INTERNATIONALCONFERENCE ON MULTIMEDIA FOR HUMANITIES

# Knowledge-based Retrieval of Multimedia documents

**Hiranmay Ghosh and Santanu Chaudhury**
Centre for Development of Telematics, Delhi, India
Indian Institute of Technology, Delhi, India

*Abstract: In this paper, we describe a knowledge based approach for retrieval of multimedia documents. We establish the similarity between the query and the documents at a conceptual level. The principle of abduction is used to determine the degree of relevance of the documents for a query. The documents that can explain the expected media patterns for the concepts presented in a query are selected for retrieval. The framework supports collection of evidences in favour of the concepts from multiple sources, namely, content analysis of the document components in multiple media forms as well as metadata associated with the documents, and for combining them. The knowledge-base required for the retrieval is partitioned into well defined functional units and is assumed to be distributed. A multi-agent distributed problem solving architecture is proposed to support the logical framework.*

## Introduction

Multimedia and internet technologies offer a great potential for on-line dissemination of information. Multimedia documents can combine textual descriptions with audio and visual portrayals and can be much more expressive than the conventional documents comprising text, and possibly pictures. The internet and web technologies allows ready access of the on-line documents across the globe. However, with the exponential growth of such collections in the recent years, retrieval of the relevant documents has become

increasingly difficult. Specifically, audio and video documents cannot be easily browsed unlike conventional documents and they put more cognitive load on the users. Thus, an efficient retrieval system is essential for such collections.In this paper, we describe a distributed knowledge-based retrieval method for multimedia documents.

The problem of retrieval deals with matching documents to queries based on some similarity measures. A user specifies some conceptual entities, representing his field of interest, in a query. The task of the retrieval software is to establish a similarity measure between the query and the documents available in one or more repositories and to present the pertinent documents in a ranked order.

With falling cost of computations, retrieval methods based on analysis of document contents and associated annotations have proved to be a viable alternative to traditional index database search. The latter suffers from the shortcoming that its success depends on the suitability of the pre-conceived keywords associated with the documents in the context of a particular query. Processing of textual annotations has the inherent advantage of computational efficiency over analysis of non-textual document contents. Examples of such retrieval methods are presented in [4,7] for image collections and [10] for video collections. These methods fail to exploit rich information contents of the non-textual media forms.

A multimedia retrieval system should analyse the information content in all media forms to compute the similarity measures between the documents and the queries. The simplest content-based retrieval systems attempt to find the similarity between the query and the media objects in terms of the media primitives, for example, the shapes and sizes in image documents [5] and sample tunes in audio (music) clips [2]. They

are indeed extensions of full-text search to non-textual media forms. These methods do not produce satisfactory results, since the similarity measures based on media primitives usually do not reflect the conceptual similarity between a query and the documents. More sophisticated pattern recognition routines attempt semantic abstraction of media primitives to recognise complex media objects, for example features of a human face in image [22], spoken words [8], handwritten words [11], human behaviour in video [14] and news items in video [17].

The retrieval systems described above are confined to a single media form. A multimedia document repository, in general, contains documents with multiple media components. Thus, there is a need to provide a unified interface that can abstract over the heterogeneity of the media forms and the associated search methods. Architectures [13,18] have been proposed to combine non-textual content analysis with annotation based search in order to benefit from the effectiveness of the former and the efficiency of the latter. The HERMES [3] retrieval system provides a framework for reasoning over multiple media forms as well as metadata. The Informedia project [20] combines search on audio and video components for retrieval from a large video library. However, the retrieval is still based on a user specified combination of media-specific objects and not on concepts that is abstracted over media forms.

We use a knowledge-based approach to develop a retrieval engine than can reason with concepts abstracted over multiple media forms. In contrast to the usual deductive approach (e.g. [21,12]), we model retrieval as a problem of abduction. Abduction is a reasoning model for constructing an appropriate explanation for a set of observed patterns. Concepts are abstract entities and cannot be directly observed in the documents. However, they manifest themselves as some observable patterns in the media components of the documents. In our model of reasoning, we identify the documents that can account for the expected media patterns as the candidates for retrieval. Our framework is general enough to combine data from content analysis of multiple media forms as well as meta-data, such as annotations, that may be associated with the documents.

