The Meaning of an Image in Content-Based Image Retrieval

Walter ten Brinke[1], David McG. Squire[2], and John Bigelow[3]

[1] Clayton School of Information Technology
[2] Caulfield School of Information Technology
[3] School of Philosophy and Bioethics
Monash University, Victoria, Australia
Walter.tenBrinke@csse.monash.edu.au

Abstract. One of the major problems in CBIR is the so-called 'semantic gap': the difference between low-level features, extracted from images, and the high-level 'information need' of the user. The goal of diminishing the semantic gap can be regarded as a quest for similar 'concepts' rather than similar features, where a concept is loosely defined as "what words (or images) stand for, signify, or mean" [1]. We first seek to establish a metaphysical basis for CBIR. We look at ontological questions, such as 'what is similarity?' and 'what is an image?' in the context of CBIR. We will investigate these questions via thought experiments. We will argue that the meaning of an image—the concept it stands for—rests on at least three pillars: what actually can be seen on an image (its ontology), convention and imagination.

1 Introduction

Before the advent of Content-Based Image Retrieval (CBIR) as a means to automatically index and query image databases, access to such collections was based on textual annotation. Annotation, however, is a lengthy, costly and cumbersome process that has ambiguous results. It depends not only on the state of mind of the annotator, but also on the situation the annotator is in. In recent years, so-called 'semantic' approaches have come to the fore. Often these approaches are based on predefined categories, for example, in an 'ontology'. The term 'ontology' has been appropriated by Information Science from Philosophy and has been given a different sense. For instance, Gruber's short answer to the question 'what is an ontology?' is that an ontology is a "specification of a conceptualization" [2], quite different from "a general theory of what there is" [3, p. 400], as twentieth century analytical philosophers use the term. Zùñiga [4] describes this appropriation and attempts to reconcile the different viewpoints. In this paper we will use the term ontology in the latter sense of 'what there is'. Perhaps, a 'specification of a conceptualization' can be be built on top of our ontology of an image.

In the early days of CBIR it became apparent that the content-based approach to image retrieval exhibited the same sorts of problems as did text-based

Information Retrieval (IR), linked to the difficulty for a user to express his/her actual information need in terms of the document representation mechanism (often hidden) used by the system. One of the major problems came to be known as the 'semantic gap', i.e. the difference between (the similarity of) low-level extracted features, such as measures of colour and texture, and (the similarity of) high-level concepts [5]. This is even more accentuated in CBIR than in IR, since the "distance" between the features (e.g. colour and texture descriptors vs. individual words) and the meaning of the document is apparently greater. Some in the CBIR community looked at IR and introduced Relevance Feedback (RF) [6] into the retrieval process [7].

The rationale of Relevance Feedback is that by incorporating a user's judgments of the retrieval results into subsequent queries, the user would be able to further refine the representation of his/her information need, and thus enable the system to narrow the semantic gap. However, little has been said about the nature of the relation between the low-level and high-level features. We have investigated that relation and believe that it is one of *supervenience* [8].

In this paper we discuss the process by which an image acquires meaning. We will argue that the meaning of an image is indeed in the mind of the beholder, but that his/her state of mind is affected by (1) the physical appearance of the image on a computer screen, (2) conventions and (3) individual capabilities, in particular imagination.

The remainder of this paper is organized as follows. We give a brief overview of the CBIR process, highlighting those aspects that have a bearing on this paper's argument. This is followed by a brief discussion of our ontology of an image, with supervenience relations between pixels, visual features, and the 'whole of the image'. We then discuss the role of convention and imagination in conceptualization. We conclude with some remarks concerning our methodology.

2 A Overview of the CBIR Retrieval Process

Without collections of digital or digitized images, CBIR would not be possible. The origins of the images and the contexts in which they were taken are largely immaterial to our argument, except that the collection is not from highly specific domains (e.g. medical imagery). That is not to say that what follows is not applicable to specific domains, but rather that what is in mind is a heterogeneous collection. It might include medical images, but also images of paintings, sculptures, landscapes, people,... whatever.

