Verbal Retrieval of Pictorial Information

Alexander Koutamanis
Delft University of Technology
Faculty of Architecture
Delft
The Netherlands

ABSTRACT

The proliferation of on-line image databases and the utility of these images have triggered the development of content-based techniques for indexing and retrieval. Most techniques are characterized by a verbal interpretation of visual patterns for query formulation. The paper describes the integration of such verbal terms for architectural pictorial information in AZILE, a natural language interface that operates through a dialogue with the user. In this dialogue the user expresses queries as normal everyday utterances. These are parsed and matched to a thesaurus of architectural terms and concepts. The meaning and associations of these terms result into a preliminary fuzzy classification of available pictorial information. The purpose of AZILE is three-fold. Firstly, it serves as an incremental refinement of the query. Secondly, it facilitates direct retrieval of suitable information in a browsing fashion. Thirdly, it supports machine learning by automatically indexing of the images with the terms identified in the user's utterances.

1 INDEXING AND RETRIEVAL OF PICTORIAL DESIGN INFORMATION

The proliferation of multimedia in computerized information has reinforced the significance of pictorial architectural documentation. The ability to create new digital visualizations and to transcribe existing analog static and dynamic images to the computer is resulting into an unparalleled wealth of visual documents that describe existing buildings, new designs and all kinds of architectural concepts, from building type schemata to illustrations of ergonomic requirements in space. Moreover, these descriptions occur at a variety of abstraction and specificity levels, varying from free-hand sketches of a building layout to large-scale dimensioned construction details and from two-dimensional line drawings to interactive animations complete with photorealistic lighting, colour and texture. The number, size and variety of such documents is evident both on the computer of an architect and on the world wide web. The Internet promises a worldwide information system, capable of uniting different sources and types of original, up-to-date and directly usable information. At the same time, design and office automation offers tools for registering and processing information in various forms and modes, as well as facilities for correlating this information and its carriers. That the promised integration, transparency and utility have yet to be achieved is purely a matter of poor choices and limited effort in the past.

The ability to generate, disseminate and retrieve information is increasingly within the reach of a large number of people who are less bounded by constraints

traditionally associated with information channels such as periodicity and censorship. The Internet is consequently portrayed as a worldwide system that will ultimately unite different sources and types of information into a ubiquitous infrastructure. What makes this infrastructure particularly appealing is that it comprises original, up-to-date information. Rather than evaluating the quality of information by the authority of the channel we are promised the possibility to evaluate the reliability and quality of the actual source (Koutamanis 1998). Moreover, we are offered ample opportunity to use this information directly and unobtrusively, e.g. by integrating the latest building components from an online database or by evaluating the conformity of a design to building codes and regulations (Koutamanis 1997).

Straightforward production and efficient dissemination of information are the hallmarks of computerization today, especially in popular areas such as entertainment. In technical and professional areas like architecture and building development is less spectacular but nevertheless sufficient as an indication of the future. Publication of design and product documentation on the Internet is available through most drawing and modeling programs but the purpose of publication generally does not go beyond communication. In particular, pictorial information is frequently treated as mere illustration embedded in dominant verbal documents that form the primary focus in indezing and retrieval. In terms of integration such practices are seemingly efficient and save time and labour. However, when considered in relation to consistency and utility, they generally fail to provide structural solutions.

Indexing and retrieval mechanisms are already an integral part of the Internet. Link lists and search engines are routinely used for retrieving information. If the query is successfully formulated, *recall* (the ratio between the number of relevant documents retrieved and the number of all relevant documents in the system) is generally high, despite the usual frustration of obsolete links. On the other hand, *precision* (the ratio between the number of relevant documents retrieved and all documents retrieved by a specific query) is lowered by the lack of clear terminologies and poor contextual matching but remains mostly adequate. Textual search engines are increasingly refining their query strategies, e.g. by integrating elements of vocabulary control, improving indexing efficiency and providing relevance feedback. Most such refinements presuppose extensive pre-processing of information.

In comparison to alphanumeric information, indexing and retrieval of pictorial information has attracted less interest than it deserves. In information generation and indexing images are treated as empty vessels, scarcely ever annotated with textual indexing terms. Consequently, image retrieval was until recently largely ignored as an imprecise, vague area. The current attention on content-based image techniques is helping reconsider imagery as information (Lopes 1996). Using image-processing and pattern-recognition techniques content-based approaches derive features from low-level attributes such as color, texture and shape directly from the image. These features form the basis for computing image distance measures (i.e. similarity) between user input and the images in a database.

