Intelligent Multimodal Interfaces for Visual Information Retrieval

PhD First Year Report

José Manuel de Castro Torres
Supervisor: Dr. Alan Parkes

Computing Department
Faculty of Applied Sciences
Lancaster University

November 2000

TABLE OF CONTENTS:

LIST OF FIGURES:

LIST OF TABLES:

ABSTRACT

This document reflects the work done by the author as a PhD student during his first year. The work can be divided in two major parts, the literature review part and the theses proposal part.

As a result of the second part, certain hypothesis are raised. The target domain of the study is Multimedia Information Retrieval or more specifically Visual Information Retrieval (VIR). The main aim is the integration of adaptable mechanisms with user models and natural ways of interaction into VIR systems.

Suitable characteristics of an ideal VIR system are proposed. The central role of user modelling in such a system is discussed. Approaches to the automated analysis of images in image retrieval systems are considered. A preliminary sketch of a model is also provided. The model is based on Bayesian user modelling techniques and dynamic Bayesian Networks.

Finally, the report presents a provisional work plan for the rest of the work with some concluding remarks.

1 INTRODUCTION

With the continuous advances in technological areas such as multimedia coding, computer networks and computer hardware, and the growing of importance of information, the need to access multimedia repositories is now arising.
In Multimedia Information Retrieval (MIR) systems the dominant indexing and querying metaphor has typically been traditional keyword and text queries. Such approaches are insufficiently rich to deal with the diversity of multimedia information types.
The new paradigm in multimedia retrieval are the Content Based Multimedia Retrieval (CBMR) systems, i.e., systems capable of extracting information about the multimedia content using a set of content analysis methods. Such systems are capable of delivering much more query power to the user since they allow the query to include information about the multimedia content (example: "find images featuring trees").
A further interesting possibility would be to tailor the system to the user's needs by incorporating knowledge of the user and application domain. This is the central motivation of the PhD project that is described in this document.
Some other characteristics that, in conjunction with content-based retrieval, can be incorporated in a MIR system in order to enhance the system performance are:

- Adoption of more natural and powerful human-computer interfaces by using multiple modes of input/output (multimodality);
- Employment of knowledge about the user in order to obtain improved query results, through the capture and management of user profiles;
- Exploitation of domain knowledge, i.e., those aspects of the application which can be adapted or are required for the operation of the adaptive system;
- Incorporation of an interaction model to represent the interaction between the user and the application.

As seen, multimodality and the system's capacity to adapt to the user at both interface and query levels, are desirable characteristics of a sophisticated MIR system.
The on going work developed in the PhD context is restricted to the problem of Visual Information Retrieval (VIR), since the results obtained are expected to be sufficiently valid to be applied in more generic MIR systems.

1.1 Information seeking behaviour of image users

In designing a VIR system, a crucial aspect is to detect and predict typical behaviour patterns of the users, and identify their needs and expectations.
To provide a full description of the uses being made of visual information in the diverse domains is extremely difficult. The following examples could be interpreted as a snapshot of the some activities that, in some way, depend on the use of images (Eakins, 1999):

- Crime prevention;
- Medicine;
- Fashion and graphic design;
- Publishing and advertising;
- Architectural and engineering design;
- Historical research.

Ornager, (1997), in her study of journalists using newspaper image archives, proposes a user typology, as follows:

- The specific inquirer who asks very narrow questions, because he/she has a specific photograph in mind;

- The general inquirer who asks very broad questions because they want to make their own choice;
- The story teller inquirer who tells a story being covered and is open to suggestions from the archive staff;
- The story giver inquirer who hands the story being covered over to the staff wanting them to choose the photograph(s); and
- The fill in space inquirer who only cares about the size of the photograph in order to fill an empty space on the page.

A study of seven picture libraries which sought to develop a general-purpose categorisation of user requests for still and moving visual images, carried out by Armitage and Enser, (1997) led to the identification of four major types of query: image content (‘find me some images of ...’); identification/attribution/provenance checking; accessibility of image/artist of work (e.g. ownership/viewing availability); and miscellaneous (e.g. requests for administrative procedures only, or unusable queries). These authors analysed some 1749 image content queries across the seven libraries involved, categorising them into 'by named artist'; 'known items'; 'unique subject' and 'non-unique subject'.

If it can be reliably established that different types of users do in fact require different styles of interaction with image retrieval systems, the task of the systems designer will be made considerably easier (Eakins, 1999).

1.2 The Ideal VIR System

The essentials of a VIR system are three main operations: *Query formulation*, *Search*, and *Information Provision*. Considering the first and third of these, we have identified the following requirements:

Query formulation

- the integration of textual and visual media
- the acquisition of knowledge about the user’s specific search requirements
- the provision of interactive assistance in formulating an appropriate query
- the use of other contextual information to increase the specificity of the query

Information Provision

- the presentation of information in the most appropriate form to represent the required information content
- the adaptation of the actual presentation of information to suit the particular user

The above requirements suggest a central role for techniques to acquire knowledge about the user and to use that knowledge to guide the query formulation and information presentation components of the system. This is based on *user modelling*. User modelling implies *adaptivity*, in that the VIR system adapts itself to suit the specific user according to knowledge it has acquired about that user.

The second major feature of a VIR system is that it must be able to gain access to the *content* of the information. As the information processed by a VIR system is predominantly pictorial, this implies that the system must be able to extract relevant features from the images in order both to determine which images best match the user’s query, and then select the most appropriate combination of images to present to the user.

The above requirements indicate that an ideal VIR system would reflect the best characteristics of both computers and the humans. Existing VIR systems tend to be very ‘low level’, i.e., extremely focused on low-level content analysis and unable to adapt to the individual user. There is thus a gap that must be filled between current VIR systems and the ideal.

1.3 Structure of the Report

The remainder of this report reflects the two major areas on which this work focuses, namely, Visual Information Retrieval (the first area) and modelling, perception and learning performed by an autonomous module that we call an 'Agent' (the second area).
The next section deals with the description of VIR and some of the aspects related, beginning with an introduction to traditional Information Retrieval systems.
Multimodal perception, machine learning and user modelling are the main subjects treated in section 3. This section suggests the importance of Artificial Intelligence when the objective is to develop 'Intelligent' Multimedia systems.
Section 4 briefly reviews some existing projects in "Intelligent" VIR systems. Some of the shortcomings of these VIR systems are also presented.
The PhD project plan is presented in section 5, stressing the major objectives and the tasks required to achieve them. Finally, section 6 provides some general conclusions about the PhD project.

2 VISUAL INFORMATION RETRIEVAL

2.1 Introduction

Traditionally, Information Retrieval (IR) is concerned with the representation, storage, organisation, and accessing of information items. In principle, no restriction is placed on the types of item handled in information retrieval (Salton, 1983). Though the media objects used by an information retrieval system can be diverse, text has historically been the principal featured media type in such systems. Other media types have been treated as highly informative sources, but have been primarily linked for retrieval based upon search of the text (Kowalski, 1997).

Multimedia Information Retrieval presupposes indexing and description of the multimedia content. This section focuses on the requirements of indexing and description and makes particular reference to the MPEG-7 standard for information representation.

Query processing, information visualization and evaluation of the information retrieval operation aspects are also discussed. Finally, a short list of projects in visual information retrieval and some concluding remarks are presented.

2.2 Visual Information Retrieval

The main purpose of Visual Information Retrieval is to retrieve images or image sequences that are relevant to a query operation. Visual information retrieval is a form of multimedia information retrieval application. It represents an extension to traditional information retrieval, were the documents retrieved are predominantly pictorial, rather than textual. Multimedia information retrieval systems can be classified in terms of the kind of media they feature, such as audio, images, moving images or video and text.

In VIR systems, it is crucial for the systems to gather knowledge about visual documents that belong to the repository. Tools to analyse visual elements such as colour, texture, object shape and spatial relationships should permit searching for visual data by referring directly to its content (Bimbo, 1999). In fact, such features, directly related to perceptual aspects of image content, together with high-level concepts such as objects, roles, events or actions, are used as clues for retrieving images that match the query.

In *query by content*, there are several approaches to query formulation. A common approach is *query by image*. In this case, an image is used as the query image, i.e., the search engine will try to retrieve images of a high degree of similarity to the reference image, according to defined measures of similarity.

A further challenge of VIR systems is to use manually applied annotations (mainly text) in conjunction with information automatically extracted from the visual content (mainly numerical information) to support the retrieval operation.

Different research fields are involved in VIR, such as information retrieval, multimedia modelling, image/video analysis and processing, pattern recognition and computer vision, multimedia databases, man-machine interaction and information visualisation, and psychological modelling of user behaviour.

In a VIR system, several subsystems can be identified: the index system, a user interface for query formulation and information visualisation, and a search engine. In the rest of this section more attention will be paid to the relevant aspects of a VIR system.

2.3 Feature Extraction and Content Based Image Retrieval

During the indexing stage of VIR, important perceptual features of the images and/or video such as colour, texture, shape, image structure, spatial relationships and motion are associated with each visual image document from the content repository. The representation of these perceptual features is a fundamental problem in VIR (Bimbo, 1999). Image analysis and pattern recognition algorithms provide the means to automatically extract numeric descriptors that give a quantitative measure of these features.