In the generic CBIR process, the first step is to extract *features* from the stored images. The most common extracted features are descriptors of colour (e.g. colour histograms [9]), texture (e.g. responses to a bank of Gabor filters [7]), and shape [5]. The main development of these features took place in the field of Computer Vision. In some cases an effort is made to choose features that coincide with theories of how the human visual system works, such as the use of Gabor filter responses [10], but in others it seems clear that there is no such correspondence (e.g. the use of one-dimensional Fourier descriptors to describe shapes [11]). Despite this, the extracted descriptors are commonly called "visual features".

The result of feature extraction is a set of numerical features. An image is represented by these features, and these are used to index the entire collection. Although the data structures and algorithms used for indexing are at the heart of any proper CBIR system, we will not discuss them further. We simply point out the difficulty of comparing disparate and perhaps incommensurate features for indexing purposes.

Another crucial part in any CBIR system is its user interface, which enables a user to compose and submit queries. The query mechanism is typically based on the Query-By-Example (QBE) paradigm, first developed for a more intuitive access to relational databases [12]. The user is asked to define his/her 'information need' by providing an example image. Refinements include selecting an image region by mouse-click [13] or drawing a sketch [14].

After submission of the query, the CBIR system compares the features of the query image with those of others in the collection. Based on some similarity measure, often the Euclidean distance, the system ranks the images, and presents the top-n to the user, usually as a linear list of images. The retrieved images are displayed in this list in decreasing order of similarity according to the measure.

At this stage the user may give relevance feedback (RF). RF is a valuable tool in CBIR, for it gives the user the opportunity to further refine his/her 'information need'. For every image in the retrieved list, the user is enabled to indicate whether a certain image is relevant to his/her query (positive feedback), or not (negative feedback). The two early main strategies for processing user's feedback in CBIR were (1) to issue a separate query for each image with feedback and merge the results, and (2) create a new, composite query based on the user's feedback by reweighing the features in the query [7]. Note that none of the strategies attempt to overcome the 'semantic gap' by some adjustment at the high level of the 'information need'. Soon it was proved that the incorporation of RF, especially the feature reweighing strategy, enhanced the retrieval effectiveness in terms of precision and recall [15, 16].

However, the semantic gap still persists. This persistence is often blamed on human perception, as humans are said to interpret the same visual content differently at different times. Moreover, the same visual content is interpreted differently by different users [17, 15]. Be this as it may, there has been little apparent attempt to comprehensively model the visual content of an image and the workings of the human brain in perceiving and interpreting this visual content in the context of CBIR. In this paper we attempt precisely that.

In summary, we distinguish three phases in the CBIR process in which the capabilities of the human brain and mind—vision, perception, cognition, memory—play, either implicitly or explicitly, a role. In the feature extraction phase, images are decomposed into visual features. During query processing the user is initially required to express his/her information need via some example image. The processing entails comparison of the extracted features of the query image(s) with extracted features of all of the images in the database. In the comparison some similarity measurement is used. The RF part makes use of the user's judgment of the relevance of the images in the retrieved set.

In this paper we concern ourselves first with the connection between pixels, visual features and the whole of the image, secondly, with the influence of convention on our concepts, and, thirdly, with the role of imagination in our conceptualization.

