Towards a comprehensive survey of the semantic gap in visual image retrieval

Peter Enser and **Christine Sandom**

School of Computing, Mathematical and Information Sciences, University of Brighton, U.K.
pgbe@bton.ac.uk, c.sandom@bton.ac.uk

Abstract. This paper adopts the premise that the 'semantic gap' is an incompletely surveyed feature in the landscape of visual image retrieval, and proposes a framework within which this deficiency might be made good. Simple classifications of types of image and of types of user are proposed. Consideration is then given in outline to how semantic content is realised by each class of user within each class of image. The argument is advanced that this realisation finds expression in perceptual, generic interpretive and specific interpretive content. This analytic framework provides the basis for the specification of a broadly encompassing evaluation study, which will employ the image/user type classification and the expert domain knowledge of selected user groups in the construction of segmented test collections of real queries, images and relevance judgements. From this study should come a better-informed view on the nature of semantic information need, and on the representation and recovery of semantic content across a broad spectrum of image retrieval activity.

1 Introduction

Within the last ten years content-based image retrieval (CBIR) has risen to prominence in the research agenda for computer science. The digitised image offers some enticing processing opportunities derived from quantifiable attributes of colour, texture and the spatial (or, in the case of moving images, spatio-temporal) distribution of shapes. With developing maturity has come a realisation of the limitations of CBIR processes in practice, however. These limitations reflect the fact that the retrieval utility of visual images is generally realised in terms of their inferred semantic content. The context for this inferential reasoning process is to be found in the distinction drawn in semiotics between the denotation, or presented form, of the image and the connotation(s) to which it gives rise [1]. It is clear that personal knowledge and experience, cultural conditioning and collective memory – the shared knowledge of a society – contribute towards that reasoning process. The CBIR community

has attached the label ‘semantic image retrieval’ to the formulation and resolution of information needs which engage that intellectual process. The sharply drawn distinction between that process and the automatic extraction of low level features from denotative pixel structures is characterised as the ‘semantic gap’[2].

Given the vast and constantly expanding scale upon which visual resources are made available via the World Wide Web, users, providers and the CBIR research community collectively have much to gain from a narrowing of that semantic gap. This paper adopts the premise that a complete survey of this feature in the landscape of visual image retrieval has yet to be undertaken, and proposes a framework within which this might be undertaken.

The approach taken in this paper is, first, to propose a classification of types of image, followed by an even simpler classification of types of user. Consideration is then given in outline to how semantic content is *realised* by each class of user within each class of image. This framework provides the basis for a proposal to set up a formal study which is intended to generate a more comprehensive evaluation of visual image retrieval paradigms than any reported thus far.

2 A Categorisation of Image Type

2.1 Documentary – general purpose

Documentary images are interpreted here as faithful representations of reality. They may be captured as a result of a photographic process, or created by some form of human or human-initiated endeavour. In their captured form they represent a momentary entrapment of reality, typical examples being photographs taken above, on or under the land and sea. In their created form they are artworks using any of a variety of materials.

2.2 Documentary – special purpose

Images in this category are a faithful representation of a specific part of some larger reality, that part being the subject of specialised analysis and not necessarily visible without special equipment. As in the general purpose case, such images may be captured or created. Typical examples of the former are medical X-rays, ultrasound scans and microscopy images, whilst fingerprints are one example of the created form of such images.

2.3 Creative

Images in this category may be placed on a spectrum of reality representation, from documentary images which have been subject to some degree of amendment or manipulation, through to completely abstract artworks.

2.4 Models

In this category may be found 2- and 3-dimensional images which model aspects of reality such as processes and geographical phenomena. Typical examples are maps, diagrams, charts, plans, architectural and engineering drawings.

2.5 Moving images

The classification shown in Figure 1 is couched in terms of still images, but may be applied to moving images as well. The animation of photographically or digitally captured sequences of any of the still image types is theoretically possible, although the footage resulting from some of the stills might have little practical significance. For example, TV news footage belongs to the class of general-purpose documentary images, and digital video microscopy is a moving image variant of the special-purpose documentary image. The feature film is the dominant, animated version of the creative image, and animated weather maps which feature in TV broadcasts are one example of the moving image version of the 2-dimensional model.