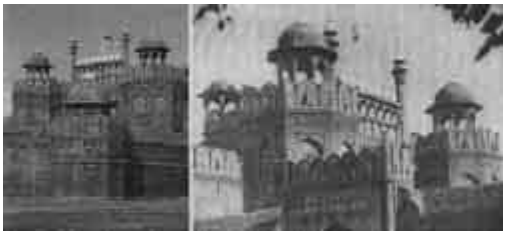
The problem of retrieval from a distributed and heterogeneous database can conveniently be modelled as a problem of distributed problem solving. A distributed approach not only provides a convenient means to handle multiple users and information sources distributed over a network, but can also utilise a distributed knowledge-base. Two distinct distributed architecture for retrieval have been presented in [15,1]. In the former, the documents themselves are encapsulated in active agents with epistemic knowledge about themselves. The agents bear a hierarchical relationships with each other representing the structure of collections. Retrieval takes place as interaction between these agents. In the second architecture, two classes of independent agents interface with the users and the document collections respectively. Some agents perform the role of mediators between the two. Our approach is similar to the latter, but is quite distinct in its use of the knowledge base. We use multiple forms of knowledge at different levels of abstraction. We partition the knowledge-base in well-defined functional units and encapsulate each in an autonomous intelligent agent. Another interesting aspect of our architecture is the use of mobile agents which can move across the network and perform requisite functions. This method of partitioning the required knowledge-base and realising it through a multi-agent architecture is a distinguishing feature of our approach.

We organise the rest of the paper as follows. Section 2 describes the framework of reasoning used in our retrieval system. It is elaborated with examples, the query and the knowledge representation schemes. Section 3 presents a distributed architecture for supporting the reasoning framework, which is followed by a summary and conclusion.

## 2. Framework of reasoning

As a motivating example, let us consider a query requesting documents pertaining to monuments of the

Mughal period. Figures 1 -3 depict some image documents which are the targets of this query. We shall use this example to illustrate the framework of reasoning as we develop it in this section.



Each of the pictures comprise some elementary patterns, such as straight lines and curves, with some spatial relationships with each other. These patterns can be detected through appropriate algorithms and their relationships can be established. But, each of the images is quite distinct from the others in terms of these elementary media features, though all of them pertain to a common theme. As a result, it is not possible to capture the commonality across these documents with pattern recognition routines that compares these documents with a sample in terms of elementary media features.

We note that a concept, such as Mughal monument, is an abstract entity and cannot be observed in a document. But, it produces certain perceptible effects in one or more media forms comprising the documents [9]. An observation model for a concept is a representation of a concept in terms of a set of perceptible media objects and their interrelationships. It can be derived from the subject domain knowledge about the concept. We can identify a concept in a document by recognising one or more of its observation models.

For example, in order to identify the monuments of the Mughal era, we resort to the knowledge about the architectural styles of that time. We note that the architecture of the Mughal period has the general characteristics of Islamic architecture and is marked by the presence of arches, bulbous domes, minarettes and Arabic inscriptions. It is distinguished from the Islamic architecture of other parts of the world by a flat facade, relatively flat domes, use of red sandstone or white marble as the building material, etc. We can associate a structure with Mughal period by observing one or more of such media objects representing these architectural characteristics.

The media objects that characterises a concept might undergo various media dependent transformations in a particular media component constituting a document. Such transformations include projective and colour transformations in images, use of synonyms or alternative spellings in text segments, speaker dependencies in speech, truncation because of obstructions, and so on. It is therefore necessary to define recognition functions that describe the media objects as a combination of media patterns invariant over such transformations in order to recognise the media objects in the documents. Knowledge about the media forms allows us to formulate the recognition functions.