In terms of efficiency and reliability content-based approaches appear to offer

significant advantages over standard databases that depend on keywords that describe image features. Indexing an image database with keywords requires large numbers trained human classifiers and unambiguous but flexible keyword conventions. However, content-based systems may also require extensive human effort in e.g. manual image segmentation (as in QBIC) or annotation (as in Chabot). Another advantage is the ability to form queries through visual interfaces, e.g. by selecting from canonical or typical images in order to retrieve the images of a particular category. However, little has been done in terms of demarcation, i.e. distinction between different types and subjects of pictorial documents.

A severe limitation of many content-based systems in their application to architectural drawings is reliance on global, low-level features such as color. Perceptual similarity between two photographic images could rely on such features but in architecture shape is generally more important, as it relates to typical architectural queries concerning type, style and spatial arrangement. This presupposes automated image segmentation, on the basis of not only color (as in VisualSEEK or Photobook) but also shape. This recovers tentative objects that combine to form the image. For example, image segmentation can recover spaces and building elements in a floor plan. The image is generally described in terms of such objects, their interrelationships and their overall configuration.

Analysis of vector or pixel images can recover important features, such as the overall shape of a floor plan or a space. For practical reasons, however, image analysis and content-based retrieval are generally restricted to a small number of salient features. This derives from a conscious effort to minimize the keywords used for indexing. Such features and keywords form the basic classification criteria for the image in the framework of specific indexing and retrieval activities. Consequently, content-based image indexing may degenerate into a deterministic search for the most economical collection of features (Gong 1998). Even when the objective is just the vectorization of a pixel image, efficiency dictates that issues that might be of relevance such as drawing style are ignored because they are deemed secondary in many searches (Baird, Bunke et al. 1992).

In architectural research analysis of pictorial information is primarily linked to two main areas of investigation, design generation and protocol analysis. Design generation traditionally places emphasis on representation and makes intensive use of frequently pictorial representations. These representations encapsulate the form of the design and related design actions or decisions. This means that the representation used by a generative system provides a comprehensive and consistent coverage of the design classes that can be produced by the system. Consequently, use of the representation for describing these design classes amounts to a simple reversal of the purpose of the representation (Koutamanis 2000; Koutamanis 2000). The resulting reversed representations provide the formal means for identifying the existence of relevant parts in a design, as well as their location.

Protocol analysis addresses similar issues to design generation but generally beyond the confines of specific design classes and without resorting to holistic

representations. Instead, the actions and products of design sessions are captured on conventional media that complement the formal session record or video recording. In recent years protocol analysis has been enriched with tools for capturing design actions in comprehensive multimedia repositories (McFadzean 1999; Gross, Do et al. 2001). Such repositories support the indexing and retrieval of specific instances and occurrences using pictorial forms and interfaces for both dynamic and static images. The analysis and retrieval supported by such tools goes beyond the possibilities offered by groupware and project management systems, where the priority is document management rather than information analysis, processing and correlation.

2 THE ROLE OF VERBAL INTERFACES

Most content-based image indexing schemes, including systems for image parsing, analysis and matching, are characterized by a verbal interpretation of visual patterns for query formulation. This is due to two main reasons. The first is the precedence of alphanumeric information processing in the computer. Existing interfaces, database management software and input hardware are predominantly oriented towards the manipulation of alphanumeric information. Scaling down the complexity of pictorial information to alphanumeric indexing terms means a significant reduction in development and usage costs, as well as connectivity with existing systems.

The second reason relates to the more critical issue of *quality control*. The explosive growth of computerization in general and in particular of the Internet with its inherent lack of quality control may offer possibilities to the producer and publisher but may also make retrieval of information time-consuming, tedious, unreliable and unproductive. This problem has been attacked in two different ways. The first is the use of expert or just informed *search intermediaries*. These either operate as domain authorities who sanction information sources or as peer judges who guide search through pre-selection (Negroponte 1995). The second way is through search engines that index Internet sites practically exhaustively and facilitate full-text natural language retrieval. A combination of the two is based on vocabulary control, i.e. the classification of indexing terms into thesauri (GAHIP 1990). These allow for non-redundant, coherent and consistent indexing and provide knowledgeable relevance feedback already from the stage of query formulation (Koutamanis 1995). At the same time they avoid the pitfalls of authoritative control of the actual information and its sources and focus instead on the improvement of information utility.