2.3.1 Colour

Colour is a sensation produced in the brain in response to the incidence of light on the retina eye. The sensation of colour is caused by differing qualities of light emitted by light sources or reflected by objects (Levkowitz, 1997).
In physical terms, colour perception arises from the spectral energy distribution of the electromagnetic radiation that strikes the retina, expressed as a function of wavelength $E(\lambda)$ in the visible range of 350-780 nm.
The human visual system reacts to differences in colour through three distinct types of photoreceptor cells, called *cones*, in the retina that have long (L-type), medium (M-type), and short (S-type) wavelength sensitivity[1]. Cones operate as filters for different ranges of wavelength. L, M and S-type peak sensitivity is 565-570 nm, 530-550 nm and 420-440 nm, respectively. Associated with these L, M and S cones are the colours red, green and blue, respectively.
It is accepted among the research community that colour is a tri-stimulus phenomenon, i.e., human colour perception takes place in a three-dimensional space. This is referred to as trichromacy. This leads to the usual method of representing colour organisation and modelling called a *colour model*.
A colour model (also *colour solid* or *colour space*) is a three-dimensional body used to represent some colour organisation according to a particular choice of three co-ordinates that describe colour (Levkowitz, 1997).
Such colour models can be instrument-oriented, i.e., model the colour sets of specific instruments such as colour monitors and printers, and be based on human visual models, i.e., provide a perceptual ordering and uniformity[2].
From the psychological point of view, perception of colour is best described by colour attributes as brightness, chromaticity and saturation.
The task of representing lower lever colour properties of an image is traditionally done by colour histograms. A colour histogram is obtained by isolating distinct image colours and counting how many pixels belong to each colour. To find perceptually similar images based on colour histograms requires the use of similarity measures that reflect human perception. In this task, different colour models can be used, different quantization and clustering methods and different similarity measures such as histogram intersection, incremental histogram intersection, weighted Euclidean distance, colour coherence vectors and others.

2.3.2 Texture

Texture can be considered to be the degree of smoothness of an object's surface. Texture affects the reflection of light, and is made more visible by shadows formed by its vertical structures. Texture is a term more generally used in pattern recognition to refer to image

[1] There is a fourth kind of photoreceptor called a *rod*. Rods are achromatic light sensitive cells, so they only see "black and white" and are responsible for night and other low light level vision.
[2] The distance in the co-ordinate system indicates the size of perceived colour differences uniformly over the whole colour space.

patches of any size that are characterised by differences in lightness. It concerns the higher frequencies in the image spectrum. Smoothed images are not usually considered to be textured images. Properties such as the size of the image patch, the number of distinguishable grey-level primitives and the spatial relationships between these primitives, are all interrelated, and all characterise a texture (Bimbo, 1999).
Depending on the scale selected to analyse a texture, a texture image can appear textureless, at a coarse scale, or neutral. Humans perceive textures through the observation of properties such as granularity, directionality and repetition.
A texture is usually represented by numerical vectors, holding measures of texture features, since it is not convenient to represent textures in non numeric terms. Image processing operators, operating in either the space or frequency domain, are used to derive measures of texture features. A texture is therefore modelled as a point in a suitable multidimensional feature space.
The auto-correlation function, co-occurrence matrices, fractals, auto-regressive models and stochastic models are among the most usual space models adopted. Some frequency-based models also exist, such as power spectrum or wavelet transform.
Most of the above models do not attempt to directly represent human perception, so they place that burden on the similarity measures adopted. In fact, for a texture model, given two different points in the multidimensional texture feature space of the model (representing two different textures), the distance between those two points does not reflect the difference in human perception of the corresponding textures.
Certain textural representation models allow the adoption of Euclidean distances that more closely represent distances between textures as perceived by humans. These models express textures in terms of contrast, coarseness and directionality, such as that used in the QBIC VIR system (Niblack, 1993). A further perceptually oriented model of texture is the one used in the Photobook VIR (Pentland, 1993) system and is based on the Wold decomposition.

2.3.3 Shape and Segmentation

A *shape* is a characteristic of an object often referred to in terms of its spatial contour. It identifies the object as a meaningful form. It is obtained as a product of the segmentation process. This process divides a visual scene into a number of individual objects or contiguous regions, differentiating them from each other and the image background.
Shapes that are represented in images are often irregular and complex, i.e., without regular properties such as, for example, circles or squares. The shape representations ought to be achieved through a set of objective and concrete features rather than a subjective textual description since they represent physical properties of the object.
The features extracted using the shape analysis tool can characterise either the global form of the shape (e.g. its area, or the orientation of its major axes), or local elements of its boundary such as corners, characteristic points, and so on (Bimbo, 1999).
As for textures, shapes can be viewed as points in the shape feature space, and mathematical distance functions can be used to compute similarity measures between two shapes in two images. In this approach, usually one of the points represents the shape feature vector of the query image and the other point is the feature vector of the shape descriptor of the visual document from the repository being tested.
Other approaches to shape representation are the relational approach and the shape through transformation approach. These are discussed in more detail in Bimbo (1999).
The properties of invariance with respect to translation, rotation and scaling are often desired for the shape descriptors, since the results of all the three operations referred are basically different views of the same shape. Then it is possible to deduce that different projections in different images of similar visual objects are similar according to the similarity function adopted.

2.4 Multimedia Content Description Interfaces

An effort to standardise description of multimedia content is being made by the Moving Picture Experts Group (MPEG) of the ISO organisation. MPEG-7, formally called "Multimedia Content Description Interface", aims to create a standard for describing multimedia content that will support some degree of interpretation of the media object's meaning, which can be passed onto, or accessed by, a device or a computer code. MPEG-7 is not application specific; rather, the elements that MPEG-7 standardises are intended to support as broad a range of applications as possible (MPEG Requirements Group, 2000).

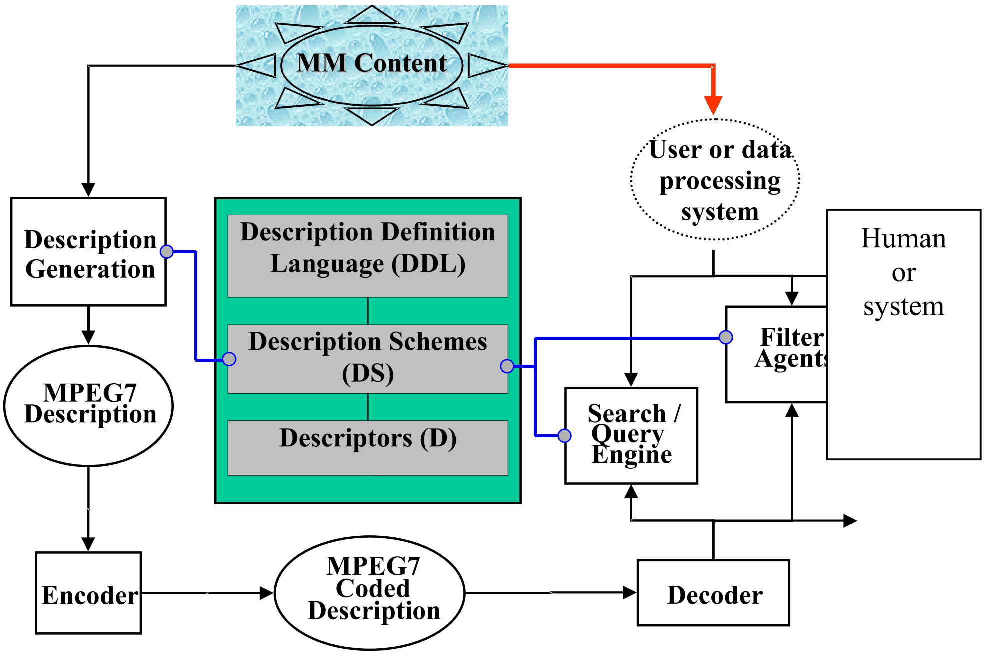

Figure 1 - An abstract representation of possible applications using MPEG-7 (from MPEG Requirements Group, 2000)

Figure 1 shows a possible MPEG-7 chain in practice. The square boxes depict tools that support processes such as encoding or decoding, whereas the circular boxes represent static elements, such as a description. The dotted boxes in the figure encompass the normative elements of the MPEG-7 standard.

2.5 The problem of Query Processing

In Visual Databases where the search is based on content, the VIR system must gather content information about the image repository, usually a set of numerical features for each image, and textual information, usually manually provided. These methods pose some problems for query formulation processing. A further class of problems concern how the image metadata is indexed, since the high dimensionality of the feature vectors involved reveals a real problem for large-scale image collections (Rui, 1998). Additional problems are inherited from traditional textual IR systems, such as the definition of similarity measures or the clustering of similar document items. The following sections discuss how these problems affect the query processing stage in VIR.

2.5.1 Similarity Measures Models

Several similarity models exist for VIR systems. One well-known model embodies an assumption that human similarity perception is based on the measurement of an appropriate distance in a metric psychological space. In this theory, it is assumed that a set of features models the stimulus properties so that it can be represented as a point in a suitable feature space.
Many studies have pointed out certain inadequacies of the feature model and metric distances. Other similarity models that exist are the virtual metric space, Tversky's model, and the transformational distances (Bimbo, 1999).
Other models have the potential to adapt to user subjectivity, such as FourEyes (Minka, 1996).

2.5.2 The problem of High Dimensional Indexing

Finding index structures that support efficient search of an image database continues to be a problem in VIR systems. The existing popular multidimensional indexing techniques feature structures such as the R*-tree and the SS +-tree, but the overheads of using these complex index structures are considerable. A more recent approach, which seems to offer better prospects of success, is the use of similarity clustering of images, allowing hierarchical access for retrieval and providing a way of browsing the database (see Celentano, 1999; Eakins, 1999).
Before one can utilise any indexing technique, it is beneficial to first perform a dimension reduction (Rui, 1998). In fact, since the dimensionality of the feature vector in a VIR system is normally very high, the embedded dimension is much lower. Rui refers to at least two techniques that achieve this: Karhunen-Loeve Transform (KLT) and column-wise clustering.

2.5.3 Ranking Algorithms

In the first Boolean systems of IR, the presentation of the items found by a query did not reflect a ranking function since all the retrieved items met all aspects of the Boolean query. In this case, hits are retrieved in either a sorted order, e.g. by Title, or in time order from the newest to the oldest item.
With the introduction of ranking based upon predicted relevance values, the status summary displays the relevance score associated with the item along with a brief descriptor of the item. This relevance score is an estimate of the search system on how closely the item satisfies the search statement, being typically normalised to a value between 0.0 (not relevant) and 1.0 (entirely relevant) (Kowalski, 1997).
Ranking models can be divided into two types: those that rank the query against individual documents and those that rank the query against entire sets of related documents (Harman, 1992).

2.5.4 Document Clustering

Cluster analysis is a technique for multivariate analysis that assigns items to automatically created groups based on a calculation of the degree of associations between items and groups (Rasmussen, 1992). In the information retrieval field, cluster analysis has been used to create groups of documents with the goal of improving the efficiency and effectiveness of retrieval, or to discover existing relationships in complex data such as the terms in a textual document collection or the feature set in an image collection.