In the example, while all the pictures contain the key components of Mughal architecture, namely, the facade, the arches, the domes and the minarettes, their shapes and relative positions appear to be quite distinct because of differences in dimensions, focus and viewing position. In particular, though the two pictures in fig. 1 depict the same structure, they appear to be quite dissimilar because of the transformations suffered by virtue of a change in the viewing angle. Similar differences are visible in the other pictures (fig. 2 and fig. 3) because of differences in viewing distances and photographic techniques. The color of the structures, that can be used to recognise the building material, might also undergo transformations because of the differences in daylight and ambient lighting conditions. In order to identify the media objects in spite of such differences, we need recognition functions which describe these media objects (and their combinations) in terms of media patterns and their relationships, that are invariant over any local transformations.

The documents could as well be identified from the associated textual descriptions (if available) and their association with other conceptual entities, for example, the names of the monuments or the emperors who have been responsible for their construction. As in image, there can be local transformations in the textual parts of the document because of use of synonyms and variations in spelling, for example the text object Red Fort may as well be described as Lal Quila in some of the documents.

Thus, a concept may be identified through evidences from multiple sources. The first step in processing a query is to deduce the media objects and hence the various media patterns and their interrelationships that might provide evidence for the specified concepts. We use conceptual knowledge as well as media knowledge to perform this step. Then, we schedule pattern recognition routines on the document components to seek evidences for the expected patterns. In the current example, we search the documents for the projective and colour invariants that characterises Mughal architecture and textual patterns that characterises other concepts of the Mughal period. The documents which provide best explanations for these media patterns are selected and are presented in a ranked order.

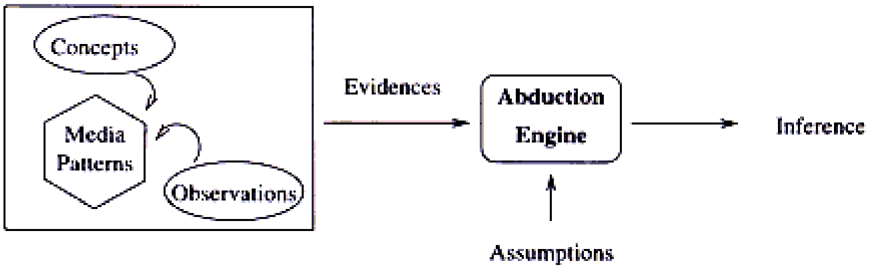
We note that the very nature of a concept defies principle of logical deduction for their identification. In the first place, concepts are abstract and are defined only through the observation models as examples. Further, a combination of media patterns may not be unique to a media object. For example, the shape of a minarette is very similar to that of a chimney, and there is a possibility of confusion between the two. Thus, the observations provide, at best, plausible explanations to the media objects caused by a concept. We follow an abductive model of reasoning in our retrieval system. We believe a document to pertain to a concept when the observed media-patterns corroborate the hypothesis.

Such reasoning often requires some underlying assumptions. For example, a long and narrow vertical structure can be interpreted to be a minarette and not a chimney, assuming that the image pertains to a construction of historic importance. We use classification knowledge to derive such assumptions. The documents are pre-classified to some broad, possibly overlapping subject classes with the help of associated meta-data. Classification knowledge helps in optimising processing of a query also. It suggests a set of documents or repositories, where the requested concepts are most likely to appear. Without this knowledge, we would need to schedule the pattern recognition routines on every document in the repositories. Such a proposition could be prohibitively costly for any non-trivial collection. The classification cannot, however, be the sole basis for retrieval since the perspective of a pre-classification may not account for all types of queries that a system might require to process.

We can formalise the principle of abduction used for retrieval as follows. If a concept C implies a set of media patterns { M }, then C can be abduced from { M } in presence of some set of assumptions { A }, i.e.

if C -> {M}, then {M} ^ {A} => C.

Figure 4 explains the basic model of an abductive engine used for retrieval.



## 2.1 Query representation

A query is the specification by the user for retrieval. We have developed a markup language for representing a query. For the purpose of brevity, we do not present the exact syntax of the query language in this paper, but bring out its essential structure.

