Intelligent Multimodal Interfaces for Visual Information Retrieval

J. M. Torres and A. P. Parkes
Multimedia Content Group
Computing Department
Lancaster University
{torres@comp.lancs.ac.uk, app@comp.lancs.ac.uk}

ABSTRACT

This paper discusses the Visual Information Retrieval (VIR) process and emphasizes the need for more natural interfaces and a better understanding of the user. The central role of user modelling and the importance of the communication channel in a VIR scenario are discussed. We identify areas of research and possible approaches for dealing with the problem of creating effective, user responsive VIR systems.

Introduction

Recent years have seen a rapid increase in the number and size of digital image repositories. Content-based image retrieval (CBIR) systems, which have appeared primarily in the last decade, feature image indexing and subsequent search on the basis of properties that are automatically extracted from the images (see Rui *et al*, 1999).

However, such CBIR systems mainly reflect the "query by example" approach, and thus do not adequately reflect end user requirements. There is a semantic gap affecting the communication between the user and the system. This fact demands new characteristics that should, in the future, be considered for emerging VIR systems.

A user-oriented approach to VIR

In typical content-based visual information retrieval (VIR) systems the user expresses a query by direct reference to low level features that the retrieved images should contain. For example, the user may specify coloured regions that the retrieved image should feature in certain locations of the image. The search engine simply matches and ranks, according to some mathematical measure of similarity, the features specified by the user to images in the image repository (see Bimbo, 1999). This approach does not reflect the user's concept space.

Many aspects of image semantics are more to do with the person who experiences the image that the image itself. This factor needs to be taken into account when building VIR systems. A VIR system needs to manipulate a representation of the user, i.e., a *user model* should play a crucial role in image retrieval.

A further perspective on VIR presupposes an *agent* (the human user of the VIR system) who communicates via mechanisms such as natural language, gestures, facial or corporal expressions, predetermined actions, etc. Such an agent has a structured memory, key parts of this being mental images, associations between concepts and sensations and/or experiences. This perspective forms the basis of our approach.

Most systems have no way of taking account of user characteristics such as the user's memory, knowledge, preferences and goals. Moreover, the channel of communication between the user and system is usually of a restricted form that is imposed by the system. It does not permit the types of communicative acts preferred by the user. Ideally, communication between a user and a VIR system should incorporate the best characteristics of both participants. In particular, the following communicative dimensions would be highly desirable:

- The adoption of a more natural dialogue through multimodal input and output;
- The acquisition of knowledge about the user, and the use of such knowledge to improve the results of queries for that user, and tailor the system and its interface for the user.

The "User Assistant" in VIR

Typical VIR systems operate according to what can be called the *pull model* of information retrieval. The user has sole responsibility for the specification of queries, and the system provides little user-sensitive assistance in this area. An alternative to this, suggested by the above observations regarding user-system communication, is the *push model*. Here, the system is more proactive, and becomes the user's *partner*, rather than *idiot savant*, as it were. This model presupposes that the VIR system learns how to detect the user's information needs and suggests and delivers information to the user accordingly. However, we do not advocate a totally push model approach; we see the most fruitful combination being yielded by a mixed initiative communication model.

The key role of User Modelling

As can be appreciated by the preceding discussion, we assert that a successful VIR will need to observe, analyse and act upon the user's patterns of behaviour. In order to design techniques with which a system can achieve this, it is useful to consider what types of behaviour are typical in VIR scenarios.

Studies described by Eakins and Graham (1999) indicate that the patterns of behaviour of VIR users depends on the users background (e.g. the user's professional activity). This background can influence such things as the terminology favoured by the user.

Contextual information can also influence the way that the user implicitly elaborates his or her image retrieval strategy. Moreover, due to a change in context, users frequently adopt different strategies within a given session.

As Eakins and Graham point out, there is strong evidence that different types of users require different styles of interaction with retrieval systems. A better understanding of how and why people use images can support the identification of distinct user types and their needs. The result will be the ability to design more effective VIR systems.

Towards effective VIR systems

There is clearly a need for research on the topics outlined above. In particular, in our own work we are considering the following:

- Developing techniques through which the system can acquire information about the user, and use this information to adapt the VIR process to the individual user's requirements.
- Providing the VIR with a multimodal interface, enabling the use of a combination of text, graphics, audio, video, etc., in the formulation of queries
- Prototyping and evaluating conversational assistants for the multimodal VIR interface.
- The modelling of both objective (low-level) and subjective (high-level) features of images, through the use of knowledge based techniques

Our approach to the first topic is the adoption of unobtrusive and probabilistic models of the user (as in Cox *et al*, 2000). We use Bayesian Networks (Horvitz *et al*, 1998) and the relevant existing inference mechanisms. The guiding principle is that the VIR system analyses the user's actions during the interaction, and this information is used to infer the user's goals and subsequently the images that best match the user expectations (Torres & Parkes, 2000):

$$\text{Observation}(U_{actions}) \rightarrow \text{Inference}(\text{U}_{\text{goals}}, \text{I}_{retrieved})$$

The user assistant, along with a natural language interface and gestural input represent the multimodal interface. The assistant's function is to decrease the communicative gap between the VIR system and the user. This requires information structures to map conceptual structures onto visual information. During interaction, the user assistant infers the main concepts that are relevant to the user and suggests the elementary visual objects that can be used to interpret content-based queries via the aforementioned information structures.