The two main types of cluster analysis methods are the non-hierarchical, which divide a data set of N items into M clusters, and the hierarchical, which produce a nested data set in which pairs of items or clusters are successively linked (Rasmussen, 1992).
The non-hierarchical methods such as the single pass and reallocation methods are heuristic in nature and require less computation than hierarchical methods. However, the hierarchical methods have usually been preferred for cluster-based document retrieval (Rasmussen, 1992).
According to Rasmussen, (1992), to proceed with a cluster analysis technique in IR one needs to solve problems such as:

- selection of the attributes on which items are to be clustered and their representation;
- selection of an appropriate clustering method and similarity measure;
- creation of the clusters which can be a computational expansive task;
- validation of the results obtained;
- constant updating of the clustering, if dynamic;
- selection of a method for searching clusters or the cluster hierarchy if the aim is to use cluster collection as the basis for information retrieval.

Minka, (1996), defines *groupings* as a set of one or more image regions ("patches") that are related in some way. He uses hierarchical clustering to form two kinds of groupings in the FourEyes system. FourEyes computes within-image groupings for each similarity measure, such as colour or texture, by using hierarchical clustering via the shared neighbour algorithm. The second grouping operation is performed across-image and these groupings are computed from a hierarchical clustering of an attribute measured over the within-image groupings. All of these within-image and across-image groupings are computed off-line, before the user interacts with the system.

2.5.5 Probabilistic Retrieval and Relevance Feedback

Relevance feedback is a process wherein users identify relevant documents in an initial list of retrieved documents, and the system then creates a new query based on those sample relevant documents (Croft, 1995). Algorithms for automatic relevance feedback have been studied in IR for more than thirty years, and are therefore well established and effective.
In text information retrieval systems, the central problem in determining relevance is to select appropriate "features" (words, phrases) from relevant documents and calculate weights for these features in the context of a new query. This problem is substantially more difficult in environments with large databases of full-text documents. In addition, people searching databases in real applications often use relevance feedback in different ways than those anticipated by IR researchers. Feedback techniques were developed to improve an initial query and assumed that a few relevant documents (all those in the top ten, for example) would be provided. In many real interactions, however, users specify only a single relevant document. Sometimes that relevant document may not even be strongly related to the initial query, and the user is, in effect, engaged in feedback-based browsing.
The above factors mean that traditional feedback techniques can be unpredictable in operational settings. Research aimed at correcting this problem is underway and more operational systems using relevance feedback can be expected in the near future. Relevance feedback techniques in conjunction with learning techniques are also an important part of IR systems that automatically build user profiles.
In VIR systems, the common approach is to use a start image as an example in the retrieval process. Following this first step, the user successively refines the query in an iterative browsing process until finding the desired image(s).
Several VIR systems incorporate *relevance feedback*. These systems automatically create new queries based on sample images identified as relevant by the user in previous queries (Rui, 1998).

2.6 Information Retrieval User Interfaces and the Information Visualisation Problem

When users approach an information access system they typically have only a fuzzy understanding of how to achieve their goals. Thus, the user interface should aid in the understanding and expression of information needs. It should also help users formulate their queries, select among available information sources, understand search results, and keep track of the progress of their search (Hearst, 1999).
Shneiderman (1997) discusses principles of user interface design. Those which are particularly important for information access include the need to provide informative feedback, the permitting of easy reversal of actions, support for an internal locus of control, reduction of working memory load, and the provision of alternative interfaces for novice and expert users.
The field of information visualisation requires new techniques to display large, abstract information spaces intuitively. Until this happens, the role of visualisation in information access will probably be primarily confined to providing thematic overviews of topic collections and displaying large category hierarchies dynamically. Breakthroughs in information visualisation can be expected to have a strong impact on information access systems.

2.7 Evaluation of Visual Information Retrieval

The most frequently used method for evaluating the retrieval effectiveness score of a VIR system is that used by the text retrieval community. This is the *precision-recall* method as seen in Table 1.

Classification of document items according to the query formulated	Relevant	Not relevant
Retrieved	I (correctly retrieved)	II (falsely retrieved)
Not retrieved	III (missed)	IV (correctly rejected)

Table 1 - traditional evaluation performed in IR according to evaluator's judgement

In this method *Recall* represents a measure of the ability of the system to retrieve all the documents that are relevant and is defined as:

$$Recall = I / (I + III)$$

Precision is a measure of the ability of the system to retrieve only the documents that are relevant and is defined as:

$$Precision = I / (I + II)$$

2.8 Related work and existing Systems

In this section, a brief list of some of the existing projects in trying to improve the efficiency of the actual Multimedia Information Retrieval systems is presented.

- PicHunter: a prototype content-based image retrieval system that uses Bayes rules to predict the user's target images (Cox, 2000);
- AlFresco: an interactive, natural-language centred system for a user interested in frescoes and paintings, with the aim not only of providing information, but also of promoting other masterpieces that may attract the user (Stock and Team, 1993);

- FourEyes: an interactive learning agent that selects and combines models based on examples from the user is included in PhotoBook, a tool for performing queries on image databases based on image content. In the FourEyes system, similarity and attribute extraction models are selected and combined according to information provided by examples gathered from the users (Minka, 1996);
- Miracle: the Miracle system offers semantic access to a large, experimental multimedia information base, using a dialogue manager in conjunction with the retrieval engine (Stein, 1997);
- "Multimodal Indexing, Retrieval and Browsing: Combining content-based image retrieval with text retrieval"[3]: This project aims to integrate innovative matching, sophisticated text retrieval, and new query formulation methods to create a powerful platform for indexing and retrieving text, images and video;
- Multimedia Analysis and Retrieval System (MARS)[4]: this system adopts a relevance feedback based interactive retrieval approach to try to bridge the gap between high-level concepts and low-level features and subjectivity of human perception of visual content (two distinct characteristics of CBIR-Content Based Image Retrieval systems). During the retrieval process, the user's high-level query and perception subjectivity are captured by dynamically updated weights based on the user's feedback (Rui, 1998);
- Content-Based Image Retrieval[5]: This joint project between the Institute and the Manchester Visualization Centre, University of Manchester, is funded by the JISC (Joint Information Systems Committee) Technology Applications Programme. Its principal aims are: a) To report on the current state of the art in content-based image retrieval, with particular emphasis on the capabilities and limitations of current technology, and the extent to which it is likely to prove of practical use to users in higher education and elsewhere, and to identify areas in which further research is required. b) To undertake an evaluation of available CBIR software systems. c) To test the applicability of CBIR methods and software systems in a small range of practical application environments, involving representatives of user groups, and in so doing generate a body of relevant experience and knowledge. d) To raise awareness of this technology in the user and developer communities through web-based demonstrator systems, widely disseminated reports and through hosting seminars and workshops (Eakins, 1999);
- VISOR[6]: The aim of this 2-year project is to investigate the information seeking behaviour of image users in specific disciplines or domains. This will lead to an increase in the understanding of how and why people seek for and use visual information.

2.9 Conclusion

This section has attempted to locate VIR in the more general field of Information Retrieval. Specific attention was given to the “multimedia content cycle”, i.e., indexing, description and retrieval. Attention was also given to the specific problem of query processing.

The continuous growth in the number of digital libraries requires the development of more sophisticated VIR systems. Techniques for automating the indexing stage are of VIR are currently being developed. Systems incorporating such techniques are capable of searching for images on the basis of content characteristics such as colour, shape or texture. However, the approach embodied in such VIR systems remains far from the human reality. Novel forms of interaction with such systems are needed, to support richer languages that enable humans to construct more natural and effective queries.

[3] Project URL: ciir.cs.umass.edu
[4] http://www.ifp.uiuc.edu/ or http://jadzia.ifp.uiuc.edu:8000/
[5] http://www.unn.ac.uk/iidr/CBIR/cbir.html or http://www.man.ac.uk/MVC/
[6] http://www.unn.ac.uk/iidr/VISOR/

The brief list of some of the ongoing projects shows that there are too many fields of research that need to be explored and that will have an impact that goes beyond the limited field of the Multimedia Information Retrieval research area.

The following section considers the nature of the interface between a VIR system and the human user. The component of the computer system that embodies elements of perception, learning capacity, reasoning and pro-activity we call the Agent. The use of Agents can be a key factor in developing sophisticated VIR systems. Important characteristics of agents such as perception and learning are discussed in the following section.

3 AGENT PERCEPTION AND LEARNING

3.1 Introduction

An agent can be interpreted as an entity that has the capacity to perceive the surrounding environment through sensors and act through actuators. An agent can be a human, a robot or simply a software component.

Ideally, for each percept sequence, a rational agent should act in order to maximize its performance measure (Russel, 1995). To do this, the agent generally uses built-in knowledge in conjunction with perceived information. If its internal knowledge or model is able to evolve over time as a result of the agent's interaction history, we say that the agent has learning capabilities.

For VIR, an agent can perform a key role in mediating or assisting the interaction between the user and the system.

Following the presentation of a generic interaction model of an adaptive system, this section discusses two characteristics that an agent based VIR system should have:

- Multimodal perception;
- Learning and user adaption.

3.2 The Generic Interaction model

The *interaction model* is an important component of a generic adaptive system, and is an abstraction representing its interaction with the user.

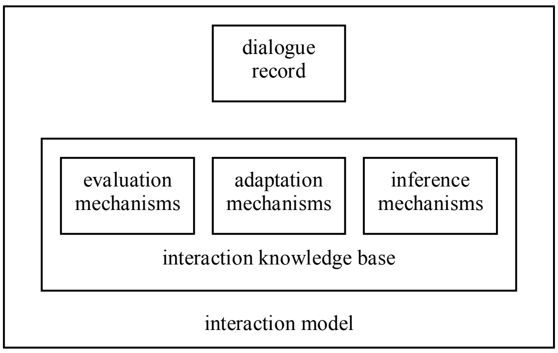

Figure 2 - An interaction model overview

In Figure 2 the *interaction model* captures and stores selected raw data from the interaction in the *dialogue record* and represents the *evaluation mechanisms*, the *adaptation mechanisms* and the *inference mechanisms* in the *interaction knowledge base*.