3 The Ontology of an Image

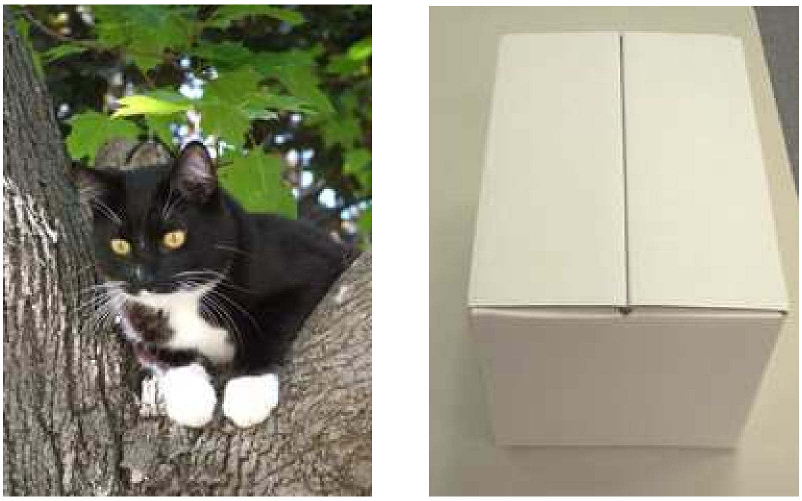
We are concerned with the image when it is displayed on an ordinary computer screen: when it offers its content to the 'beholder to be', as it were. The most common digital representation of an image is in pixels (picture elements), where a pixel consist of a value for the each of the colour components red (R), green (G) and blue (B), and two-dimensional—a computer screen is flat—spatial coordinates that define its position within the *whole* of the image. In short, a pixel can be represented with a five dimensional vector: (R, G, B, x, y). Pixels are the basic entity in our ontology of an image.

As the purpose of a CBIR-system is to facilitate retrieval of images similar to some user's information need, based on his/her visual experience of the image, comparing images solely based on their pixels will not do, for visual experience is robust to many small changes in the values and positions of pixels. Between the whole of the image and the pixels is a mediating level: the level of visual features. The third entity is seemingly the most obvious: the 'whole of the image'. Note that by the 'whole of the image', we do *not* mean simply a collection of pixels, rather we refer to the entity capable of being perceived by a human observer.

The relation between the pixels, the visual features and the 'whole of the image' is one of supervenience. We understand the relation as succinctly put by the philosopher Lewis: "supervenience means that there *could* be no difference of the one sort without a difference of the other sort" [18, p. 15]. In our case the 'one sort' are the visual features and the 'other sort' the properties of the pixels. Note that 'difference' extends to both the colour of the pixels and their spatial coordinates. In short, the visual features supervene on the pixels. Since supervenience is transitive, the 'whole of the image' also supervenes on the pixels [8, 19].

Many of the features used in CBIR systems do not have the property that the 'whole of the image' supervenes on them: it is possible to have a change in the 'whole of the image' *without* a change in the feature . We would thus choose not to call them 'visual features' in the sense above. Many features used in CBIR have the property that radically different images (in terms of both the pixels and their perception) lead to identical features. A classic example is the colour histogram: any permutation of the pixels of an image leads to another image with the same colour histogram. The vast majority of these will look like noise to a human observer. We contend, therefore, that designers of CBIR systems should strive to employ features that are visual features, in the sense that the 'whole of the image' supervenes on them.[1]

4 Convention



(a)

(b)

Fig. 1. Two images

Figure 1 presents two images. The image on the left is of a cat. We can be so specific because in our lifetimes we have learned that the word 'cat' refers to an animal which looks a lot like the one in the image. We can see the animal in the image, because the dominant visual features have the appropriate properties, in turn because the pixels have certain colours at suitable locations. The representation of the cat is central in the image and grabs our attention. There are further visual features, such as the texture we associate with the bark of a tree, and in the background we see the texture of leaves, but we do not regard these features as critical in determining what this image is about.

If, however, the pixels had had different relational properties, then it might have been a different animal altogether. For example, Figure 2 is not that of a cat, although our visual experience is dominated by black and white pixels as in Figure 1.

The right-hand image in Figure 1 is that of a box. We know, because the visual features (edges, angles between edges, shading) makes the shape visible

[1] There is perhaps a caveat here. It is sometimes desirable that features be invariant under certain transformations of the pixels (e.g. rotation, scaling. etc.) that would be perceived as a change in the 'whole of the image'.