3. A Categorisation of User Type

For each of the four classes of image a variety of *specialist* users can be identified. For example, among the class of general purpose documentary images, archive photographs of urban scenes may be sought by transport historians and historical geographers. Surgeons, conservators and art historians are among the diverse users of X-ray photographs, an example of special purpose documentary images. Among the various types of artwork which contribute to the class of creative images, games designers will make use of computer-generated images; and astronomers will use (celestial) charts and maps within the class of 2-dimensional images. Other examples of specialist users of particular types of image are shown in Figure 1.

User type Image type	Generalist	Specialist
Documentary - general purpose		
archive photo – urban scene	general public	transport historian; historical geographer
aerial photo	general public	town planner; military intelligence agent
contemporary photo - society	general public	journalist; fashion designer
Documentary – special purpose		
X-ray photo	n/a	Surgeon; conservator, art historian
microscopy image	n/a	microbiologist
ultrasound scan	n/a	obstetrician
fingerprint	n/a	police officer
facial identification photo	[general public]	immigration officer; police officer
Creative		
painting	general public	art historian; curator
trademark; logo	general public	trademark lawyer
tapestry	general public	textile designer
computer-generated graphic.,	general public	graphic designer; cartoonist; games software programmer
Model		
map	general public	geographer; civil engineer; archeologist
chart	general public	astronomer; navigator; genealogist
technical drawing	general public	architect; engineer
diagram	general public	engineer; graphic designer; technical author

Fig. 1 A classification of image types and user types, with example members

Generalist users, characterised in Figure 1 as the general public, may seek general purpose documentary images, 2-dimensional models and creative images, expressing visual information needs which we may expect to be different from those of the specialist users. Generalist users would not normally engage with special-purpose documentary images, although passport photographs – a variety of facial recognition image – and foetal scans are exceptions to the general rule.

4. Realisation of the semantic content of an image

The two-way classification of image and user types shown in Figure 1 provides a framework within which the semantic gap in visual image retrieval may be analysed. For each class of image we consider what is meant by semantic content, and further consider how both types of user perceive that semantic content.

4.1 General purpose documentary images

The semantic content of general purpose documentary images is multi-layered. This multi-layering has been described in different ways. The art historian Panofsky, working with creative images, identified 'pre-iconographic', 'iconographic' and 'iconologic' levels of expression [3], which Shatford's generalisation in terms of generic, specific and abstract content, respectively, made amenable to general purpose documentary images [4]. Shatford is more particularly associated with the *of-ness* and *about-ness* of image content, the former derived from Panofsky's *factual* pre-iconography and iconography levels, and the latter from his *expressional* pre-iconography and iconography levels of analysis [4]. The of-ness and about-ness content finds alternative expression in *hard* and *soft* indexing [5], an approach which resonates with the *perceptual* and *interpretive* layers of meaning postulated by Jörgensen [6]. In her study of human pictorial image perception, participants – asked to identify the attributes present in an image – also recognised *reactive* content. The latter was defined in terms of subjective response to the image (such as uncertainty and pleasure) and is not further considered here. Perceptual attributes were those named as interpretation-free responses to a visual stimulus, and correspond broadly with the generic/pre-iconographic category of Shatford/Panofsky.

Interpretive attributes in the Jörgensen study "are those which require both interpretation of perceptual cues and application of a general level of knowledge or inference from that knowledge to name the attribute" [6]. In the case of people, interpretive qualities include the nature of the relationship among depicted persons, their mental or emotional state, or their occupation, for example. Such attributes do not conform with Panofsky's analysis since, although they are generic features, they are not interpretation-free. The specific name by which the person was known would also be an interpretive attribute, whereas Shatford/Panofsky distinguish such specific/iconographic content from the generic subject matter featured in an image. Jörgensen's interpretive attributes also encompass the abstract/iconological category of pictorial content specified by Shatford/Panofsky.

A simple characterisation of the semantics of an image in terms of perceptual and interpretive content is attractive. On the other hand, studies of user need for general purpose documentary images, both still and moving, have revealed a high incidence of requests for specific, named features such as objects, places, events and people [7-12]. There would appear to be some value, therefore, in adapting the Jörgensen model for our present purposes,

by recognising semantic content in terms of perceptual, *generic* interpretive and *specific* interpretive attributes.