A query consists of three parts, namely, domain of search, condition for retrieval and restrictions on the retrieval, if any. The first and the third parts specify the boundaries of the network for search, restrictions on media forms to be retrieved and any restrictions on real time or computing resources to be used. The second part of a query represents the condition for retrieval. The condition is an expression involving concept-descriptors and concept-modifiers. A concept-descriptor is a single term or a phrase describing a concept, such as monument. A concept modifier operates on a concept descriptor resulting in a new concept. We envisage three types of concept modifiers.

1. **Specialisation**: A concept descriptor is modified by another to represent a specialisation of the former (as in object oriented paradigm), for example, monument.Mughal which means a specialisation of monuments that belong to the Mughal era.

2. **Property association**: A property has a name and is associated with a value. The value may be symbolic, numeric or a vector. It may also be a pointer to a sample, for example, shape = <SampleShape>. A property can be associated to a concept, for example, monument <parameter> height = tall <parameter> represents a tall monument.

3. **Branching**: Two concept descriptors may be associated using a connective to indicate a new concept, distinct from either of the two, for example, <of> philosophy, Buddha </of> represents philosophy of (propagated by) Buddha, i.e. Buddhism.

A concept descriptor with one or more concept modifiers is treated as a single complex concept, which is processed by the conceptual knowledge as a whole. The condition part of the query is a logical combination of simple or complex concepts. The documents which can provide plausible explanations for the concepts specified in the query constitute the response from the system. The quality of retrieval is determined by the constraints imposed, besides the correctness of the algorithms used.

## 2.2 Conceptual knowledge

We use conceptual knowledge to derive the observation models for the concepual entities in the queries. An observation model describes a concept in terms of perceptible media objects and their interrelationships. The various observation models are associated to a concept through an expertise about the concepts acquired over many actual observations. The concepts encountered by an average person during a lifetime is vast, and the amount of commonsense reasoning performed by him defies today's limit of technology. We define our retrieval problem within the boundaries of a definite subject area. We assume a closed world model where there is a finite number of concepts.

The concepts in a subject domain are not isolated, but are related to each other. These relationships can be utilised for similarity based reasoning rule-and-similarity. One interesting relationship is subsumption, which means that a concept is completely or partly covered by another. The subsumption relationship can be used for top-down property inheritance, unless explicitly overruled. For example, the concept Mughal architecture is subsumed in the more general class of Islamic architecture and inherits the observation models of the latter. Subsumption relationships can also be used for retrieval using examples. Retrieval of documents pertaining to the subsumed items implies retrieval of documents pertaining to the subsuming concept as well. For example, a retrieval request for Mughal architecture, can be also satisfied by retrieving documents that are similar to, say, the Red Fort and Jama Masjid.

Another relationship that can be profitably be utilised for retrieval is association of concepts. For example, we might retrieve samples of Mughal architecture using the names of the emperors who have been responsible for the construction of such monuments. Any document that is associated to a Mughal emperor and contains the structure of a monument may be considered to contain a sample of Mughal architecture.

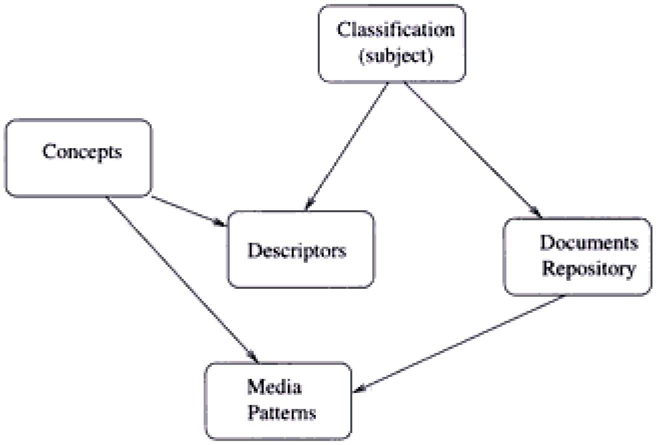
## 2.3 Another example: railway steam engine

