Visual Information Retrieval through Interactive Multimedia Queries

Former Title:

Intelligent Multimodal Interfaces for Visual Information Retrieval

PhD Second Year Report

José Manuel de Castro Torres

Supervisor: Dr. Alan Parkes

Computing Department

Faculty of Applied Sciences

Lancaster University

March 2002

Table of Contents:

List of Tables:

1 INTRODUCTION

Recent years has seen the rapid development of tools to effectively process digitised visual data. Many applications, ranging from cultural repositories to medical or stock photograph archives have also appeared. Visual information retrieval is thus a very important research area (Bimbo, 1999). The current project focuses on interactive query formulation in Visual Information Retrieval (VIR) systems.

Thus far, the work has focussed on the development of models of user interaction, on the incorporation of new types of multi-sensorial input, and on bridging the gap between low-level image descriptions and semantically-oriented descriptions in VIR systems. The second year has seen the development of a much more focused work programme, dedicated to developing techniques for interactive multimedia query formulation in VIR systems. It was decided that the original focus of the work was too broad. However, the current focus includes elements of the three areas of investigation mentioned immediately above. The title of the project has been changed to reflect the more precise new direction of the work.

2 RESEARCH DESCRIPTION

2.1 Overview of Main Goals

VIR has seen the proliferation of techniques for automatically extracting relevant features from visual material. This has led to so-called Content Based Image Retrieval Systems (Eakins, 1999; Niblack, 1993; Pentland, 1993; Rui, 1999). Such advances have made it possible to index large collections of digital visual data automatically, a process that hitherto required manual effort.

However, despite such advances in VIR, there are still many challenges in trying to deploy more user-oriented systems, largely because the output from image analysis processes continues to consist of mainly low-level features (regions, colour, texture, etc). There is a considerable gap between such features, which tend overwhelmingly to support a "Query by example" approach, and the ways in which people tend to classify and refer to images. Ideally, then the process of composing a query needs to involve collaboration between the user and the system, and the VIR system should interactively assist the user in specifying a query in ways that bridges the gap between the low level descriptors and the semantically – oriented descriptions favoured by humans. Moreover, we believe that the queries themselves can be used to augment the image descriptions (which may be different for different users, or because users have different perspectives on meaning of the images), and thus enrich the indexing scheme available to the VIR system.

In the above respect, the work remains true to a claim made in the first year report, *viz.* "*an ideal VIR system would reflect the best characteristics of both computer and human*".

The central aim of the current project can be stated as:

The incorporation into a VIR system of techniques for interactively associating conceptual descriptions with automatically extracted visual information, along with an ability to represent and reason about the user will enhance VIR systems and bring us closer to the hypothetical "ideal" VIR system.

It can be seen from the preceding statement that user modelling is of crucial importance to the project. The success of the interaction between an end user and the VIR system partly depends on the way that the system acquires and reasons with the available information provided explicitly or implicitly by the human user.

2.2 Research Aims

The research aims of the work are to:

- Analyse contemporary content-based image retrieval solutions, and characterise their weaknesses in terms of user-orientation ability to deal with image semantics
- Survey the trends in content-based image retrieval and, determine the techniques required to increase and enrich the interactivity and expressive power of such systems
- Develop techniques that reduce the semantic gap between human-oriented descriptions of visual information and feature-level automatically extracted information. Apply the techniques to enrich both human-system interaction and image descriptions in a prototype VIR system.
- Provide the system with the ability to acquire information about the user (related both to user type and individual user characteristics), and use this information to adapt the visual information retrieval to the individual user's requirements.
- Incorporate the above into a VIR prototype system that supports the user in creating multimedia queries composed of user-annotated images.

For the proposed work, the exploration of the relationship between textual information (the conceptual level) and visual information (the content level) is of crucial importance as described in the following section.

2.3 Proposed Approach

One characteristic of human beings is the ease with which they can maintain mental representations of concepts and conceptual relationships. Many concepts can, of course, be associated with physical and visual entities.

A structure that connects the textual labels to the visual information can help in the reduction of the semantic gap between the concept space and the low level features extracted from the multimedia content

One of the drawbacks of using textual descriptions in VIR systems is that these are usually manually applied, resulting in the following:

- The textual information associated with visual data tends to be subjective and dependent of the human annotator, especially if the annotations are not subject to any restrictions. For example, if we ask 10 people to describe the content of a photo, we would certainly almost certainly obtain 10 significantly different descriptions;
- The manual annotation of large visual repositories is very costly in terms of time and human resources.

Despite these cons, the use of textual information during the search or retrieval of visual information allows the user to express his query in a much more semantically richer way. Moreover users may describe images in different ways, not least because the description depends to a large extent on the *purpose* for which an image is required by the user. We propose that the (textual) statements a user might make about a retrieved image, i.e. about the properties of the image that made it relevant as a result of the initial query, could, if linked to the image, become part of the systems representation of that image. These enriched descriptions would enable the same user (and perhaps other users) to subsequently query the VIR system with much richer, mixed text and image queries.