How a user recognises perceptual and generic interpretive matter in an image is a cognitive phenomenon which is, as yet, incompletely understood [13,14]. Greisdorf & O'Connor, reporting that "the research consensus points to an integrated cortical process involving, at least, perception and cognition", suggest that the visual impression engendered by the sensory stimuli is first cognitively matched to some form of syntactic equivalence. First time viewers, moreover, "appear to determine initially what the image represents to them before making any evaluations of its topicality, meaning and utility in regard to an information need" [13]. It might seem reasonable to suggest that the user recognises perceptual and generic interpretive content by low-level features within the image; shape may be especially significant, complemented by colour and texture, bringing to bear a previously learned linguistic identifier to generate meaning. For example, a paddle steamer may be detected within an image by matching an outline (despite the complexities of occlusion, orientation, perspective, etc.), with memorised profiles associated with the verbal tag 'paddle steamer'.

One might expect the same to be true for specific interpretive features, which, it is postulated, the user also recognises by feature matching underpinned by a defining linguistic identifier. Thus, we recognise President Bush when we see his image. Furthermore, we continue to recognise him in creative images, even when his features are subject to some degree of denotational degradation under the influence of the cartoonist's pen – an aspect of recognition by components which has been investigated in the psychology literature [15]. Whatever the perceptual processes involved it would seem to be the case that *identification* is dependent upon the prior existence – and knowledge by the user – of a defining linguistic tag. However, we heed Eakins' warning that too many investigators have made unwarranted assumptions about the nature of perceptual similarity [16], and lend support to the call for further research into the psychophysical aspects of human perception [13,14].

As an added complication, the process of identification may involve *context*, recognition of which would seem to invoke high-level cognitive analysis supported by domain and tacit knowledge. Contextual anchorage is an important role played by text annotation within the image metadata [1,5].

When the focus of interest lies with the abstract content of the image – the client wanting images of suffering or happiness, for example - shape may be of limited use, unless we use

recognition of features within an image to recall scenes within our memory which invoke the appropriate cognitive response . Colour might be significant, since it can be an effective communicator of mood. We are likely to be dependent, however, on the presence within metadata of an appropriate textual cue which conditions our *interpretation* of the semantic content of an image.

Both generalist and specialist users will realise semantic content of general-purpose documentary images in terms of perceptual and interpretive attributes. Such attributes in the case of the specialist user may involve linguistic identifiers of high specificity drawn from the specialised vocabulary of the subject domain in question.

4.2 Special purpose documentary images

The image might take a variety of forms, but realisation of semantic content is likely, again, to reveal a heavy dependency on an intial detection of primitive features. In some applications – those which employ images captured by scanning technology or microscopy, for example - the spatial distribution of shapes, colours and textures may be particularly significant. For the specialist user, the existence and spatial distribution of such features within the image may lead to inferential reasoning about some external condition for which the image is a passive signal; examples would be the presence of a tumour within an organ of the body, or a structural mass beneath the surface of the land or water.

The user's inferential reasoning may thus establish *significance* and *identification* - semantic properties which, it would seem clear, depend on linguistic identifiers for their realisation. Again, such identifiers may be highly specific and drawn from the specialised vocabulary of a particular subject domain.

4.3 Creative images

Creative images assume a greater variety of physical forms than is the case for general-purpose documentary images; some indication of this variety may be found in [17]. In many cases the realisation of semantic content does not depart in any significant way from the analysis presented in 4.1 above. However, as the accuracy with which reality is represented in this class of images is progressively relaxed towards totally abstract form, the more significant become the perceptual and generic interpretive attributes. Correspondingly, the

more potentially effective becomes the application of CBIR techniques to this class of image, for both generalist and specialist users.

There are specialised types of creative image, the perceptual attributes of which predominate over the interpretive, or in which interpretive attributes are missing. These are context-free images which do not have the foreground/background disambiguation problem which, from the CBIR perspective, bedevils 'real scene' images. In such cases CBIR can be a powerful tool for retrieving images which are similar, on the basis of some chosen metric, to a target image. For the specialist user in particular, significant applications in trademark matching [2,18,19], and experimental work in fabric design pattern matching [2] have been reported.