We illustrate the framework of reasoning with another example. Let us assume a query specifying a concept engine.steam, i.e. a steam engine. From the conceptual knowledge, we find that the steam engine is a special type of a railway engine. All railway engines have the common characteristics of possessing some wheels and running over a pair of rails (with rare exceptions like the mono-rail and magnetic levitation trains). A steam engine inherits them. Over and above, a steam engine has the special characteristics of a specific shape (comprising a boiler, a chimney, etc.), its huff-and-puff sound, its cloud of smoke and whistle. A steam engine can be recognised in a multimedia document when one or more of these characteristic media-objects are recognised. An observation model for a steam engine could be its body shape (consider a close view of a stationary engine as in the left picture of fig. 5). Another observation model could comprise its huff-and-puff and a smoke cloud (consider a distant engine as in the right picture of fig. 5). A third observation model could be its whistle when the engine is not at all visible.

Each of the media objects in the observation models might undergo media specific transformations, because of variations from model to model, observation positions and other local noise factors. Recognition functions provide means to recognise these objects in spite of such differences. Some example documents that could be retrieved with these observation models are (1) a video-clip depicting distant smoke over a grassland with the huff-and-puff of a passing steam engine*, (2) a picture of the Fairy queen, and (3) a document on history of Indian Railways containing pictures of various makes and models of railway engines.

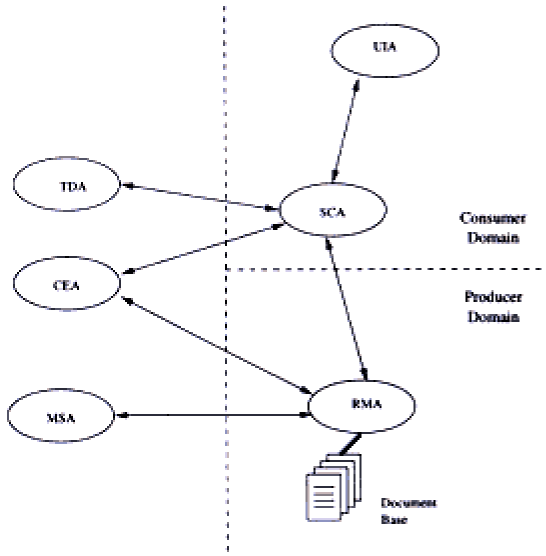
## 3. Architecture for retrieval

Our model of abductive reasoning relies on fusion of information from multiple sources. This necessitates operation of reasoning process at multiple levels of abstraction. At each level of abstraction, we need a different form of knowledge for reasoning. At the lowest level, we achieve media abstraction using the recognition functions. Recognition functions encapsulates structural knowledge about the media form. They can cope up with the heterogeneity of the media forms and format and can identify combinations of complex media-objects inspite of local variations. We use conceptual knowledge for deriving the search specification from the conceptual terms in the queries. Domain expertise establishes the relationships across the concepts in the domain of discourse. It defines the observation model for the concepts in terms of media specific objects and their properties. Collection expertise encapsulates a logical view of the document collections and thus, provides the necessary framework of assumptions for the abductive reasoner. It classifies the documents into broad subject classes so that a coarse grained similarity can be established between the documents and the queries in terms of some common descriptors. The interaction between these knowledge components is shown in fig. 6. Above all, we need a planning module to utilise the other forms of knowledge.



We model retrieval as a problem of distributed AI to exploit the possibility of a distributed knowledgebase. Besides, a distributed model can conveniently cope up with multiple user locations and distributed information sources. We implement the retrieval system with a multi-agent architecture. An agent is a software entity that has its own independent set of beliefs, capabilities, choices and commitments. It functions autonomously in an environment in which other processes take place and other agents possibly exist [16]. In a multi-agent system, problem solving is achieved through coordination of autonomous and possibly heterogeneous set of agents.