The *dialogue record* is a low-level trace of the user's observed behaviour that is maintained for as long as required by the adaptive system. As an abstraction of the interaction, it may contain records of sequences of keystrokes, mouse clicks, menu selections, time stamps or other elements that describe the structure of the dialogue between a user and an interactive computer system.

The *interaction knowledge base* interprets the dialogue record and describes the possible inferences, the feasible interaction evaluations and the possible interaction adaptations. While the user model and the domain model define what can be inferred and adapted, it is the

interaction knowledge base that combines these models, making inferences about the user and adapting the system.
The attributes represented in the interaction model must be defined either in the user model or in the domain model and this is what can be used to make inferences about the user's beliefs, plans, goals, characteristics or profile data.
In addition to the above, the data obtained may also be used by the system to evaluate its own inferences, adaptations and recommendations and correct its own behaviour using reflective mechanisms or techniques such as relevance feedback.

3.3 Multimodal Perception

In Multimodal systems, natural input modes such as speech, pen, touch, hand gestures, eye gaze and head and body movements are co-ordinated with multimedia system output (Oviatt, 1999).
Multimodal systems represent a research-level paradigm shift away from conventional windows-icons-menus-pointers (WIMP) interfaces providing users with greater expressive power, naturalness, flexibility and portability (Oviatt, 1999).
With multimodality, the entropy of the information stream provided by the user to the system increases dramatically compared to more traditional human-computer interfaces. Moreover, multimodal systems must evolve on the basis of multidisciplinary efforts, most notably involving the field of cognitive science.

3.3.1 Multimodal Frameworks

The complexity of Multimodal interfaces raises several problems that are not present in more traditional user interfaces. Several communication channels between user and system operate in parallel, hopefully to resolve the ambiguities arising during interaction. The system must be able to co-ordinate and synchronise the communication channels in order to deal with the multiple information streams provided by the multiple modes of input.
One approach to dealing with the problem of interpreting multimodal input is the adoption of an agent-based framework (Cheyer, 1998). Cheyer et al., (1998) proposes an Open Agent Architecture (OAA) that provides a general purpose infrastructure for constructing systems composed of multiple software agents written in different programming languages and running on different platforms. The core services of OAA are responsible for domain-independent co-ordination and routing of information and services. These services can be classified in three areas: agent communication and co-operation, distributed data services, and trigger management.
Multimodal architectures play a major rule in multimodal systems and the challenge will be for them to handle the time-critical nature of parallel interdependent input signals, while being optimised for error avoidance and robustness (Oviatt, 1999).
Oviatt also points out that that cognitive science will play an essential role in guiding the design of robust multimodal systems. One of the most critical subsystems, the conversational subsystem, is discussed in the next section.

3.3.2 Conversational Systems

A conversational system is one that supports a natural dialogue between the human user and a software application. A key factor in developing conversational interfaces is "understanding the model of human communication" (Trower, 1997), so that it is exploited when suitable, and rejected as a mode of communication in those situations where it is inappropriate. Essentially, conversational interfaces employ a spoken conversation metaphor to structure interactions. This metaphor "allows a system to create a shared context and protocol with a

user to communicate its abilities, constraints, and level of understanding" (Hayes and Reddy, 1983).
Furthermore, as Roy et al (1997) point out, "an effective conversational approach utilises directive and time-out prompts, implicit, explicit and context-free confirmations, and modelling of dialogue states and transitions to better represent conversational flow" (Roy et al, 1997).
According to Langley et al (1999), conversational interfaces allow users to engage in spontaneous, interactive conversations, incrementally arriving at the desired information in far fewer steps than the steps required to search a ranked list of results of a search performed online. They suggest developing a system that mimics the type of human interaction in which humans engage in a natural dialogue to arrive at an answer.
Stein et al., (1997) assert that one of the most important tasks in information retrieval is the clarification of information needs. This clarification, obtained through interaction between the user and the system, is meta-communication about dialogue goals and strategies and it can be modelled as conversational interaction.

3.4 Learning and Adaptation in Computing Environments

The term *User modelling* refers to the process of gathering information about users in order to provide services or information adapted to their specific requirements (McTear, 1993).
The main goal of an intelligent interface is its capability to adapt, i.e., dynamically change without the direct interaction of the user. Such systems change aspects of their structures, functionality or interface in order to accommodate the differing needs of the users over time (Benyon et al, 1987). An adaptive computer system gathers information about the user and uses this information to adapt itself so that the user can perform tasks more efficiently.
In a multimedia information retrieval application context, information about both the user and the utilisation context can be precious in the formulation of the query.

3.4.1 User Modelling

A *user model* is a set of beliefs about the user. User Models are important since they represent information about the user so that a system can operate in more effective way. In the case of human-machine interaction, this model is stored in the machine and represents an attempt to model salient aspects of the user.
Kay, (1995), asserts that the user in user-machine interaction can be considered as two subsystems. The first, a *private area* or $\text{user}_{private}$, represents those aspects of the user that are not accessible to the machine. The second subsystem, user_{shared}, represents the shared aspects of the user such as the commands that the user has entered and other user actions afforded by the interface.
User models can be constructed by the system explicitly, i.e., the system asks to provide information so it can create or update the user model. An alternative approach is for the system to construct the user model in an unobtrusive way based on its own analysis of the user's behaviour.
One important approach to user modelling is through the adoption of *stereotypes*. Using a gradual construction mechanism, communities of users are dynamically identified so that any one user can be associated in varying degrees with one or more stereotypical communities of users.
User modelling has been studied as part of many different research topics such as knowledge representation, planning, natural language understanding and generation, intelligent teaching systems (here, user models are sometimes called *student models* or *learner models*) and intelligent information retrieval.
For Information Retrieval, Belkin, (1997) identifies three actors: the *user*, the *mediator* and the *information source*. In order to solve certain specific problems of information retrieval,

Belkin suggests that IR can, and indeed should, be considered to be a dialogue. Each actor must have dynamic models of the others for the system to operate effectively. Belkin introduces the notion of ASK-based retrieval where the information retrieval action is seen as a process of solving an Anomalous State of Knowledge (ASK) by the user. This state is transmitted to the system through some communicable structure, which is used by the system to support the retrieve operation.
Further information retrieval systems based on cognitive user modelling such as THOMAS and GRUNDY (and also ASK, mentioned immediately above) are described in more detail by Ellis, (1990).
The user model in an information retrieval system should support two different aspects of user needs:

- Short-term information needs: the information gathered from the user from the queries submitted in several sessions can be used to infer better results for a specific query within a given session;
- Long-term information needs: the system should be capable of suggesting new and potentially interesting material added to the repository after learning about the main topics of interest to the user.

A further technique, used recently as part of the *Office Assistant* in the Microsoft Office'97 application suite, is Bayesian User Modelling (Horvitz, 1998). The aim of this approach is to analyse the interaction of the user over time and to model the user's time varying needs in the context of the use of the Microsoft Office software application.
MPEG-7 (discussed earlier) defines a model for user preferences with respect to multimedia material. The model is instantiated through the *UserPreference* Description Scheme. User preference descriptions can be correlated with media descriptions to locate desired content. Correspondence between user preference and media descriptions is intended to facilitateaccurate and efficient personalization of content access and content consumption (MPEG MDS Group, 2000).

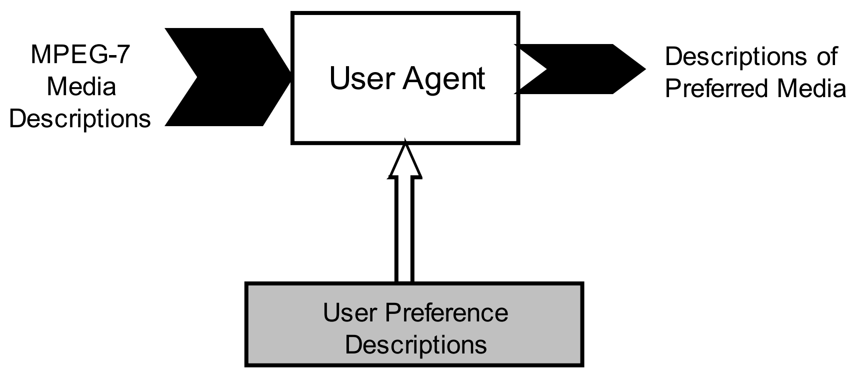

Figure 3 - A generic usage model for user preference and media descriptions (extracted from document (MPEG MDS Group, 2000)).

User preference descriptions are used by consumers (or their agents) for accessing multimedia content that reflects the user's personal preferences. A generic usage model is depicted in Figure 3, where a user agent takes media descriptions and user preferences as input and generates a filtered output containing descriptions of media that meet personal preferences. In specific applications, the output may be media locators of preferred media, or a summary of an audio-visual program where the type of the summary satisfies the user's summary preferences. For example, a user may prefer to view only the goals of a soccer match, but another user may prefer a 30-minute highlight summary of the match (MPEG MDS Group, 2000).

3.4.2 Machine Learning

In a general sense, learning can be defined as the process of improving one's ability to carry out a task. The kind of learning discussed here is cognitive acquisition (such as learning to recognise an object by acquiring a description of it), rather than motor skill refinement, such as learning how to ride a bicycle.
Learning can be viewed as the combined process of inferring and memorising. Inference is deemed to be any kind of reasoning or knowledge transformation. That is, inference results in the production of new knowledge (Michalski, 1994).
Functionally speaking, learning will often consist of filtering data to obtain 'interesting' information and then refining this in some way so that 'useful' facets remain. Thus data is distilled into 'knowledge', whence it is stored for future use (Dutton, 1996).
An ideal VIR system should be capable of using the knowledge acquired from previous interactions with a specific user or perhaps with other users to better cope with the retrieval needs of the that user in future interactions. In order for this to be possible, some learning mechanism must be incorporated in the VIR system.
In the following sub-sections, several machine learning techniques are briefly discussed.