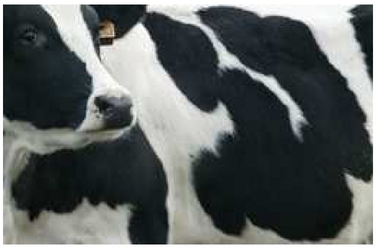


Fig. 2. A cow

that we call ‘box’. Hardly anybody will be able to *not* make a distinction between the images of Figures 1 and 2. Had our information need been, ‘cat’, ‘box’ and ‘cow’, we would have found images for them. We recognize the images, because we have learned that the images represent what we denote by the proper nouns. Here some sort convention is at work.[2]

Whatever the ‘convention’ is exactly, it is clear that it is some sort of agreement that influences the behaviour of people. Amongst the strongest conventions are the linguistic conventions, for example, that the word ‘cow’ in the English language refers to something that very much looks like the animal depicted in Figure 2. It is almost impossible to go against this convention when using—speaking, writing—English.

Other examples of conventions, include driving in a car on the left-hand side of the road in Australia and the kinds of clothes to wear for particular occasions. What all conventions have in common is that when somebody does not abide by them, there is a penalty. Driving on the right side in Australia will cause havoc for the driver and other traffic users, and wearing the wrong clothes to a function may cause exclusion in some social circles. Conventions regulate human behaviour without direct law and law enforcement [20].

5 Imagination

Consider the images in Figure 1 again. What concept do these images represent? Obviously, in isolation the left-hand image could represent the concept ‘cat in

[2] However, conventions can only work on things that are in existence, or at least that are *imagined* to be in existence. Also at work are non-conventional features of the environment. The fact that the world we live in contains cats and cows is is extrinsic to the pixels, but it feeds into the determination of whether the pixels constitute an image of a cat or a cow. In a thought experiment one might imagine a world in which there never were and never will be any such things as cats and cows. We could imagine the arrays of pixels in Figures 1 (a) and 2 arising in that world. Then the images would have all the same visual features, but they would not be images of a cat or a cow. The meaning of the image also depends on what exists in the world outside, not just the pixels on the screen.

a tree' and the right-hand image the concept 'closed cardboard box on a flat surface'. However, we deliberately put the two images together to express one concept. What that concept is depends on your frame of reference at the time you see the images. That frame of reference includes your knowledge, recollection of past experiences, your state of mind, and your mind's ability for imagination.

We offer two alternative concepts that can be derived from the two images. Both concepts are affected by what the images actually represent, explained by the supervenience relation, by the beholder's imagination, and knowledge and past experience. The first concept is that of 'a cat that is ill and has to go to the veterinary surgeon'. The beholder had a cat that very much looked like the one on the image. In its last days the cat succumbed to an illness that was treated by a veterinary surgeon. The cat was transported in a white cardboard box, such as the one on the image.

The second concept is that of the famous 'Schrödinger's Cat' thought experiment that Schrödinger conceived to attack the apparent absurdity of the notion of superposition in quantum theory being extended to macroscopic objects [21, p. 328]. The juxtaposition of the images of a cat and a box in Figure 1 might bring to mind Schrödinger's Cat for an observer with appropriate knowledge and imagination.

As a further illustration, suppose that the left-hand image of Figure 1 is the image of the cow (Figure 2). It is a big ask of the beholder's imagination, but, again, not impossible, to think of Schrödingers cat.

From these examples, we conclude that not 'anything goes' in imagination with respect to interpretation of the images in CBIR. Imagination is at least bounded by what is actually 'in' the images, what is there to be seen by the beholder.

A possible lesson for CBIR is that, when presenting images side by side to a user, the juxtaposition of certain images may well affect the user's interpretation of the meaning of the images, and, thus the user's judgment of the relevance of the images with respect to his/her 'information need'.

