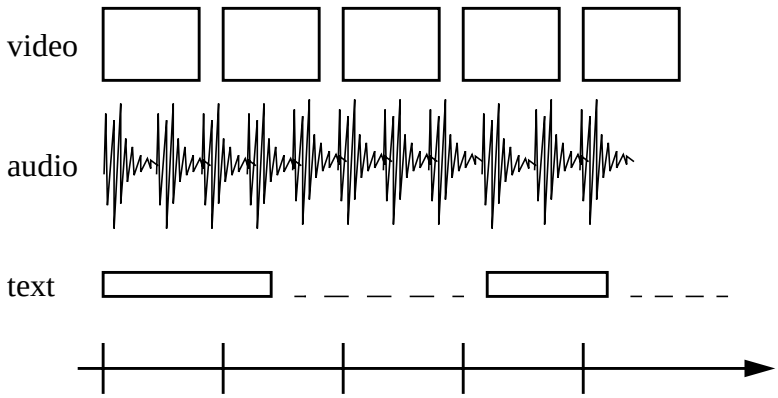
3. MULTIMEDIA DOCUMENTS STRUCTURE

The structure of a multimedia document can be quite complex if we consider the possibility of hypermedia links, spatial and temporal synchronization mechanisms, various data types etc... For our purpose, we initially consider a simplified model where a multimedia document consists of one or several flows of information that are aligned along the same timeline. A simple example is the recording of a conference, or a TV program. The flows of information can be:

- video, taken by one or several cameras,
- audio, captured by one or several microphone,
- text, captions or notes, either prerecorded or typed on the fly, or manually added later to the recording,
- other kind of events, such as computer activity, external captors etc...

There may be several flows of a given type, for example if we record a conference using multiple cameras, or multiple microphones, or if a TV program contains

captions in different languages etc...



Simplified Multimedia Document Structure

4. MULTI-AGENT RECOGNITION

Following are examples of simple agents that are relevant to our example:

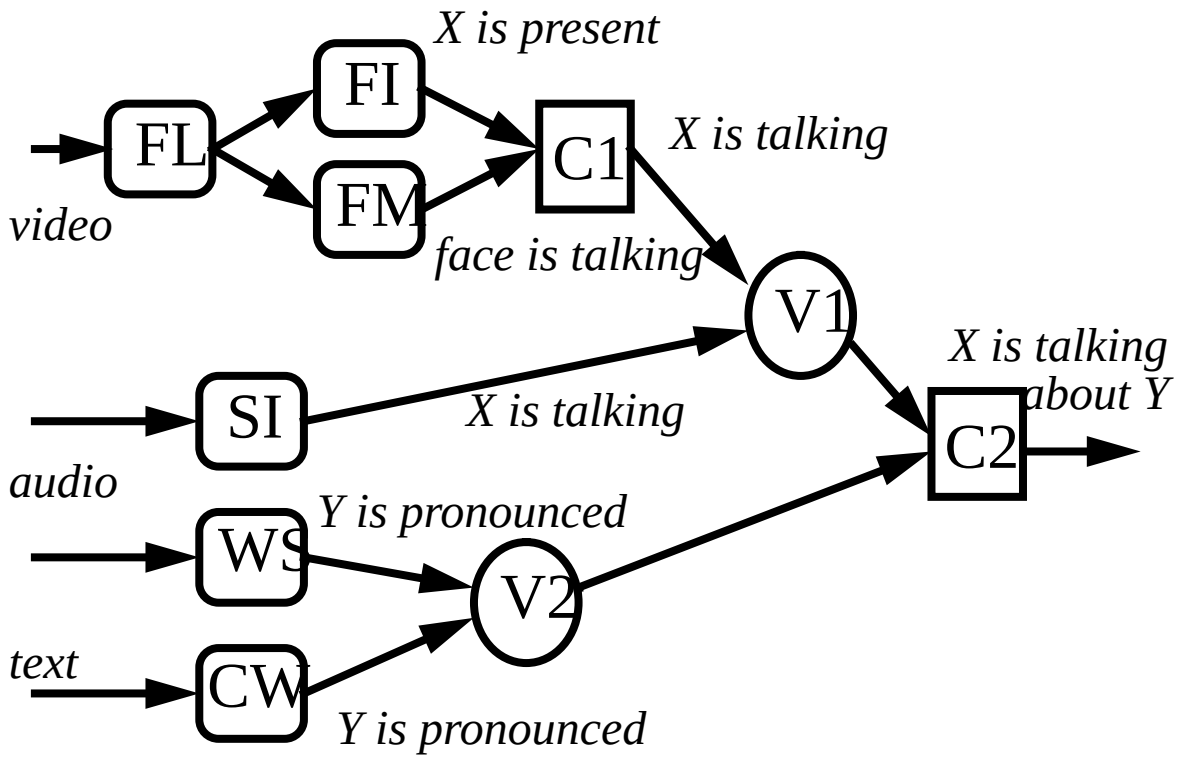
- Face Location (FL) agent: takes a video as input and produces a list of regions in the image that are classified as faces of people,
- Face Identification (FI) agent: takes as input an extract of the video which is supposed to be a human face and identifies the person (against a given database),
- Face Movement (FM) agent: takes as input an extract of the video which is supposed to be a human face and checks if the person is actually speaking (by analyzing lip movement),
- Speaker Identification (SI) agent: takes as input an audio signal and determines the instants where a given speaker is talking,
- Word Spotting (WS) agent: takes as input an audio signal and determines the instants where certain words (from a predefined vocabulary) are pronounced,
- Caption Word (CW) agent: takes as input a text stream (words with timecodes) and determines the instants where certain words appear.