3.4.2.1 Connectionist methods (Neural Nets)

The *perceptron* was one of the earliest neural network models (Rosenblatt, 1962). A perceptron models a neuron by taking a weighted sum of its inputs and sending the output 1 if the sum is greater than some adjustable threshold value, otherwise it sends 0.
The *n* inputs and *n* connection weights are typically real values, both positive and negative. If the presence of some feature *i* tends to cause the perceptron to fire, the weight *i* will be positive; if the feature *i* inhibits the perceptron, the weight *i* will be negative. The perceptron itself consists of the weights, the summation processor, and the adjustable threshold processor. Learning is a process of modifying the values of the weights and the threshold. The weight of the threshold can be thought of as the propensity of the perceptron to fire irrespective of its inputs.
A perceptron computes a binary function of its input. Several perceptrons can be combined to compute more complex functions.
A group of perceptrons can be trained on sample input-output pairs until it learns to compute the correct function.
According to Rich, (1991), connectionist models can be used for pattern recognition, pattern transformation and dynamic inference.

3.4.2.2 Symbolic Learning

There are several types of symbolic learning that essentially vary in the amount of knowledge required and the level of inference performed. The less information required to the beginning of the process the better, and the greater the level of inference, the less emphasis is placed on a user or teacher (Dutton, 1996). Moreover, the greater the level of inference, the more potentially complex and/or voluminous the derived knowledge becomes (Dutton, 1996).
Several types of symbolic learning exist, such as rote, learning by taking advice, deduction and induction.
Induction is usually characterised as learning through search of a knowledge space as defined by the particular knowledge representation used. Inductive learning can be supervised or unsupervised.
Unsupervised induction implies that an algorithm is supplied with, or acquires, unclassified examples, and is required to decide which examples belong to which classes, and indeed what

the classes are. Unsupervised induction can be performed using clustering or discovering (Dutton, 1996).
Supervising learning relies upon a set of pre-defined training examples, which, ideally, are indicative of the concepts to be learned. The algorithm learns to discriminate between the given concepts.

3.4.2.3 Statistical Methods (Bayesian)

In general, statistical techniques entail the construction of a probabilistic model to describe a 'sample', from which it is possible to draw conclusions about a population in general. The estimation of the parameters, such as mean and variance, of a given probability density function for a variable is one common application.
The statistical work that is of more interest to Machine Learning mostly concerns inductive inference, where one forms hypotheses based on the data. Here, inference is based directly upon probabilities within a given sample of observations (Dutton, 1996).
Two types of technique for investigating relationships between variables are:

- Regression analysis: this attempts to determine the nature of the existing relationship between two variables;
- Bayesian techniques: these describe decisions, for example, in terms of risk or losses, an example being 'one incorrect prediction may be more costly than another'. A Bayes rule classifier is said to be optimal in that it depicts the minimum error one can expect in a domain.

Statistical methods are often *parametric*, in that they make assumptions about the data in use (Dutton, 1996). For example, the form of the probability density function may be given *a priori*. Some parameters may be estimated from the current sample, often with good results, provided that the sample is large enough for underlying assumptions to remain justifiable. Parametric methods therefore leave themselves open to criticism concerning the validity of assumptions and their generality (Dutton, 1996).

3.5 Intelligent Visual Information Retrieval

Intelligent Visual Information Retrieval, as a specific case of more generic intelligent multimedia retrieval, goes beyond traditional hypertext or hypermedia environments to provide content-based indexing of multiple media and management of the interaction with these materials by representing and reasoning with models of the media, user discourse and task (Maybury, 1997).
New query and browsing paradigms of multimedia repositories will need to consider the evolution of user interfaces that support several types of human-computer interaction. However, as information systems and human-computer interaction become increasingly complex, the entire process of information search, retrieval, and relevance assessment will need to be actively supported by easy-to-use multimodal user interfaces (Stein, 1997).
One of the more demanding problems in VIR systems is the impedance mismatch between users' expectations and high-level concepts and the low-level feature representation usually adopted in these systems. In general application domains, there is no direct link between the high-level concepts and the referenced low-level features (Rui, 1998). Addressing this problem requires an alliance between several research fields, especially Artificial Intelligence and Computer Vision.
The use of multimodality in an Intelligent VIR system presents some significant advantages that remain largely unexploited, such as:

- The amount of user information gathered: a potentially huge amount of information about the user is available to the Intelligent VIR system, i.e., the bottleneck of the information/communication channel from the user to the system is released from the interface;

- Robustness: since multiple modalities of input are used, the system can be designed to be more "error resilient";
- Modality complementarity: that the interaction takes place through multiple models of discourse reflects the fact that the user's natural communication uses multiple modes of communication in a complementary way;
- Expressiveness: Multimodality considerably increases the potential for extremely powerful queries;
- Naturalness: A multimodal Intelligent VIR provides human-computer interaction that is closer to real world communication. Interaction with such systems is thus more natural for the user.

3.5.1 Multimodal Query Formulation Scenarios

In a generic multimedia retrieval system, in principle, any type of AV material may be retrieved by means of any type of query material. This means, for example, that video material may be queried using video, music, speech, etc. It is up to the search engine to match the query data and the MPEG-7 description. (MPEG Requirements Group, 1999). The following briefly describes several potential multimodality/multimedia information retrieval scenarios:

1. *Music*: A sequence of notes played on a keyboard result in the retrieval of pieces of music containing similar sequences. Alternatively, the music retrieved may have certain properties in common with the original sequence (mood, etc.)
2. *Graphics*: A line drawing is used to retrieve images containing similar graphics, logos, ideograms,...
3. *Image*: Objects, colour patches or textures are specified, the retrieved items being used by the user to compose his or her own images of interest.
4. *Movement*: A description of the movements of objects and temporal and spatial relationships between objects is used to retrieve an animation or video featuring those objects.
5. *Scenario*: A description of a sequence of actions results in scenarios containing similar action sequences.
5. *Voice*: An audio excerpt featuring the performance of a particular musician is used to retrieve video, audio and other media objects featuring that musician.

3.5.2 Adaptivity and User modelling in VIR Systems

Generically, *adaptation* means that the system can sense the surrounding environment and take actions proactively to carry out its tasks more effectively.
An *adaptive system* is a system that changes its functionality or interface in order to accommodate the differing needs of the users over time (Benyon et al, 1987).
In order to this adaptation process be possible, some information about the user must be captured, stored and processed. This sequence of operations can be called user modelling. As already mentioned, *user modelling* can be defined as the process of acquiring knowledge about a user in order to provide services or information adapted to their specific requirements (McTear, 1993; Kay, 1995). Belkin (1997) and Ellis (1990) describe approaches to user modelling in information retrieval systems.

3.5.2.1 Adaptivity at the emotional level: Affective Computing

Emotional and rational behaviour are intrinsically associated in the human mind and recent neurological studies indicate that the role of emotion in cognition is essential (Damasio, 1994). In fact, Damasio asserts that emotions play a critical role in rational decision-making, in perception, in human interaction, and in human intelligence.

The importance of emotion, in conjunction with advances in computer science in the ability of computers to express and recognise emotional states opens a broad research area called "affective computing" (Picard, 1995).
Most of the actual computer systems fail when interacting with the emotional part of the users. The new proposed metaphors of human-computer interaction, multimodal in nature, open new categories of computer systems (II, II and IV of Table 2) and the way these interact with the human user as can be seen in Table 2 (taken from Picard, 1995).

Computer System	Cannot express affect	Can express affect
Cannot perceive affect	I	II
Can perceive affect	III	IV

Table 2 - Four categories of affective computing, focusing on expression and recognition

Affective recognition can be applied several areas. Picard (1995) refers to the example of the use of affective recognition to support searching an image archive on the basis of emotional states. An example query could be "find digital photographs reflecting a particular *mood*".
A further application of affective computing could be in the designing of computer systems that could respond to the human user in a more effective way since they could pay attention to the user's affective expression. Picard says that "the online collection of this information could not only lead to more relevant feedback to the user, but could also be of great use in consumer satisfaction studies, not to mention in the development of more pleasurable and effective learning experiences".
In the context of Intelligent VIR systems, an ideal VIR system would incorporate characteristics of affection both at the perception and expression level. The ideal VIR system would belong in category IV of Table 2.

3.5.2.2 Adaptivity in VIR Systems

In a VIR system, user modelling and adaptivity are needed to support the following:

- *adaptation of the queries*

The user's query may be adapted by the system in order to meet that user's specific needs as identified by the user model.

- *adaptation of the information presentation*

Determining the suitability of images is based not only on the content of the images, but also on how that content relates to the user's goal and purpose.

3.6 Conclusion

This section has discussed two important characteristics of rational software agents: perception and learning.
The multimodal perception in a VIR system can play an important role in the interaction between a user and a VIR system. In fact, the more powerful and broader the communication channel is, the more expressive power is available to the user to formulate the queries.
A VIR system must be capable of learning, in order to support its retrieval processes. The process of creating models of diverse users can also be very helpful in enabling the VIR system to adapt to the specific user.
In the next section, the PhD project plan is presented based on the ideas expressed along these last 2 sections.

4 PROPOSED RESEARCH AND WORKPLAN

4.1 Introduction

This section is dedicated to the presentation of the proposal for the research work to be developed during the context of the author's PhD. The work plan is a corollary of the main ideas presented in the previous sections.
The next section introduces the main goal and the more specific objectives for this project. The subsequent sections illustrate the overall approach adopted in the first prototype to be developed. The program of work set up to achieve the defined goals is described in point 4.7. Finally, some general conclusions about the project plan are drawn.

4.2 Research focus

The PhD work described in this document is positioned in the area of Multimedia Information Retrieval. The more specific area of Visual Information Retrieval has been chosen to test and validate the main ideas defended.
The thesis defended in this work is that the incorporation into a VIR system of: 1) more natural or multimodal ways of interaction between the user and the system; and 2) representation and reasoning with models of the media, user discourse and task, and user modelling can enhance VIR systems and bring us closer to the hypothetical "ideal" VIR system.
The proposed research will focus mainly on the development of the following novel techniques:

- The ability of the system to acquire information about the user, and use this information to adapt the information retrieval to the individual user's requirements.
- Enabling the user to communicate, via a multimodal interface, to the system the types of information he or she wishes to retrieve (using at the very least a combination of text, graphics, audio and video in the query formulation). The initial prototype will consider only a simplified conversational interface.
- The prototyping and evaluation of dialogue managers and conversational assistants for multimodality interfaces. Initially, these dialogue managers will be simple in that they will impose constraints on user dialogue in order to reduce the complexity of the dialogue. The user will interact with the system via dialogue templates.