6 Future work

In future work we will investigate the ideas developed in this paper in implementing a CBIR system based on Formal Concept Analysis (FCA) [22]. We hypothesise that a CBIR system based on some form of concept analysis would have the potential to improve retrieval results. This expectation is based on a metaphysical viewpoint in which the locus of concepts is in the mind of the beholder. This viewpoint goes back to Aristotle (384–322 BC) for whom real properties and relations, 'universals', are in the things themselves and only the mind's faculty of abstraction gives access to these universals.

In our ontology an image is represented by features that supervene on the pixels. In FCA, a formal context is defined as a set of objects, a set of attributes, and a relation between them. A 'formal concept' consists of all the objects that have certain attributes in common (via the relation), and vice versa. The idea is to create a formal context in which the objects are images and the attributes are binary predicates that indicate whether an image has or does not have a certain feature.

The result of FCA is a lattice, the vertices of which are concepts. Edges connect concepts, leading to more general concepts in one direction (up), and more specific in the other (down). We hope that using this lattice in the user interface of a CBIR system will help to reduce the semantic gap, because the user will be able to compare his/her high-level concepts with the formal concepts. Even though the latter may not coincide with the former, we do not have to transform high-level concepts to the low-level features. Furthermore, by presenting images as part of a concept rather than in a linear list, we also hope to circumvent the 'juxtaposition' effects of a ranked list.

7 Conclusion

We have described the main phases of the CBIR retrieval process, namely visual feature extraction, similarity measurement, and query processing with relevance feedback We have drawn attention to our ontology of an image, in which the 'whole of the image' supervenes on visual features, and the visual features supervene on the pixels. We have noted that many image descriptors—features—used in current CBIR systems do *not* have the property that the 'whole of the image' supervenes on them. We suggest that it it is in fact desirable that features employed in CBIR be true visual features, and that the supervenience relation is a useful way of thinking about and making this distinction.

We also made clear that RF strategies do not compare images on the level of information need. Perhaps one of the reasons the 'semantic gap' persists.

We followed these discussions with brief thought experiments addressing convention and imagination. These further highlight the importance of the beholder in determining the meaning of an image. We conclude that to overcome the 'semantic gap', the human, with all its mental dispositions, should get a more interwoven role in the CBIR retrieval process, beyond current relevance feedback. We see potential for increasing the influence of the user, for example, in the measurement of similarity [23] and in a different representation of the retrieval results, perhaps without regular juxtaposition of images.

We believe that an interdisciplinary approach, such as that taken here, can lead to fruitful results. It has led to a deeper understanding of 'convention' and 'imagination', and given that understanding gained, we could derive some likely practical consequences. This paper has focused more on the philosophical aspects. To establish the case with computer scientists will require experimental results. In the section on future work we have sketched the experimental CBIR system we are developing to meet this requirement. We note that such interdisciplinary work is something of a balancing act: to great a depth of explanation and detail from the first discipline risks incomprehension by the latter, too shallow risks being seen as trivial by the former.