We envisage different classes of agents in our system. We briefly describe the classes of agents in our architecture in the following paragraphs. A more detailed account of the architecture is presented elsewhere [6]. The Media Search Agents (MSA) encapsulate the recognition functions and provide media abstraction. An MSA specialises in a certain media form and possesses adequate structural knowledge to autonomously recognise complex patterns in that media form in spite of local variations. MSAs are implemented as mobile agents, so that they can travel to the nodes holding document collections and operate upon them. The advantage of this approach is two-fold. First, the volume of the multimedia data being usually large, it is more economic to move the agents to the data as compared to moving the data to the agents. Second, as the agents instantiates themselves at different nodes and utilises its computing power, a single processing node does not become a computational bottleneck. The Thematic Descriptor Agents (TDA) encapsulate domain knowledge and provide for query refinement. A TDA deals with the concepts in the closed world of a subject domain. It associates some perceptible properties with the concepts. These properties form the basis of the observation models. Besides, a TDA establishes relationships between the concepts enabling property inheritance. The Repository Manager Agents (RMA) encapsulate the knowledge about the physical collection. It deals with the directory structures, file-formats and their identification and the containment relationship between the documents and document components. They control access to their respective repositories. In our architecture, a physical collection need not be localised on a single node but may span over an arbitrary network, typically over a LAN. The Collection Expert Agents (CEA) provide a logical view of the document collections. They incorporate a subject classification and relate the subject classes to a set of documents. The scope of a CEA spans over one or more physical collections. The Search Coordination Agents (SCA) have the capability of reasoning in the distributed environment. An SCA performs the overall planning and coordination for solving a retrieval problem. The plan is dynamically evolved during the course of a retrieval process. Planning also accounts for optimisations and resource constraints that may be applicable to a query. The User Interface Agents (UIA) provide the human-machine interface for the system. It is implemented as two interacting modules. A front-end mobile module may be downloaded on the user machine using a standard browser, such as the Netscape. It provides a user-friendly graphic interface for accepting the query, for presenting the search results and for other management functions. A back-end module manages the user data, such as the registration information, user profile and preferences, etc. A Registration Agent (RA) allows the other agents in the system to be aware of each others presence, so that they may communicate to each other.



The interaction between the agents is shown in fig. 7. The agent based architecture allows for dynamic growth of the system with least software distribution and configuration management overheads. New information sources can be accomodated and new functionality can be realised in the system by the way of addition of new agents with appropriate capabilities. We have implemented these agents over Java Remote Method Invocation (Java-RMI) services, but it could be implemented equally well over any other mechanism, such as CORBA or even plain and simple TCP/IP.

## 4 Conclusion

We have proposed an abductive framework for reasoning for retrieval of multimedia documents. The method proposed in this paper can retrieve documents based on concepts abstracted over multiple media forms. We use conceptual and media knowledge to derive combinations of media patterns that can be used to identify the concepts presented in a query and are invariant over media specific variations. The documents that provide plausible explanations to the expected media patterns are selected for retrieval. The method of partitioning the knowledge base required for retrieval is a unique feature of our approach. It has resulted in modelling retrieval as a problem of distributed AI using well-defined functional units of knowledge. We have proposed a multi-agent architecture for the retrieval system with mobile agents to improve the computational performance.

Daniel E. Atkins, William P. Birmingham, Edmund H. Durfee, Eric J. Glover, Tracy Mullen, Elke A. Rundensteiner, Elliot Soloway, Jose M. Vidal, Raven Wallace, and Michael P. Wellman. Towards inquiry-based education through iteracting software agents. IEEE Computers, 29(5):69--76, May 1996.

Thom Blum, Doglas Keislar, James Wheaton, and Erling Wold. Audio databases with content based retrieval. In Mark T. Maybury, editor, Intelligent Multimedia Information Retrieval, pages 113--135. AAAI Press, 1997.

A. Brink, S. Marcus, and V.S. Subrahmanian. Heterogeneous multimedia reasoning. IEEE Computer, 28(9):33--39, September 1995.