References

Bimbo A. 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*. To Appear in IEEE Transactions on Image Processing, VOL. 20, 2000

Eakins J. P.; Graham, M.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 D. & Rommelse K. (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.

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.

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 of 2000.

Intelligent Multimodal Interfaces for Visual Information Retrieval

J. M. Torres and A. P. Parkes
Multimedia Content Group
(url: www.comp.lancs.ac.uk/computing/research/mcg/index.php)
Computing Department
Lancaster University
{torres@comp.lancs.ac.uk, app@comp.lancs.ac.uk}

The Context

- There is a growing need for easy access to multimedia repositories in general and image databases in particular
- Visual Information Retrieval (VIR) systems have evolved from traditional annotation-based systems to content-based systems

The Problems

- But these systems are still lacking in the interaction stage, i.e., they cannot perceive the user's goals and the user finds it difficult to understand the system and exploit its full potential

The Ideal Visual Information Retrieval System

- An ideal VIR system would reflect the best characteristics of both computers and the humans
- Desirable additional characteristics of future VIR systems are:
  - More natural and powerful human-computer interfaces using multiple modes of input/output (multimodality);
  - Using knowledge about the user 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.

The Related Technologies/Research Areas

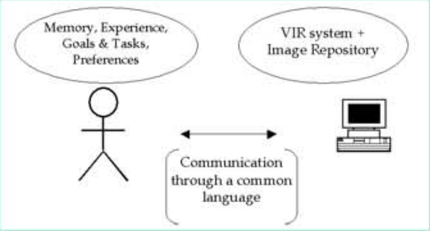
- Bayesian User Modelling and Inference
- Multimodality and Natural Interfaces
- Multimedia Description: MPEG-7

The Aims of this Research

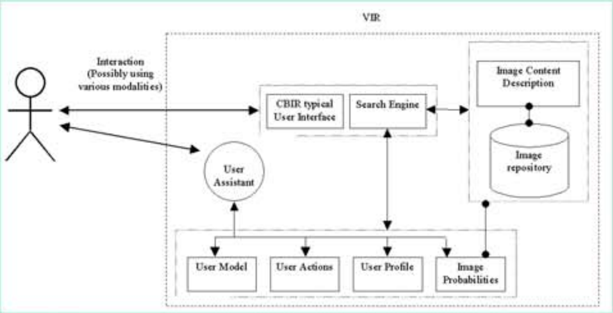
- Develop techniques to enable the system to acquire information about the user, and use this information to adapt the information retrieval to the individual user's requirements.
- To develop multimodal interfaces to enable the user to communicate to the system the types of information he or she wishes to retrieve (using at the very least a combination of text, graphics, audio and video in the query formulation).
- Prototyping and evaluation of conversational assistants for multimodal interfaces.
- Model of both objective (low-level) and subjective (high-level) features of the multimedia content, through the use of knowledge based techniques

A perspective on VIR

VIR can be considered as a communicative act between the human agent and the VIR system:


Memory, Experience, Goals & Tasks, Preferences
VIR system + Image Repository
Communication through a common language

The architecture of the VIR prototype


VIR
Interaction (Possibly using various modalities)
CBIR typical User Interface
Search Engine
Image Content Description
Image repository
User Assistant
User Model
User Actions
User Profile
Image Probabilities

An example

In this example a history researcher is researching images about the medieval history of the UK. 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]-> Uprofile(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,….]->Ucontext(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]->Ugoal(looking_for_images_featuring_duels_of_mounted_knights)*

Following this dialogue with the system, the system suggests that constructs a sketch with knights and horses using the content-based image retrieval tools. The system could propose a first version of the content-based query by providing sketches of the knights and horses (using elementary visual objects from a visual database that serves as the basis of constructing a visual query), the correct textures and colours and suitable spatial arrangements of the various objects.

The interface agent

A user assistant that communicates through a natural language style of interaction

The user modelling approach for VIR systems

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

- 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, Si, 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:

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

Bayesian Networks to model statistical dependencies

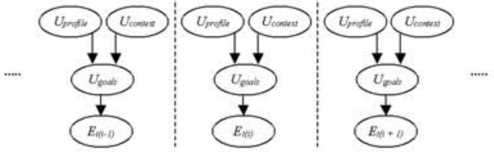
Temporal Dependencies

The arguments of the function f evolve over time, so a mechanism is required to deal with this. 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 timestamped 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.


$U_{profile}$
$U_{context}$
U_{goals}
$E_{t(i-1)}$
$E_{t(i)}$
$E_{t(i+1)}$
.....
.....