This list should provide a reasonable understanding of what a simple agent may be. From these examples, we can construct a multi-agent system that is able to perform the

recognition of the complex event:

"Person X is talking about topic Y".

Assuming that X is a definite person, and that topic Y is illustrated by a keyword, a possible structure of such a system is indicated in the figure below:



The video is analysed by the Face Location agent to detect face, which are identified by the Face Identification agent. At the same time, the Face Movement agent is able to detect that the person is talking. The combination of the identity of the person and its lip movement gives an indication of the event *"X is talking"* based on the video source. In parallel, the Speaker Identification agent analyses the audio signal to detect when *"X is talking"*. The video and audio informations are then combined to formulate an evidence for *"X is talking"*. Similar operations occur with the other agents. This example show that there are three different ways by which simple agents can be combined together:

- *succession*: when the output of one or several agents are used as input to another agent. This is the case when the Face Location agent provides information to the Face Identification agent.
- *validation*: when several agents provide informations about the same event, and that their advices have to be combined in a single hypothesis. This is the case when we

want to combine the output of the video and audio identification agents.

- *composition*: when the hypotheses formulated by various agents have to be combined to form a more complex hypothesis. This is the case when we combine the information *"X is talking"* with the information *"word Y is being pronounced"*.

The last two operations are performed using validation and combination agents that perform the necessary combinations. We should note that the design of such a multi-agent may be done in several ways. For example, we could build a much simpler system by relying on the audio only for the identification of the person. Also, a design might make some assumptions about the adequacy between the events that are detected and the information we are looking for. This happens for example when we assume that a person is speaking when some lip movement is detected on its face, or when a topic Y is characterized by the occurrence of a given keyword. Note also that we are using instances of generic agents. For example, the Word Spotting agent that appears in the previous system is an instance which has just the simple task of detecting a specific word Y, while other instances might be used in other situations.

We can summarize our approach by the following aspects:

- a catalog a simple agents,
- three ways of combining these agents into a complex system, with the introduction of validation and combination agents,
- a decomposition of complex events into combinations of simple events, and the corresponding design a multi-agent system.

5. PROBABILISTIC FRAMEWORK

The previous structure can be operated using AND-OR mechanisms for the different types of combinations of hypotheses. However, the tasks that are considered here involve uncertainty. Pattern recognition techniques are not 100% reliable, so that any decision made by an agent is only made with a certain degree of confidence. Uncertainty also occur when we transpose an event into the detection of certain features, for example, when we replace the search of a certain topic by the detection of a given keyword (the topic could actually be discussed without this keyword appearing, or this keyword might appear

in other contexts).

To handle uncertainty, we propose to use a probabilistic approach to weight the various hypotheses that are proposed by different agents and their combinations. This approach is an extension of the Causal Probabilistic Networks introduced in [Andersen et al., 1989]. Probabilities in this context have some nice characteristics. They provide a sound theoretical framework. Probabilistic models have already been applied successfully to various domains of pattern recognition, image, speech or even natural language. Finally, the probabilistic laws naturally match the possible combinations of agents that we have defined previously. This is because the structure of the multi-agent system arises from the decomposition of complex hypotheses, in a similar fashion that complex probabilities are decomposed into simple ones.