Tat-Seng Chua, Hung-Keng Pung, Guo-Jun Lu, and Hee-Sen Jong. A concept based image retrieval system. In Proceedings of the 27th Annual Hawaii International Conference on System Sciences}, 1994.

Myron Flicker, Harpreet Sawhney, Wayne Niblack, Jonathan Ashley, Qian Huang, Byron Dom, Monika Gorkani, Jim Hafner, Denis Lee, Dragutin Petkovic, David Steele, and Peter Yanker. Query by image and video content: The {QBIC} system. IEEE Computers, 28(9):23--32, September 1995.

Hiranmay Ghosh and Santanu Chaudhury. A distributed architecture for concept-based multimedia documentretrieval. In Ravi Mittal, K.M. Mehata, and Arun K. Somani, editors, Proceedings 5th international conference on Advanced Computing (ADCOMP), pages 362--369, Chennai, December 1997.

Eugene J. Guglielmo and Neil C. Rowe. Natural-language retrieval of images based on descriptive captions. IEEE Transactions on Information Systems, 14(3):237--267, July 1996.

Gareth Jones, Jonathan Foote, Karen Sparck Jones, and Steve J. Young. The video mail retrieval project: experiences in retrieving spoken documents. In Mark T. Maybury, editor, Intelligent Multimedia Information Retrieval, pages 191--214. AAAI Press, 1997.

Hannu Kangassalo. Conceptual level user interfaces to data bases and information systems. In Hannu Jaakkola, Hannu Kangassalo, and Setsuo Ohsuga, editors, Advances in information modelling and knowledge bases}, pages 66--90. IOS Press, 1991.

Inderjeet Mani, David House, Mark T. Maybury, and Morgan Green. Towards content based browsing of broadcast news video. In Mark T. Maybury, editor, Intelligent Multimedia Information Retrieval, pages 241--258. AAAI Press, 1997.

R. Manmatha and W. Bruce Croft. Word spotting: indexing handwritten manuscripts. In Mark T. Maybury, editor, Intelligent Multimedia Information Retrieval, pages 43--64. AAAI Press, 1997.

Sherry Marcus and V.S. Subrahmanian. Foundations of multimedia database systems. Journal of the ACM, 43(3), may 1996.

Virginia E. Ogle and Michael Stonebraker. Chabot: retrieval from a relational database of images. IEEE Computers, 28(9):40--48, September 1995.

Alex Pentland. Machine understanding of human behaviour in video. In Mark T. Maybury, editor, Intelligent Multimedia Information Retrieval, pages 191--214. AAAI Press, 1997.

Michail Salampasis, John Tait, and Chris Bloor. Co-operative information retrieval in digital libraries. http://osiris.sund.ac.uk/cs0msa/bcsir96.ps , May 1996.

Yoav Shoham. Agent-oriented programming. Artificial Intelligence, 60:51--92, 1993.

Stephen W. Smoliar and Hong Jian Zhang. Content-based video indexing and retrieval. IEEE Multimedia, 1(2):62--75, Summer 1994.

Rohini K. Srihari. Automatic indexing and content based retrieval of captioned images. IEEE Computers, 28(9):49--56, September 1995.

Ron Sun. Robust reasoning: integrating rule-based and similarity-based reasoning. Artificial Intelligence, 75:241--295, 1995.

Howard D. Wactlar, Takeo Kanade, Michael Smith, and Scott M. Stevens. Intelligent access to digital video: Informedia project. IEEE Computers, 29(5):46--52, May 1996.

Atsuo Yoshitaka, Masahito Hirakawa, and Tadao Ichikawa. Knowledge-assisted retrieval of spatiotemporal content in multimedia databases. International Journal of Software Engineering and Knowledge Engineering, 7(3), 1997.

Atsuo Yoshitaka, Setsuko Kishida, Masahito Hirakawa, and Tadao Ichikawa. Knowledge-assisted content-based retrieval for multimedia databases. IEEE Multimedia, 1(4):12--21, Winter 1994.