In the proposed approach to the above problems, a knowledge base that links a textual thesaurus with a visual object database plays a central role. This knowledge base, essentially

as "*Visual Thesaurus*", will be used as the user interactively indexes retrieved images (during interactive query formulation).

The VIR system will use the visual thesaurus to assist the user in the construction of a visual template that captures the most important features of the visual content desired. These captured features can be semantically meaningful features, such as traffic on a motorway, or more related with the content, such as a landscape featuring. The visual template obtained is primarily composed of visual objects, gathered from the Visual Thesaurus, that are indicated, explicitly or implicitly, as relevant by the end user. The visual template will be personalised by individual users with the help of an image annotation tool.

Ideally, the system should be capable of assisting the user in the construction of queries supported by textual expressions such as: "I am looking for photos featuring a clean blue sky with the sun on the left side and featuring some trees and houses"

Once the user has completed the initial version of the visual template, the VIR system will search the visual database and present the set of images that satisfy this query. This search will not only consider automatically extracted features but will also consider any text annotations (including those previously applied by users engaged in interactive query formulation).

The automated extraction of feature-based descriptions is not a research aim of the work. The prototype will therefore incorporate existing techniques, such as the perceptual features of visual content defined in the emerging MPEG-7 standard (MPEG Requirements Group, 2000).

Once the initial search (described above) has been performed, the next phase is the interactive query formulation cycle. There are two main features of this cycle. Firstly, the user provides feedback to the VIR system indicating the degree of relevance of each item displayed. An outcome of this aspect of the interactive process (the other aspect being the annotation process described above) is the enrichment of the system's representation of the images.

Typical operations that a user could perform after the creation of the first version of the visual template are:

- Selecting which images are the most relevant from the result set presented (feedback information: unordered or ordered selection of one subset of the display set) see Cox, (2000) for an example;
- Selecting one or more groups of relevant images within the result set (feedback information: selection of one or more subsets of the display set);
- Associating each image of the result set with values for evaluation parameters related to the visual template constructed (feedback information: selection one or more subsets of the display set according to value);
- Validation of textual associations provided by the VIR system;
- Association of textual information with one or more images of the result set, using terminology from the thesaurus.

A typical session of visual information retrieval performed by an end user may involve the following sequence of activities:

1. The VIR system identifies the user
2. The user and the system interact with each other to compose the first version of the query

2.1 The VIR aids in the construction of a query using the available descriptors. One of the results expected from this interaction stage is a visual model that will be used to query the repository through low-level descriptors.

3. The VIR displays the initial results to the user.
4. The user interacts with the system to indicate which results are the more relevant and providing more detailed information such as which part of the query a given image relates to.
5. The user and system then work together to improve the quality of the results of the visual retrieval session. However, as side affect of this is that the VIR system augments its image descriptions.

3 PROGRESS REPORT

3.1 Main Achievements

A VIR prototype, currently under development, will serve to demonstrate, evaluate, and refine the research aims described above. The following tasks, proposed in the first year report, are completed or almost finished, and are highly relevant to the current work, despite the different initial overall aim:

- Literature Review Report in: Visual Information Retrieval, Multimodality, Multimedia Modelling, Agents, Learning and Probabilistic Reasoning
- Elaboration of the requirements of an "ideal VIR system" and presentation of an initial approach (Torres & Parkes, 2000)
- Implementation of a Java software framework for the prototype
- Specification and design of the following components of the prototype: probabilistic model, dialogue model, user interface, metadata model, and inference engine, and subsequent refinement of the model (Torres & Parkes, 2001)
- Implementation of a relevance feedback mechanism using a an interaction model that is a simplified version of that discussed in section 2.3
- Implementation of a software framework in Java for indexing (on the basis of automatic feature extraction) image content
- Implementation of a simple Java application for indexing (on the basis of automatic feature extraction) of image content. At the present time, only basic image processing techniques, such as colour analysis, are being considered.
- Selection of the textual thesaurus to be incorporated into the prototype. Initial investigations have focused on WordNet (Miller, 1990), though other thesauri are being considered.

3.2 Future Developments

The following developments are ongoing:

- Adaptation of the textual corpus into the prototypical VIR architecture
- Creation of the model of knowledge structure to relate the textual information to the visual information
- Compare various techniques to gather implicit, explicit or common user information
- Evaluate the prototype through task-based trials and comparative qualitative evaluation with other systems

3.3 Estimated Work Plan for the remaining time

The current section presents the task (T) and deliverables (D) of the research project. Table 1 summarizes the work carried out thus far. Table 2 provides the work plan for the remaining duration of the project.