Future research topics related to the above are:

- The modelling of both objective (low-level features) and subjective features (high-level features) of the multimedia content, through the use of knowledge based techniques. This is intended to narrow the semantic gap between high-level concepts used by humans in everyday life and the low-level features extracted from multimedia content by automatic techniques;
- The exploitation of metadata representations of different media such as text, graphics, audio and video in order to create meaningful associations between different media objects.

Some important aspects relevant to the effectiveness of the query process involve the exploitation of:

- Cross-reference information between the several types of media in the multimedia repository (text, graphics, audio and video);
- Sophisticated and more natural user interfaces;
- Knowledge based representations of multimedia;
- User models and contextual information about queries.

4.3 A User Model for VIR

This section sketches a formal model that represents an initial attempt to address the issues of user modelling, adaptivity and multimodality discussed above. The model is based on Bayesian User Models and Bayesian Networks (Horvitz, 1998, Cowell, 1999). First, the model is described. Following this, a simple worked example is presented to illustrate how the proposed model might be applied in a VIR system.

4.3.1 Sketch of a user modelling approach for VIR systems

The problem domain of Visual Information Retrieval consists of the following objects:

- A User (*U*), typically a human who has the goal of retrieving a non empty set of images;
- The set of images presented in the repository $I = \{i_1, ..., i_n\}$;
- A subset, *S*i, of *I*, containing images that totally or partially satisfy the user's goal;

To enable the system to compute the solution to the VIR problem, i.e., derive the set S_i, it must have access to all of the required information. We represent this information as the following function:

$$f: (U, I) \rightarrow S_i$$

It can be seen from the above that the set S_i computed by the VIR system depends on both the user and on the repository.
The information about the user can be viewed as:

$$U = (U_{goals}, U_{actions}, U_{profile}, U_{context})$$

When using the system, the user may have certain specific goals (U_{goals}) concerning the images to be retrieved. The user attempts to reach these goals through a set of interactions with the system ($U_{actions}$). User characteristics such as general interests, cultural information, and so on, are held in $U_{profile}$. Finally, $U_{context}$ features contextual information that may be relevant to the result S_i.

Temporal Dependencies

The arguments of the function *f* evolve over time, so a mechanism is required to deal with this. The temporal variability of *I* can be relaxed, i.e., during a visual information retrieval session one can consider that the set of images remains constant.
The independent variables U_{goals}, $U_{profile}$ and $U_{context}$, can be seen as relevant time-dependent information. The variable $U_{actions}$ can be considered to be an array of time stamped actions.

Dynamic Graphical Model

The actions performed by the user, $U_{actions}$, are perceived by the VIR system as a temporal series of events, E_{ti}. These events are the way in which the user demonstrates his or her goals to the VIR system. We therefore establish a dependency relation between U_{goals} at time t_i and event E_{ti}.
The user goals also depend on $U_{profile}$ and $U_{context}$. In general, a goal, or set of goals, arise on the basis of a user requirement, and are related to problems revealed by the situational context. How a user maps these problems and needs onto goals also depends on that user's psychological profile. Given these considerations, a further dependent relation can be established from the pair $U_{context}$, $U_{profile}$ to U_{goals} at time t_i.

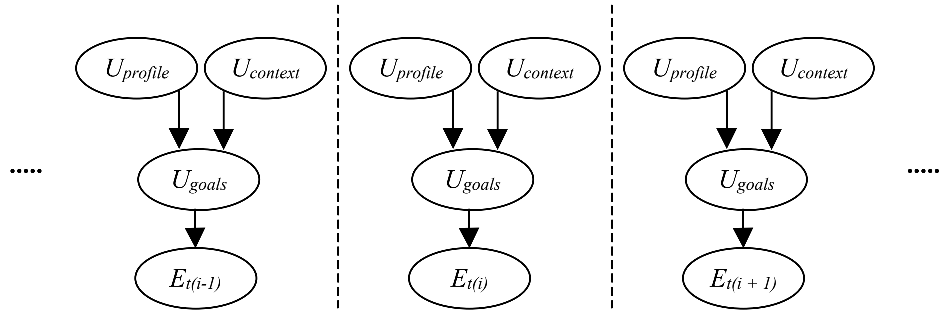

Figure 4 - Dynamic Belief Network representing the interaction between the User and the VIR system

Statistical reasoning is a possible approach to modelling the user in the VIR system. In such an approach, each element of *P(I),* i.e. each possible set of images, has an associated value representing the probability that the set of images satisfies the goal of the user. This probability value is updated as time passes on the basis of the interaction between the user and the VIR system. The VIR system uses all possible information gathered from the user and determines which element S_i of set *P(I)* satisfies the user's goal. This can be an iterative process the termination of which can be determined either by the user, declaring that S_i is valid, or the system, which has computed a probability for S_i that is greater than a preset threshold (set in conjunction with the user). A graphical representation of the VIR framework proposed is presented in Figure 4.

Graphical models that can be used to represent the type of problems considered here, namely, Dynamic Bayes Nets (DBNs), can be used to represent random variables that evolve over time. DBNs allow the state of the system to be represented as a set of hidden and observed states in terms of state variables, among which there can be complex interdependencies. Graphical models are graphs in which the nodes represent random variables and the arcs represent direct dependencies between these random variables. When the arcs are directed, such graphical models are called *Bayesian Networks* or *Belief Networks.* In such networks, each node is associated with a conditional probability distribution. The graphical structure provides a convenient means of specifying the conditional independencies, and hence represents a compact parameterisation of the model.

For VIR, given the observed user actions and using a prior graphical model, the objective is to infer the user's goal and ultimately to calculate S_i, i.e., the set of target images. The User's goal is a time-dependent hidden variable (an array of hidden variables) of the model that the system must compute in order to compute S_i. The other two time-dependent hidden variables $U_{profile}$ and $U_{context}$ are also relevant in the determination of S_i, and ultimately will be inferred from the action sequence of the user. In generic terms, the system observes $U_{actions}$ and attempts to compute:

$$P(U_{goals}, U_{profile}, U_{context} \mid U_{actions})$$

Following the above inference, is is then possible to compute S_i:

$$f: ((U_{goals}, U_{actions}, U_{profile}, U_{context}), I) \rightarrow S_i$$

Bayesian models have been used to diagnose user needs in several applications, enabling those applications to incorporate user-modelling capabilities (Horvitz, 1998).

Finally, the system should be designed so that it gathers as much information as possible during interaction with the user. This is a strong reason for the adoption of multimodal interfaces in VIR systems. The presented framework is intuitively adaptive, since it acquires information about the user to aid in the satisfaction of the user's goals. Though the image of the function *f* has been specified above as a set of target images S_i, this could be extended to deal with a set of personalised styles of interaction and interaction modalities P_i:

$$f: ((U_{goals}, U_{actions}, U_{profile}, U_{context}), I) \rightarrow (S_i, P_i)$$

The above parameters can be used to determine such features as the ways in which the information is presented to the user, the user's preferred interaction modalities, the user's preferred query formulation styles and other adaptive facilities.

4.3.2 The application of the model: a worked example.

A key problem with a model such as the one proposed above is to represent U_{goals}, $U_{profile}$, $U_{context}$, and $U_{actions}$ in a stochastic world (to determine the domain of each random variable, empirical studies have to be performed). Further difficulties arises in creating and updating the Conditional Probability Table (CPT) at each node Our initial approach is to assume that for U_{goals}, $U_{profile}$, and $U_{context}$, some of the information must be inferred (on the basis of statistical reasoning) and the rest is to be explicitly gathered. Following this approach, U_{goals} can be specified as follows:

U_{goals} = *Inferred User Goals from the Bayesian Networks + Explicit User Goals from the query formulation dialogue*

In order to suggest how the above model may be applied, we now present a simple worked example. The example user is carrying out research into the history of the Kings of England in the 19th century, and requires images to illustrate his work. The following is a description of the relevant features that might be found in the model.

The first, very important, step is to determine the interaction space of the user, i.e., the types of user actions that are possible, such as conversational acts, providing profile information in a form, selecting a subset of a displayed set of images, and so on. Note that in a more restricted interface, the interaction space may be the typical WIMP menu-based scenario. In this case, the system is unable to unobtrusively develop a knowledge-level user model (e.g., to infer that the user is a student in history, and so on). The system is likely to need to request the user to explicitly provide such information.

In systems with a rich dialogue style, the system might infer that our user was a history student from keywords extracted from a discussion with the user (e.g. {history, university, teacher,…}), i.e.

Kwords{history, university, teacher}-> $U_{profile}$*(history_student)*

During the same conversation, the user might use the keywords {19th century, King, England, thesis….}, and the system could infer:

Kwords{XIX century, King,…}->$U_{context}$*(thesis_about_english_monarchy_of_19th_century)*

The user might later refer to the keywords {fight, duel, horse, knight} and the system could infer the goal:

Kwords{fight, duel, horse, knight}->U_{goal}*(looking_for_images_featuring_duels_of_mounted_knights)*

The above refers to the *knowledge* level of the proposed user model. Lower level features would also need to be represented in the model. For the purpose of this paper, some of the more interesting of these lower level features will be properties of images used by the system to create content-based associations between images that take account of the user's conceptualisation of these similarities. For example, for our history user, the horses in the images are of interest only because they are ridden by 19th century knights. One approach would be carrying out feature extraction on the images to enable the user to indicate the objects of interest (in our example, the horse and the rider, say). The system creates a combined specification of the features related to all of the objects in the image that the user has indicated. This specification is then used by the system as part of its evidence when subsequently retrieving related images in that session. An approach based on the user's notion of similarity can be found in the PicHunter system (Cox *et al.*, 2000)

4.4 A proposal of a generic architecture

A possible architecture for the VIR system is now described. It is argued that the adoption of an intelligent multimodal interface can actually increase the relevance of retrieved items in a visual information retrieval scenario. This leads to the incorporation into the architecture of an agent that is capable of sensing, assisting and conversing with the user during his interaction with the VIR system.