The main purpose of a search intermediary is to guide search. This may amount to pre-structuring of queries, as in categorized link lists, or take place as *relevance feedback*. Evaluating precision and recall in retrieval presupposes a clear understanding of the ways retrieval takes place, the structure and content of a query and, above all, of the subject matter in the databases searched. A near-perfect example of a search intermediary capable of providing relevance feedback is a knowledgeable

librarian who can parse the searcher's utterances, evaluate the searcher's actions and their results, and propose dialectically improvements in terms of search formulation and strategy. Such feedback relates not only to knowledge and understanding of the library structure but also to episodic knowledge of e.g. a book that may contain a relevant chapter. Dramatic improvements in recall and above all in precision are made possible by relevance feedback that combines these two aspects.

The role of search intermediaries is also critical with respect to information comprehensiveness and consistency. A fundamental consideration in the development of approaches to the indexing and retrieval of pictorial information is the structure of this information. "Structure" refers not only to the organization of image databases but also to the correlation of pictorial and other information towards *compound representations* that describe a wider spectrum of aspects, as well as relationships between these aspects. Compound representations go beyond multimedia and address the correlation of different information carriers and forms into unambiguous and complete descriptions. These generally emerge in a bottom-up fashion by means of dynamic links between elf-sufficient, integral partial representations, which may remain in the possession or custody of a single party. Especially in group processes such compound representations present a welcome alternative to procrustean institutional standardization, as well as to superficial document management schemes.

In this framework verbal information and verbal interfaces using natural language play a central role. This role is somehow surprising amid the multimedia possibilities of the Internet and current operating systems. The popularity of chatting on the Internet and the predominantly verbal content of e-mails attest to two main functions of natural language. The first is the ability to oscillate between the formal and the informal, to mix the specific and the abstract, to combine precision and association. The flexibility of natural language in information retrieval is of paramount importance for the definition of effective queries and for the improvement of queries following relevance feedback. The second function is the all-encompassing role of language in human culture. There are few concepts and situations that are not amenable to full verbal description. Even when an image is preferable in terms of precision and overview, as e.g. an orthographic floor plan for the description of the layout of a building, natural language is adequate for describing most aspects, such as the geometry of a space, and essential for correlating the image with external information, as in describing emotions arising while entering the space or evaluating the acoustics of the space.

It can be argued that these two functions are integral to human and machine intelligence. The similarities between the Turing test and assumed identities or role-playing in Internet chat rooms are not coincidental. The ability to create a verbal description of an image or a situation supports communication at a variety of abstraction levels, each with different relationships to background or external knowledge, in addition to its relationships with other levels. Interaction between the parties involved in communication relies heavily on these relationships with respect to aspects like efficiency and reliability. This corresponds with the view of indexing and retrieval as communication between the indexer, the searcher and all human contributions integrated

in the structure and content of the information system, especially because it stresses the significance of implicit information and external information that must be made available on an if-needed basis.

3 AZILE

The development of AZILE, a natural language interface that operates through a dialogue with the user, was motivated by the above considerations. AZILE integrates verbal terms for architectural pictorial information in a chat-like environment that refines queries and provides relevance feedback. In this environment the user expresses queries not by means of formalized combinations of terms and operands but as normal utterances such as "I am looking for floor plans with cross-shaped central spaces". Such utterances contain the same information, albeit in an informal manner that is subject to interference by syntactical variations or anomalies. At the same time, they offer significantly more flexibility and ambiguity in the description of a query than formalized structures. As a result, they are less constrained by the prejudices that may underlie classification of search terms or stored information.

AZILE presupposes an indexed collection of pictorial information. This can be a loose collection of architectural images indexed by just a couple of terms, such as the name of the architect, the name of the building, the building type and style. Equally acceptable are structured databases that cover a complete class of designs, fully described in terms of components derived from image segmentation or a generative process (Koutamanis 2000; Koutamanis 2000). It is also possible to combine different types and structures of pictorial collections. All that matters is that the images are indexed by one or more architectural terms.