4.4 Models

A particular characteristic of this class of image is the presence of symbols with domain-specific meanings, shapes and delineated regions. Typically, the spatial distribution of, and relationships between, the components are highly significant. Texture and, especially, colour may also be significant, and embedded text is a frequently encountered additional feature.

In the context of a particular image, these components combine to act as a surrogate of its semantic content. For example, a diagram of a city subway system uses a range of symbols and lines to represent the reality of a particular public transport infrastructure; the diagram 'means' the transport system. The image lends itself to a single, correct - one might say objective – interpretation. For this class of images, as with context-free creative images, it is the perceptual, or denotational content which is significant.

In many cases, retrieval will be a matter of recovering a specific artefact on the basis of its title or other unique identifier. However, for the specialist user in particular, CBIR offers a promising approach to searching digital archives for similar versions of some target image, on the basis of a primitive attribute, most obviously shape [2]. A specific application in engineering drawings is one such example [20].

5. Retrieval evaluation study proposal

In the preceding sections a framework has been outlined within which a comprehensive survey of the 'semantic gap' might be undertaken. Such a comprehensive survey would seem

overdue: the scale on which visual image retrieval activity is conducted is now very significant indeed, and must continue to grow as access to ever larger quantities of image material is liberated.

As yet there has been no study which acknowledges the full plurality of image and user types, as represented in the framework shown in this paper. There have been a limited number of end-user needs studies which have taken a partial view, including the use of general-purpose documentary images by generalist and specialist users [7-9,12,21-24]; also the use of creative images by specialist users [25,26]. Subject requests recorded by generalist and specialist users of general-purpose documentary film and video has also been reported [10,11]. But user needs analysis of special-purpose documentary images and of models awaits even this level of attention.

In order to arrive at a better informed view of the semantic gap and the possibilities for bridging it there would seem to be merit in the specification of a broadly encompassing evaluation study. Central to this study is the construction of segmented test collections of real queries and images, the segments structured in accordance with the image/user type classification outlined above. The assembly of the test collections, together with the accompanying relevance/pertinence judgements, will employ the expert domain knowledge of selected user groups. From this endeavour should come a better appreciation of the incidence of perceptual, generic interpretive and specific interpretive content of image requests across a broad spectrum of image use.

From the same evaluative platform will come an informed view on the representation and recovery of perceptual, generic interpretive and specific interpretive content by means of a consistently-applied indexing strategy, including both textual and non-textual metadata. From such an informed view will emerge, hopefully, any prospect for narrowing the 'semantic gap' in visual image retrieval.

References

1. Barthes, R.: The Elements of Semiology. Cape, London, (1967)
2. Eakins, John P.; Graham, Margaret E.: Content-based Image Retrieval. A report to the JISC Technology Applications Programme. Institute for Image Data Research, University of Northumbria at Newcastle, Newcastle upon Tyne,, (1999). http://www.unn.ac.uk/iidr/. Accessed February 2003.
3. Panofsky, E.: Studies in Iconology: Humanistic Themes in the Art of the Renaissance. Harper & Rowe, New York, (1962)
4. Shatford, S.: Analyzing the Subject of a Picture: A Theoretical Approach. Cataloguing & Classification Quarterly, 5(3) (1986), 39-61
5. Krause, M. G.: Intellectual Problems of Indexing Picture Collections. Audiovisual Librarian, 14(2), (1988) 73-81