Figure 5 proposes an architecture of a multimodal VIR system. The user interacts with the system through a *user interface agent* plus a typical multimedia information retrieval system in order to formulate a query and perceive the evolving results.

The *Multimedia Content Description* is assumed to be MPEG-7 compliant, and is important to the process of matching the query with the multimedia information. It is also assumed that tools defined by MPEG-7 are available to automatically index multimedia content and extract low-level features in order to populate the multimedia content description database.

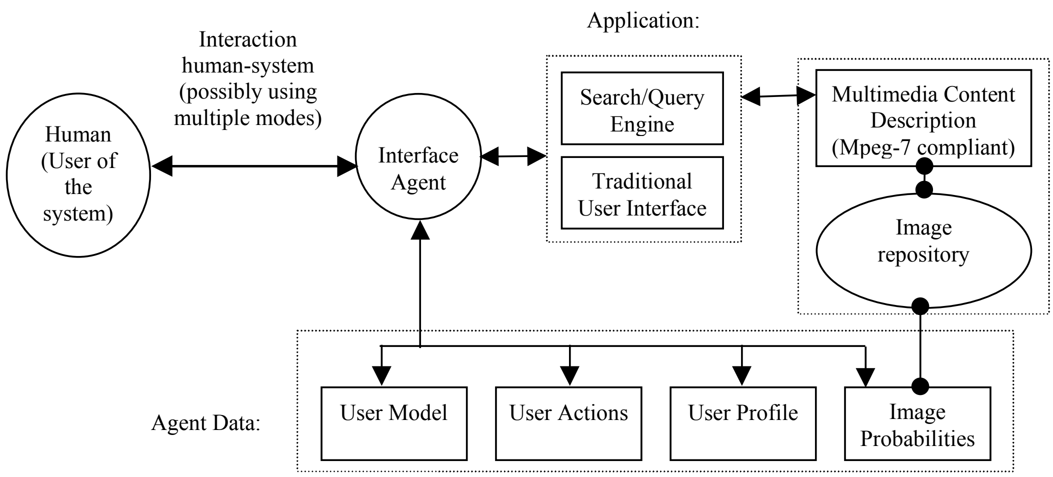

Figure 5 - Proposed application architecture

In Table 3 we can see the types of data that are stored and managed by the VIR system. The first four tables presented are all managed by the agent and are used to support the user modelling capability of the system. The user model is based on Bayesian user modelling techniques and dynamic Bayesian Networks. The last table (Image Descriptors and Image repository) is fixed, i.e., is not to be changed by VIR system but only by the Indexing subsystem (not considered here).

User Model (Bayesian):	Conditional Probability Table 1	CPT 2	…	CPT n
User Actions:	Action 1	A 2	…	A n
Explicit User Profile:	Name	Age	…	Interaction History of other sessions
Probabilities associated with the images:	Image Probability 1	IP 2	…	IP n

Image Descriptors and Image repository:

D11	D21		Dn1
:	:	…	:
D1k	D2k		Dnk
Image File pointer 1	IFp 2		IFp n

Table 3 - Data stored in the system

When the agent is functioning (is on), the search/query engine will use both information gathered from the user during query formulation (from the Application), information from the user model, and session information about the user (the sequence of user actions during that session), to improve the results of the query (Figure 6).

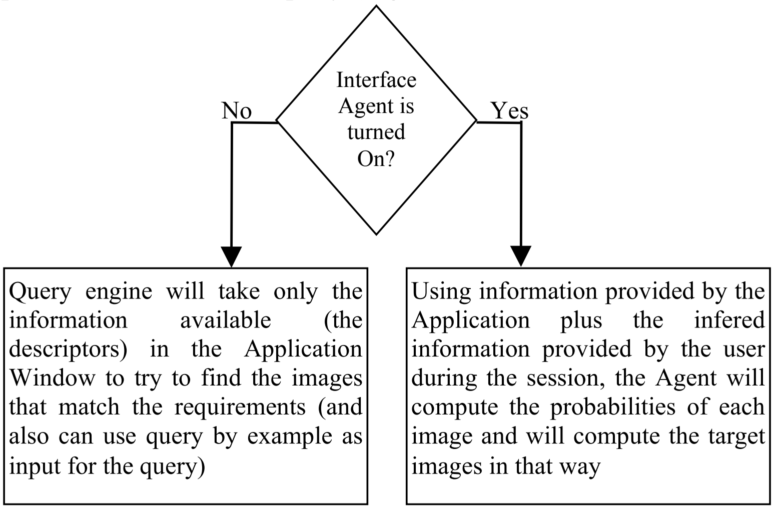

Figure 6 - Behaviour of the System with and without the Agent

In fact, since the relevance of an attribute for a query is not always obvious, it is easier for the system to learn from the user how to best match the expected result. The adoption of machine learning to automatically and dynamically select and combine features to satisfy a query seems to be a promising approach.

4.5 The user interface agent

The *user interface agent* represents an attempt to provide complexity abstraction and intelligent assistance in a self-contained software agent that communicates with the user through the multiple modalities of the user interface. This agent can learn individual user

preferences and tendencies to provide automated assistance, thereby acting as "intelligent assistant".
This agent will adapt to individual users by observing their behaviour, i.e., multimodal interactions with the system, and constructing representations of that behaviour. The agent would then store the behaviour using a dynamic knowledge representation that recognises relationships between behaviours and captures uncertainty in a cognitive (and probabilistic) model of the user.
The user interface agent will act on observed behaviour according to user-defined levels of assistance through a mutual multimodal dialogue.
In the prototype context, the agent tasks can be summarised as:

- To sense all user inputs (including the user inputs into the Application User Interface module);
- To attempt to proactively help the user (initially via a constrained dialogue environment);
- To record all user actions and manage the user profile;
- To update the probabilities associated with each image in the repository during the session.

4.6 Development of the prototype

The prototype application to be developed will be closely linked with a research project being undertaken at Inesc-Porto (David, 1999). The visual data to be used for the purpose of the prototype is that of a photograph archive (Portuguese Photograph Centre).
In the implementation of the prototype, most of the work is expected to be devoted to the development of: inference and learning mechanisms, adaptive search engines, user and domain knowledge bases, discourse models, and user interfaces.
The development environment is expected be based on open architectures and feature operating systems (Linux), programming platforms (Sun's Java, Gnu's C++ compilers and other necessary programming tools) and other software platforms and modules. It is also assumed that MPEG-7 Descriptors, Description Schemes and Description Definition Languages will be standardised and ready to be used in conjunction with tools to automatic index multimedia content and extract low-level features (MPEG Requirements Group, 1999)[7].
Such facilities will be used to populate the MPEG-7 multimedia content description database.
The first approach to the prototype will use a simple architecture that is likely to feature:

- A User interface developed in Java;
- User Model and Image descriptive data stored in a relational database (MS-Access for MS-Windows or MySql for Linux) as presented in Table 3;
- Communication between the application and data through JDBC or ODBC connectivity protocols.

4.7 Estimated Workplan

The current section is devoted to the presentation of the tasks (T) and corresponding deliverables (D) expected for the project.
The associated diagrams that follows gave an estimated timetable for the four years of the project duration.

[7] MPEG-7 standard is expected to be ready during the 2001 year.

List of tasks and deliverables:

T1) Generic bibliographic search and reading in topics as: Visual Information Retrieval, Multimodality, Intelligent Multimedia Information Retrieval, Software Agents, Multimedia Modelling, Learning and Probabilistic Reasoning;

T2) Refinement of the PhD research topic;

T3) Focused bibliographic search and reading;

D3.1) Literature Review Report

T4) Elaboration of the requirements, objectives and detailed work plan for the implementation part of the PhD;

D4.1) Paper presented on a workshop (Torres & Parkes, 2000)

T5) Writing of the First Year Report;

D5.1) First Year Report

T6) Software specification of the first prototype to be developed: probabilistic model, dialogue model, user interface, metadata model, and inference engine;

T7) Development of the first version of the prototype;

D7.1) First version of the prototype

D7.2) Published Paper

T8) Writing of the Second Year Report;

D8.1) Second Year Report

T9) Analysis of the first version prototype and necessary revisions;

T10) Writing of the Thesis part related with literature review;

T11) Development of the prototype;

D11.1) Final version of the prototype

D11.2) Published Paper

T12) Analysis of the prototype and necessary revisions;

T13) Usability tests definition/preparation;

T14) Usability tests;

T15) Analysis of the results of the usability tests;

D15.1) Published Paper

T16) Writing of the Thesis part related with the work developed;

D16.1) Thesis

First Year Workplan:

Task:	Month/Year											
	10	**11**	**12**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
	1999			**2000**								
T1	■	■	■	■	■	■						
T2						■	■	■				
T3							■	■	■	■	■	■
T4									■	■	■	■
T5												■

Second Year Workplan:

Task:	Month/Year											
	10	**11**	**12**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
	2000			**2001**								
T3	■	■	■	■	■	■	■	■	■	■	■	■
T5	■	■										
T6	■	■	■	■								
T7	■	■	■	■	■	■	■	■				
T8											■	■
T9									■	■	■	■

Third Year Workplan:

Task:	Month/Year											
	10	**11**	**12**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
	2001			**2002**								
T3	■	■	■	■	■	■	■	■	■			
T8	■											
T9	■	■										
T10			■	■	■	■	■	■				
T11							■	■	■	■	■	■

Fourth Year Workplan:

Task:	Month/Year											
	10	**11**	**12**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
	2002			**2003**								
T11	■	■	■	■	■	■						
T12				■	■	■						
T13					■	■						
T14							■	■				
T15									■	■		
T16									■	■	■	■

4.8 Conclusion

This section has presented the research aims and work plan of the author's PhD work.
An approach to the problem of Visual Information Retrieval has been presented. In this approach, a generic user model has been discussed. An initial design of the architecture has also been presented.
The work plan presented is provisional, and is expected to be modified as necessary during the course of the project.
The main work for the next few months will concern the development of the initial prototype, to validate some of the ideas. At the same time, the theoretical background will be developed through additional and more focused readings. Time will also be dedicated to the refinement of the expert model of an image retrieval user. This will be carried out through case studies and identification of patterns of behaviour of real users of image retrieval systems.