References

1. Floridi, L.: Glossary of technical terms. Web page (2004) Last access: 26/01/2006. http:/www.blackwellpublishing.com\\/pci/downloads/Glossary.pdf.
2. Gruber, T.: What is an ontology. Web page (1993) Last access: 26/01/2006. http://www-ksl.stanford.edu/kst/what-is-an-ontology.html.
3. Mautner, T.: The Penguin Dictionary of Philosophy. revised edn. London: Penguin Books (2000)
4. Zùñiga, G.L.: Ontology: Its transformation from philosophy to information systems. In: Second International Conference on Formal Ontology and Information Systems, Ogunquit, Maine, USA (2001) 187–197
5. Smeulders, A., Worring, M., Santini, S., Gupta, A., Jain, R.: Content-based image retrieval at the end of the early years. IEEE Transactions on Pattern Analysis and Machine Intelligence **22**(12) (2000) 1349–1380
6. Rocchio, J.: Relevance feedback in information retrieval. In Salton, G., ed.: The SMART retrieval system: experiments in automatic document processing. Prentice-Hall, Englewood Cliffs, US (1971) 313–323
7. Squire, D.M., Müller, W., Müller, H., Raki, J.: Content-based query of image databases, inspirations from text retrieval: inverted files, frequency-based weights and relevance feedback. In: The 11th Scandinavian Conference on Image Analysis (SCIA'99), Kangerlussuaq, Greenland (1999) 143–149
8. ten Brinke, W., Squire, D.M., Bigelow, J.: Supervenience in content-based image retrieval. In: Proceedings of the Third International Conference on Formal Ontology in Information Systems (FOIS2004), Torino, Italy (2004) 298–304
9. Swain, M.J., Ballard, D.H.: Color indexing. International Journal of Computer Vision **7**(1) (1991) 11–32
10. Daugman, J.G.: Two-dimensional spectral analysis of cortical receptive field profiles. Vision Research **20**(10) (1980) 847–856
11. Zhang, D., Lu, G.: A comparative study of Fourier descriptors for shape representation and retrieval. In: Proceedings of the 5th Asian Conference on Computer Vision, Melbourne, Australia (2002) 646–651
12. Zloof, M.: Query-by-Example: a database language. IBM Systems Journal **4** (1977) 324–343
13. Carson, C., Thomas, M., Belongie, S., Hellerstein, J.M., Malik, J.: Blobworld: A system for region-based image indexing and retrieval. In Huijsmans, D.P., Smeulders, A.W.M., eds.: Third International Conference On Visual Information Systems (VISUAL'99). Number 1614 in Lecture Notes in Computer Science, Amsterdam, The Netherlands, Springer-Verlag (1999)
14. Flickner, M., Sawhney, H., Niblack, W., Ashley, J., Huang, Q., Dom, B., Gorkani, M., Hafner, J., Lee, D., Petkovic, D., Steele, D., Yanker, P.: Query by image and video content: The QBIC system. IEEE Computer (1995) 23–32
15. Rui, Y., Huang, T.S., Ortega, M., Mehrotra, S.: Relevance feedback: A power tool in interactive content-based image retrieval. IEEE Transactions on Circuits and Systems for Video Technology **8**(5) (1998) 644–655
16. Müller, H., Müller, W., Marchand-Maillet, S., Pun, T., Squire, D.M.: Strategies for positive and negative relevance feedback in image retrieval. In: Proceedings of the 15th International Conference on Pattern Recognition, Barcelona, Spain (2000)
17. Doulamis, N., Doulamis, A.: Evaluation of relevance feedback schemes in content-based in retrieval systems. Signal Processing: Image Communication, In Press, Uncorrected Proof, Available online 21 December 2005. Accessed: 25/01/2006 (2005)

18. Lewis, D.K.: On the Plurality of Worlds. Blackwell Publishing, Oxford, UK (1986)
19. Johansson, I.: Inference rules, emergent wholes and supervenient properties. In: European Computing and Philosophy Conference (ECAP2005), Västerås, Sweden (2005)
20. Lewis, D.K.: Convention. Blackwell Publishing, Oxford, UK (1969)
21. Schrödinger, E.: Die gegenwartige situation in der quantenmechanik. Naturwissenschaften **23** (1935) 807–812, 823, 844–849 English translation: John D. Trimmer, Proceedings of the American Philosophical Society, 124, 323–38 (1980).
22. Ganter, B., Wille, R.: Formal concept analysis : mathematical foundations. Springer-Verlag, Berlin, Germany (1999)
23. ten Brinke, W., Squire, D.M., Bigelow, J.: Similarity: measurement, ordering and betweenness. In: Proceedings of the Eighth International Conference on Knowledge-Based Intelligent Information and Engineering Systems (KES2004). Special Session on Similarity Measures for Content-Based Multimedia Retrieval, Wellington, New Zealand (2004) 996–1003