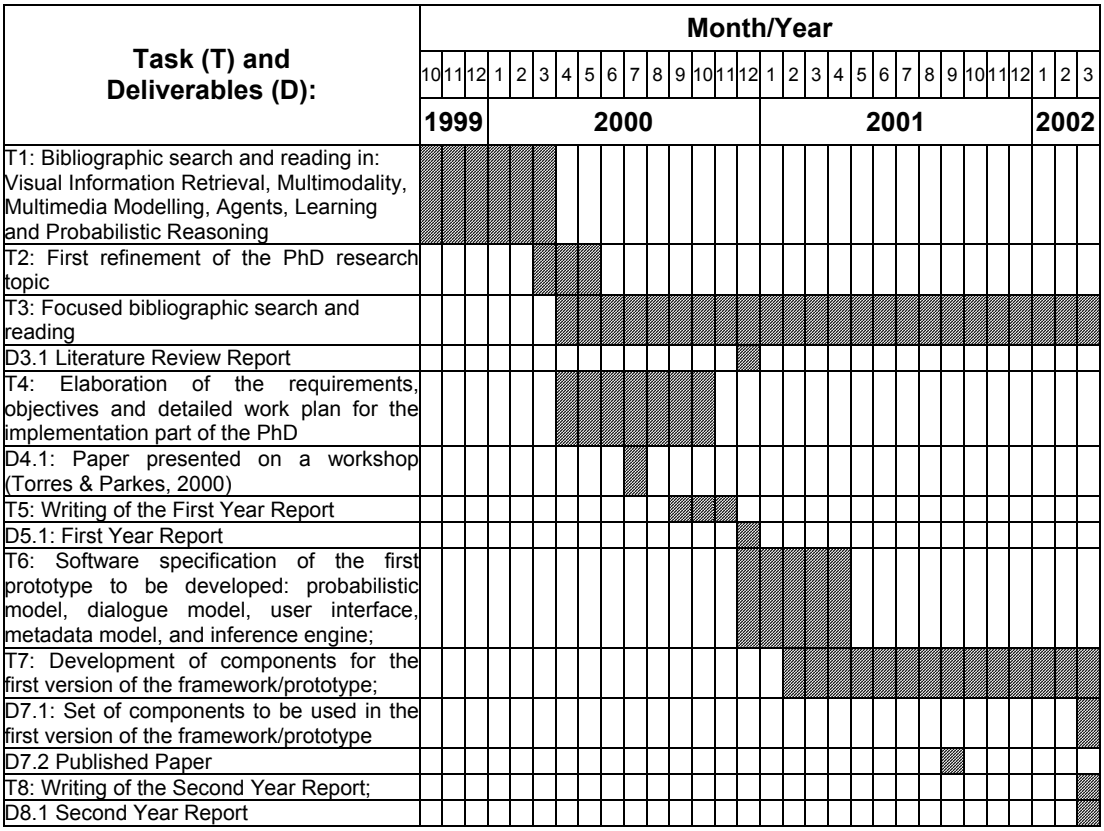

Table 1 - Completed Work-Plan Timetable

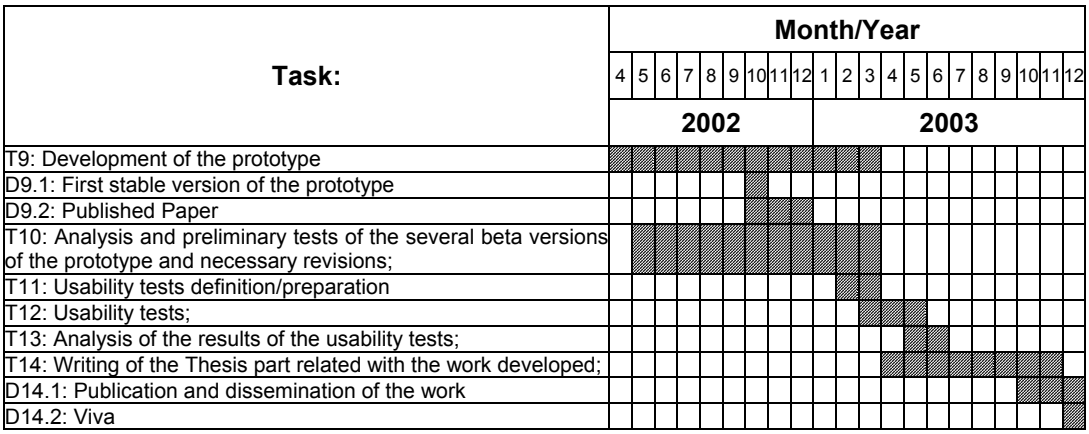

Table 2 - Estimated Work-Plan Timetable for the remaining time

4 CONCLUSIONS

This document has described a refinement of the original research aims and describes the achievements of the first two years of the research. It has discussed developments carried out in the implementation of the VIR architecture and the correspondent prototype that will be used to validate the proposed solutions in the field of Visual Information Retrieval. It has also presented a work plan for the remaining phase of the research.

References

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

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).

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.

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

Miller, George A., Richard Beckwith, Christiane Fellbaum, Derek Gross and Katherine J. Miller (1990): *Introduction to WordNet: an on-line lexical database* In: *International Journal of Lexicography* 3 (4), 1990, pp. 235 - 244.

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 (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.

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.

Rui Y., Huang T.S. & Chang S.F (1999): *Image Retrieval: Current Techniques, Promising Directions and Open Issues*, Journal of Visual Communication and Image Representation, Vol. 10, March 1999, pp. 39-62.

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.

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/

Torres, J.M., Parkes, A.P. (2001): *Intelligent multimodal interfaces for visual information retrieval*, ECDL-2001, September 2001, Darmstadt-German.