5 CONCLUSIONS

This document has focused on the description of the various aspects of the author's PhD project. Two major areas were surveyed: Visual Information Retrieval and perception, learning and reasoning in software Agents.

This report stresses the need for user modelling and adaptivity in effective VIR systems. A model based on Bayesian networks and Bayesian user modelling has been sketched out, and has been demonstrated, using a worked example, of how such a model may be applied in a VIR. The example showed how various components of the user model are populated as knowledge of the user is acquired through dialogue with the user and information about the user's actions. The initial design of an architecture for the proposed VIR prototype was also presented.

A provisional work plan has also been presented. Initial results suggest that the plan is realisable in the given timescale.

REFERENCES

Armitage, L and Enser, P G B (1997): *Analysis of user need in image archives*.Journal of Information Science, 23(4), 287-299.

Belkin, Nicholas J. (1997): *User Modeling in Information Retrieval*, Tutorial presented in Sixth International Conference on User Modeling, Chia Laguna, Sardinia, June 1997 (url: www.um.org/um_97/contents.html).

Benyon, D. et al. (1987): *Modelling users' cognitive abilities in an adaptive system*, 5th Symposium EFISS, Plenum Publishing.

Bimbo, Alberto del (1999): *Visual Information Retrieval*, Morgan Kaufmann Publishers, Inc. San Francisco, California, 1999.

Celentano, Augusto, Sabbadin, Stefano, (1999): *Multiple Features Indexing*, European Workshop on Content-Based Multimedia Indexing 1999, Toulouse, France.

Cheyer, Adam, Julia, Luc and Martin, Jean-Claude (1998): *A Unified Framework for Constructintg Multimodal Experiments and Applications*, CMC'98: Tilburg (The Netherlands), pp 63-69.

Cowell, R., Dawid, A. P., Lauritzen, S. L., Spiegelhalter, D. J. (1999): *Probabilistic Networks and Expert Systems*, Springer-Verlag New York, 1999.

Cox I.J., Miller M.L., Mink, T.P. Papathomas T.V. & Yianilos P.N. (2000): *The Bayesian Image Retrieval System, PicHunter: Theory, Implementation and Psychophysical Experiments*. IEEE Transactions on Image Processing, 2000.

Croft, W.B. (1995): *What Do People Want From Information Retrieval?* in d-lib Magazine , November, 1995.

Damasio, A. R. (1994): *Descartes' Error: Emotion, Reason, and the Human Brain*. New York, NY: Gosset/Putnam Press, 1994.

David, Gabriel (Project's leading investigator) (1999): *Project title: MetaMedia - Meta-Informação para Arquivos Multimédia (Reference: PRAXIS/C/EEI/13208/98)*, primer institution: INESC - Porto, project funded by the Science and Technology Foundation (url: www.fct.mct.pt).

Dutton, David M. and Conroy, Gerard V. (1996): *A review of machine learnin*g, The Knowledge Engineering Review, Vol. 12:4, 1996, pp. 341-367.

Eakins, John P.; Graham, Margaret E. (1999): *Content-based Image Retrieval - A report to the JISC Technology Applications Programme*, Institute for Image Data Research, University of Northumbria at Newcastle, January 1999 (www.unn.ac.uk/iidr/research/cbir/report.html).

Ellis, David (1990): *New Horizons in Information Retrieval*, The Library Association, London, 1990.

Harman, D. (1992) *Ranking Algorithms*, in Frakes, William B. and Baeza-Yates, R. (Eds.) *Information Retrieval - Data Structures and Algorithms*, Prentice-Hall, 1992.

Hayes, P., Reddy, D., (1983): *Steps towards graceful interaction in spoken and written man-machine communicatio*n, International Journal of Man-Machine Studies, No. 19, pp. 231-284.

Hearst, Marti (1999): *User Interfaces and Visualization* in Baeza-Yates, R. and Ribeiro-Neto, B. (Ed.), *Modern Information Retrieval*, ACM Press, New York, Addison-Wesley, 1999, pp. 257-323.

Horvitz, E., J. Breese, D. Heckerman, D. Hovel, and K. Rommelse (1998): *The Lumiere Project: Bayesian User Modeling for Inferring the Goals and Needs of Software Users.* Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence, July 1998.

Kowalski, Gerald (1997): *Information Retrieval Systems: Theory and Implementation*, Kluwer Academic Publishers, 1997.

Kay, Jude (1995): *Vive la difference! Individualised interaction with users*, IJCAI'95, Montreal, August, 1995.

Langley, P., Thompson, C., Elio, R., Haddadi, A., (1999): *An Adaptive Conversational Interface for Destination Advic*e, in Proceedings of the Third International Workshop on Cooperative Information Agents, Uppsala, Sweeden: Springer, July, 1999 (http://www.isle.org/~langley/papers/place.cia99.ps).

Levkowitz, Haim (1997): *Color Theory and Modeling for Computer Graphics, Visualization, and Multimedia Applications*, Kluwer Academic Publishers, Boston/Dordrecht/London, 1997.

McTear, M. (1993): *User modelling for adaptive computer systems: a survey of recent developments*, Artificial intelligence review 7, 157-184.

Michalski, R. (1994): *Inferential Theory of Learning*, *in Machine Learning: A Multistrategy Approach*, Michalski, R. and Tecuci, G. (Eds.), vol 4, Morgan-Kaufmann, 1994.

Minka, Thomas (1996): *An Image Database Browser that Learns From User Interaction*, MIT Media Laboratory, Cambridge MA, Master of Engineering Thesis, technical report TR #365.

MPEG Requirements Group (1999): *MPEG-7 Context, Objectives and Technical Roadmap*, Doc. ISO/MPEG N2861, MPEG Vancouver Meeting, July 1999.

MPEG MDS Group (2000*): MPEG-7 Multimedia Description Schemes XM (Version 2.0)*, Doc. ISO/MPEG N3246, MPEG Noordwijkerhout Meeting, March 2000.

MPEG Requirements Group (2000*): MPEG-7 Overview (Version 2.0)*, Doc. ISO/MPEG N3349, MPEG Noordwijkerhout Meeting, March 2000.

Niblack, W., R. Barber, W. Equitz, M. Flickner, E. Glassman, D. Petkovic and P. Yanker (1993): *The QBIC Project: Querying images by content using colour, texture and shape*, in SPIE 1908, Storage and Retrieval for Image and Video Databases, February.

Ornager, S (1997): *Image retrieval: theoretical and empirical user studies on accessing information in images* in ASIS '97: proceedings of the 60 th ASIS Annual Meeting, vol. 34, 202-211.

Oviatt, Sharon (1999): *Ten Myths of Multimodal Interaction*, Communications of the ACM, November 1999/Vol. 42, No. 11.

Pentland, A., R.W. Picard and S. Sclaroff (1993): *Photobook: Tools for Content-Based Manipulation of Image Databases*, Technical Report no. 255, MIT Media Lab Perceptual Computing Section.

Picard, R. W. (1995): *Affective Computing*, M.I.T. Media Laboratory Perceptual Computing Section Technical Report No. 321, November, 1995.

Rasmussen, E. (1992) *Clustering Algorithms*, in Frakes, William B. and Baeza-Yates, R. (Eds.) *Information Retrieval - Data Structures and Algorithms*, Prentice-Hall, 1992.

Rich, Elaine and Knight, Kevin (1991): *Artificial Intelligence - second edition*, international edition by McGraw-Hill, 1991.

Rosenblatt, F. (1962): *Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms*, Washington, D.C.: Spartan Books, 1962.

Roy, D., Sawhney, N., Schmandt, C., Pentland. A., (1997): *Wearable Audio computing: A Survey of Interaction Techniques*, Technical Report No. 434, MIT Media Lab, April, 1997 (http://www.media.mit.edu/~nitin/NomadicRadio/AudioWearables.html)

Rui, Yong; Huang, T. S.; Ortega, M.; Mehrotra, S. (1998): *Relevance Feedback: A Power Tool for Interactive Content-Based Image Retrieval*, IEEE Transactions on Circuits and Systems for Video Technology, Vol. 8, No. 5, September 1998, pp. 644-655.

Russel, S., Norvig P. (1995): *Artificial Intelligence, A Modern Approach*, Prentice-Hall International Editions, 1995.

Stein, A.; Gulla, Jon A.; Muller, A.; and Thiel, U. (1997): *Conversational Interaction for Semantic Access to Multimedia Information*, in Maybury, M. T. (Ed.), *Intelligent Multimedia Information Retrieval*, AAAI Press/MIT Press, Menlo Park, California/Cambridge, Massachussets/London, England, pp. 399-421.

Salton, Gerard and McGill, Michael J. (1983): *Introduction to Modern Information Retrieval*. McGraw-Hill Book Company, 1983.

Shneiderman, Ben (1997): *Designing the user interface: strategies for effective human-computer interaction*. Addison-Wesley, Reading, MA, 1997.

Stock, O. and The AlFresco Project Team. 1993. *AlFresco: Enjoying the Combination of NLP and Hypermedia for Information Exploration* In Intelligent Interfaces, Maybury, M. T. (Ed.), AAAI Press/MIT Press, Menlo Park, California/Cambridge, pp. 195-224.

Torres J.M. & Parkes A.P. (2000): *User modelling and adaptivity in visual information retrieval systems*. Workshop on Computational Semiotics for New Media, University of Surrey, Surrey, UK, June 29-30, 2000. http://www-scm.tees.ac.uk/users/p.c.fencott/newMedia/

Trower, T, (1997): *Creating Conversational Interfaces for Interactive Software Agents* , In Proceedings of the Conference on Human Factors in Computing Systems (CHI '97), Tutorial 31. Atlanta, Georgia, ACM Press, March, 1997.