The indexing terms can obviously vary, especially if the collection is mixed. Even within a single natural language like English and for a single, established area such as architecture there is tremendous scope for synonymy and variation in time, location and usage. Variation problems are resolved by referring the collected indexing terms to a controlled vocabulary that normalizes the terms by placing them in the framework of a comprehensive taxonomy. The most interesting controlled vocabularies are *thesauri* which combine polyhierarchical structure with extensive cross-referencing between terms in different branches and at different levels (Koutamanis and Mitossi 1992; Koutamanis 1995). In a thesaurus a term is normally described by:

- *The preferred term*: the term best describing the concept in the given cultural context, e.g. "reinforced concrete".
- *Scope notes*: an explanation of the breadth and depth of the preferred term, e.g. the kinds of reinforced concrete covered by the term.
- *Non-preferred terms:* obsolete or otherwise incorrect terms that may still be used under certain circumstances, such as "ferroconcrete" in descriptions of Auguste Perret's Rue Franklin apartment block.
- *The immediately broader term* in the local hierarchy (or hierarchies) of the term

- *The immediately narrower terms*
- *Otherwise related terms,* i.e. terms belonging to distant levels or distant local hierarchies

The extent and comprehensiveness of the *Art and Architecture Thesaurus* (GAHIP 1990), as well as its availability on the Internet and as an indexing structure, make it an obvious choice as a reference structure for AZILE. However, practical constraints meant that the initial controlled vocabulary for AZILE is an own adaptation of part of the *Architectural Keywords* of the British Architectural Library (BAL 1982), which has been developed in the framework of earlier related research. The choice of thesaurus is relevant to the performance of AZILE but not to its structure.

The use of an architectural thesaurus as a reference structure may classify and elucidate architectural terms employed by the user of AZILE but cannot assist in the parsing of the complete verbal input of the user. This is altogether avoided by means of a rudimentary approach that uses a minimum of linguistic knowledge. Using a relatively small knowledge base of predefined adaptable templates it is possible to select the significant components of the user's utterances and their qualifications and use them as triggers for appropriate reactions (Weizenbaum 1976). These reactions are similarly based on a small number of predefined adaptable templates that give the impression of understanding English. The reason for doing so is to help the user reformulate, refine or apply a query in a flexible, informal manner. This does not imply that this particular manner is the ultimate solution to pictorial or architectural retrieval. It merely represents an additional mode worth investigating.

The significant components in the user's utterances are matched to the terms of the thesaurus. The matching returns an initial evaluation of the query in terms of conformity with the thesaurus. The results of the evaluation are not presented to the user but nevertheless used for selecting images from the available collection. If the query terms are found, then the corresponding images are offered to the user. For example, an initial query such as "I am looking for floor plans with cross-shaped central spaces" in the context of a Palladian collection would elicit the reaction "Would you like to have a look at Palladio's Villa Foscari?" If the query terms are not present in the available collection, the matching of the query with the thesaurus provides the means for relaxation (substitution by broader terms) and fuzzification (substitution by relative terms). Relaxation and fuzzification are applied recursively until the query can select images from the database.

Selection is facilitated by a parallel operation performed in the background of the dialogue with the user. The terms in the initial query and their broader, related and non-preferred terms (i.e. terms that might be used in relaxation or fuzzification) are used to arrive at a preliminary classification of available pictorial information into fuzzy sets that can combine to produce a wide spectrum of reactions in the dialogue with the user. However, this does not mean that AZILE returns a comprehensive enumeration of retrieved item, as this would not be in character with the chosen manner of communication with the user. AZILE reactions are reasonably well-founded yet

opportunistic suggestions that prompt the dialogue and serve the following purposes:

1. *Incremental query refinement:* the thesaurus can associate a query term such as "Palladio" with "Classicism" and through this expand on the themes of central spaces and cross-shaped forms. AZILE presents the results in a gentle and indirect fashion, through leading reactions like "Are you interested in Classical buildings with cross-shaped spaces?" prior to making eventual concrete suggestions to browse available images. This allows for reflection and consideration of several alternatives.
2. *Direct retrieval of suitable information:* if the user chooses to accept an AZILE suggestion he can proceed directly to browse the selected image. From there he may go on browsing, depending on the structure of the pictorial database. If at any point he chooses to return to the dialogue with AZILE, a comparison is made between the originally selected image and the last image visited. The indexing terms of the two are matched to the thesaurus in order to ascertain their relationship (direct or indirect through a broader or related term). Acting on the assumption that the reason for returning to AZILE is dissatisfaction with the images visited, AZILE makes suggestions that preclude returning to the last image. For example, if the last image is one of Villa Stein, it is assumed that the user is not interested in Modernist architecture or in associations between Palladian villas and Corbusian villas.
3. *Support machine intelligence*: user utterances and reactions may contain query terms that have not been used to index the selected images. These terms are added to the image associations and used tentatively in subsequent queries. Consequently, every image in the collection is indexed by a number of initial terms and an increasing collection of possible terms as a result of user selections. Possible terms may obtain the same status as initial ones if extensively used by several users.

4 FURTHER DEVELOPMENT

AZILE has been developed as a self-sufficient system requiring only access to a thesaurus of architectural terms and a collection of architectural pictorial information but has been envisaged as part of a multimodal information system with several alternative interfaces and indexing / retrieval mechanisms. Integration of AZILE in this system is one of the priorities in its further development. This entails a closer correlation with the other modules of the system and in particular a comprehensive analysis of relations between verbal terms, pictorial elements and formal representations. The results of this analysis should form the basis for retrieval on the basis of image segments, i.e. with reference to specific building or spatial elements in the building representation. Also migration towards other the *Art and Architecture Thesaurus* seems inevitable.

Beyond the confines of this information system, AZILE can be coupled to an

Internet search engine, preferably one with content-based retrieval capabilities. This could provide useful insights into the effectiveness of efficiency of the approach without the benefit of a known pictorial collection that can be pre-processed in the background. Similarly useful are insights into the relationship between user queries, the thesaurus and the possibilities of existing content-based techniques.

The main internal improvement concerns machine intelligence. Negative user reactions to AZILE suggestions should modify the parameters of the predefined reaction templates so as to accommodate a wider spectrum of possibilities. They could even cause the mutation of a template so as to cover a new category or property. Such developments are currently stimulated by the increasing interest in verbal interfaces in various kinds of bots on the Internet.

5 REFERENCES

Baird, H. S., H. Bunke, et al., Eds. (1992). *Structured document image analysis.* Berlin, Springer-Verlag.

BAL, B. A. L. (1982). *Architectural keywords.* London, RIBA Publications.

GAHIP, G. A. H. I. P. (1990). *Art and architecture thesaurus.* New York, Oxford University Press.

Gong, Y. (1998). *Intelligent image databases.* Boston, Kluwer.

Gross, M. D., E. Y.-L. Do, et al. (2001). The design amanuensis. *Computer aided architectural design futures 2001*. B. de Vries, J. van Leeuwen and H. Achten. Dordrecht, Kluwer.

Koutamanis, A. (1995). Recognition and retrieval in visual architectural databases. *Visual databases in architecture. Recent advances in design and decision making.* A. Koutamanis, H. Timmermans and I. Vermeulen. Aldershot, Avebury.

Koutamanis, A. (1997). Multilevel representation of architectural designs. *Design and the net*. R. Coyne, M. Ramscar, J. Lee and K. Zreik. Paris, Europia Productions.

Koutamanis, A. (1998). Information systems and the Internet: towards a news counter-revolution? *4th Design and Decision Support Systems in Architecture and Urban Planning Conference.* Eindhoven.

Koutamanis, A. (2000). Recognition of spatial grouping in rectangular arrangements. *Design and decision support systems in architecture. Proceedings of the 5th International Conference.* Eindhoven, Eindhoven University of Technology.

Koutamanis, A. (2000). Representations from generative systems. *Artificial Intelligence in Design '00.* J. S. Gero. Dordrecht, Kluwer.

Koutamanis, A. and V. Mitossi (1992). "Design information retrieval." *Delft Progress Report* **15**(2): 73–86.

Lopes, D. (1996). *Understanding pictures.* Oxford, Clarendon.

McFadzean, J. (1999). Computational Sketch Analyser. *Architectural computing: from Turing to 2000.* A. Brown, M. Knight and P. Berridge. Liverpool, University of Liverpool.

Negroponte, N. (1995). *Being digital.* London, Hodder & Stoughton.
Weizenbaum, J. (1976). Computer Power and Human Reason: From Judgement to Calculation. San Francisco, Freeman.