Let us assume that the simple agents will provide an output which is basically a probability distribution of the detection of certain events. For example, the Speaker Identification agent will provide a probability for the two hypotheses:

X is currently talking

not (X is currently talking)

Of course, this evaluation should exist for each time t, so that the output is a continuous curve indicating p_t("X is talking")

In general, an agent B will receive as input the outputs $a_1 a_2 \ldots a_n$ of other agents $A_1 A_2 \ldots A_n$. In a deterministic approach, it would produce an output b. In the probabilistic approach, it will produce a probability distribution over the set of possible outputs b, $p(b|a_1 a_2 \ldots a_n)$. The global probability for the occurrence of the outputs of B is then:

$$p(b|a_1 a_2 \ldots a_n) \cdot p(a_1 a_2 \ldots a_n)$$

While the second term is in general difficult to compute because the outputs of $A_1 A_2 \ldots A_n$ can be correlated, we believe that in most cases, it can be simplified under reasonable assumptions. Following are some examples of such computations:

- when agents are combined in succession, their probabilities can be multiplied. For

example,

$$p(\text{X is present}) = p(\text{face is identifed as X}) \cdot p(\text{face is located})$$

- when agents are combined by a validation agent, a possibility is to interpolate the probability distributions of the various inputs. For example:

$$p(\text{X is talking}) = \lambda \cdot p_{audio}(\text{X is talking}) + (1-\lambda) \cdot p_{video}(\text{X is talking})$$

- when several agents are combined by a composition agent, and we can assume that the inputs are independent, the output distribution is the product of the input distributions:

$$p(\text{X is talking about Y}) = p(\text{X is talking}) \cdot p(\text{Y is pronounced})$$

6. MULTI-AGENT SEARCH

The last topic that we want to mention is the way the computation can be performed efficiently in such a multi-agent system. Because pattern recognition techniques are often computer-intensive, and the number of hypotheses that they could generate is enormously high, special care should be taken to organize the computation between the various agents. If the computation is performed in a brute force manner, each agent will be run over the whole data, their results later combined by their successors, etc... until all outputs from all agents have been computed. This obviously to a large amount of computation. A distributed problem solving mechanism would activate agents in an adequate order on adequate portions of data, to build progressively the desired results, and would take into account various pieces of knowledge to prune non-useful computations. Although we have not yet implemented such a mechanism that would better organize the activity of different agents, here are some preliminary considerations on certain factors that such an approach should take into account:

- the cost of using an agent: if an agent is using specific hardware that is not shared with others, it might be unimportant to optimize its utilization, since its hardware resources would not be utilized by others. On the contrary, if several agents are using common computing resources, they will have to compete and it will be important to rationalize the order in which they are activated, specially if certain agents require substantially larger computation than others,
- the possible correlation between certain agents: for example, if a person is talking,

there should be some speech recorded on the audio track, thus the Face Movement agent could be only applied to periods where a Sound Detection agent found some audio activity (as SD probably requires less computation that FM),

- the probability of partial hypotheses: those with higher probability should be extended first, and a pruning mechanism should be able to discard hypotheses when their probability falls below a predefined threshold,
- the possibility of fast rough estimations: pattern matching techniques often allow for fast matching procedures that provide a rough estimate of the exact probability of an hypothesis, with low computanional cost. When the estimate is low when compared to other hypotheses, the real matching procedure will never be called for this hypothesis.

We envision that these constraints could be formalized inside a scheduler that would analyze the current set of hypotheses and their probabilities to decide which agents to activate next. However we have not yet started any experimental work in this area.

7. CONCLUSION

We have presented a framework for realizing multimedia indexing using a multi-agent approach, where simple agents are combined using three possible combination paradigms. We have shown that probabilities can be used to effectively evaluate the hypotheses that are generated by such combinations. Finally, we have given some requirements for an efficient implementation of search algorithms in such multi-agents systems.

8. REFERENCES

Andersen, S. K., Olesen, K. G., Jensen, F. V., and Jensen, F. (1989). Hugin: a shell for building bayesian belief universes for expert systems. *Proceedings of the 11th IJCAI*.

Arman, F., Depommier, R., Hsu, A., and Chiu, M.-Y. (1994). Content-based browsing of video sequences. *ACM Multimedia Conference*, pages 97–103.

Corkill, D. D. and Lesser, V. R. (1987). Distributed problem solving. In *Encyclopedia of Artificial Intelligence*. John Wiley.

Gabbe, J. D., Ginsberg, A., and Robinson, B. S. (1994). Towards intelligent recognition of multimedia episodes in real-time applications. *ACM Multimedia Conference*, pages 227–235.

Jensen, F. V., Christensen, H. I., and Nielsen, J. (92). Bayesian methods for interpretation and control in multi-agent vision systems. *Applications of Artificial Intelligence X:*

Machine Vision and Robotics, SPIE Proceedings Series, 1708.

Levis, A. H. (1993). *Modelling and design of distributed Intelligence Systems*. Kluwer Academic Publishers.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufman.

Smoliar, S. W. and Zhang, H. (1994). Content-based video indexing and retrieval. *IEEE Multimedia*, pages 62–72.

Teodioso, L. and Bender, W. (1993). Salient video stills: Content and context preserved. *ACM Multimedia 93*, pages 39–46.

Wilcox, L., Smith, I., and Bush, M. (1992). Wordspotting for voice editing and audio indexing. *Proceedings of Computer Human Interaction*, pages 655–656.