6. Jörgensen, Corinne: Indexing Images: Testing an Image Description Template. Paper given at the ASIS 1996 Annual Conference, October 19-24, 1996. http://www.asis.org/annual-96/ElectronicProceedings/jorgensen.html. Accessed February 2003.
7. Enser, P.G.B.: Query Analysis in a Visual Information Retrieval Context. Journal of Document and Text Management, 1(1), (1993) 25-52
8. Enser, P.G.B.: Pictorial Information Retrieval. (Progress in Documentation). Journal of Documentation, 51(2) (1995) 126-170
9. Armitage, L.H.; Enser, P.G.B.: Analysis of User Need in Image Archives. Journal of Information Science, 23(4) (1997), 287-299
10. Sandom, C.; Enser, P.: VIRAMI – Visual Information Retrieval for Archival Moving Imagery. In: Bearman, D.; Garzotto, F. (eds.): ichim01 International Cultural Heritage Informatics Meeting: Cultural Heritage and Technologies in the Third Millennium, Politecnico di Milano, Italy, September 3-7 2001, 141-152
11. Enser, P. and Sandom, C.: Retrieval of Archival Moving Imagery - CBIR Outside the Frame? In: Lew, Michael S.; Sebe, Nico,; Eakins, John P. (eds.): Image And Video Retrieval. International Conference, CIVR 2002, London, UK, July 18-19, 2002 Proceedings. Springer, Berlin, (2002) 202-214
12. Markkula, M.; Sormunen, E.: End-user Searching Challenges Indexing Practices in the Digital Newspaper Photo Archive. Information Retrieval 1(4), (2000) 259-285
13. Greisdorf, Howard; O'Connor, Brian: Modelling What Users See When They Look at Images: A Cognitive Viewpoint. In: Journal of Documentation. Vol 58. No. 1, (2002) 6-29.
14. Rui, Y.; Huang, T.S.; Chang, S.-F.: Image Retrieval: Current Techniques, Promising Directions, and Open Issues. Journal of Visual Communication and Image Representation 10, (1999) 39-62.
15. Biederman, I.: Recognition-by-components: A Theory of Human Image Understanding. Psychological Review 94(2), (1987) 115-147.
16. Eakins, J.P.: Content-Based Image Retrieval – What's Holding It Back? Paper given at the ASCI [Advanced School for Computing and Imaging] 2002 Conference, Lochem, the Netherlands, June 19-21 2002.
17. Graham, Margaret E.: The Description and Indexing of Images. Report of a survey of ARLIS members, 1998/1999. (1999). http://www.unn.ac.uk/iidr/ARLIS/arlisrep.htm. Accessed February 2003.
18. Eakins,, J.P.; Graham, M.E.; Boardman, J.M.: Trademark Image Retrieval by Shape Similarity. IEEE Multimedia, 5(2), April-June 1998, 53-63
19. Wu, J.K. et al.: Content-Based Retrieval for Trademark Registration. Multimedia Tools and Applications 3, (1996) 245-267.
20. Eakins, J. P. Design Criteria for a Shape Retrieval System. Computers in Industry, 21 (1993), 167-184
21. Conniss, L.R; Ashford, L.R; Graham, M.E.: Information Seeking Behaviour in Image Retrieval. VISOR 1 Final Report. Library and Information Commission Research Report 95. Institute for Image Data Research, University of Northumbria at Newcastle' Newcastle upon Tyne, (2000)
22. Fidel, Raya: The Image Retrieval Task: Implications for the Design and Evaluation of Image Databases. New Review of Hypermedia and Multimedia. 1997 (3) 181-199
23. Ornager, Susanne: The Newspaper Image Database: Empirical Supported Analysis of Users' Typology and Word Association Clusters. In Fox, E.A.; Ingwersen, P.; Fidel R.; (eds): SIGIR 95, Proceedings of the 18th International AGM SIGIR Conference on Research and Development in Information Retrieval. Association for Computing Machinery, New York, (1995). 212-218
24. Markkula, M.; Sormunen, E.: Searching for Photos - Journalists' Practices in Pictorial IR. In: Eakins, J.P.; Harper, D.J.; Jose, J. (eds) The Challenge of Image Retrieval: papers presented at a Workshop on Image Retrieval, 5 February 1998, University of Northumbria at Newcastle, UK. University of Northumbria at Newcastle, Newcastle upon Tyne,, (1998)

25.Hastings, S.K.: Query Categories in a Study of Intellectual Access to Digitized Art Images. ASIS '95: Proceedings of the 58th ASIS Annual Meeting, Vol.32, (1995), 3-8

26. Hastings, S.K. Evaluation of Image Retrieval Systems: Role of User Feedback. In: Sandore, B. (ed.) Progress in Visual Information Access and Retrieval. Library Trends 48(2), (1999) 438-452.

This paper was originally published in the proceedings of the Second International Conference on Image and Video Retrieval; CIVR 2003, held in Urbana-Champaign, IL, USA, July 24-25, 2003. Published: Bakker, E.M. *et al.* (eds.) Lecture Notes in Computer Science, Vol. 2728. Berlin: Springer-Verlag, 2003, 291-299.