Benchmarks for storage and retrieval in multimedia databases

D.A. Forsyth[a]

[a]Computer Science Division, U.C. Berkeley, Berkeley, CA94720

ABSTRACT

There is a substantial body of research on computer methods for managing collections of images and videos. There is little evidence that this research has had important impact on an any community yet. I use an invitation to speak on a topic on which I am not expert to air some opinions about evaluating image retrieval research. In my opinion, there is little to be gained in measuring current solutions with reference collections, because these solutions differ so widely from user needs that the exercise becomes empty. The user studies literature is not well enough read by the image retrieval community. As a result, we tend to study somewhat artificial problems. A study of the user needs literature suggests that we will need to solve deep problems to produce useful solutions to image retrieval problems, but that there may be a need for a number of technologies that can be built in practice. I believe we should concentrate on these issues, rather than on measuring the performance of current systems.

Keywords: Content based image retrieval, image databases, multimedia databases, computer vision, object recognition

1. WHAT IS THE TASK? RETRIEVAL VS. BROWSING VS. ORGANISING COLLECTIONS VS. IMAGE DATA MINING

Large collections of digital pictures seem to spring up quite easily. Some collections of pictures are being digitized in the hope of better conservation, easier distribution, and better access. Others are intrinsically digital; examples include individual collections of family photographs (which can be big, and digital); the web, which is a big, disorganised collection; home videos (again, some collections are big and many are now digital).

Tools for interacting with collections of documents or of data are now quite sophisticated. Typically, one can search a collection using various kinds of text matching; one can cluster collections of text; and one can use data mining techniques. Data mining involves using statistical fitting procedures to look for trends that were not previously known (this useful pastime used to be known as "exploratory data analysis", a less exciting name, and is sometimes called "data dredging" by those who disapprove). Generally, a significant component of the value of a collection comes from the presence of such tools. To see why this might be, imagine visiting a large secondhand book shop that has its books sorted by, say, the colour of the dust-jacket; even though the collection may be very large, it's hard to imagine that you'd use the shop unless you were desparate.

It is currently difficult to organize or search image collections in a satisfactory fashion, meaning they are somewhat analogous to a poorly organised bookshop. The difficulty lies in building appropriate representations of the image information. The usual strategy of people who own commercial picture collections is to try and annotate each picture by hand (or use others' annotations).[1–6] This has its difficulties: preparing a good text description of an image can be very difficult (for examples, see[3, 6]). Furthermore, some collections are enormous (collections of tens of millions of pictures are listed in,[2] a paper published before the whole digitization industry really got going). Indexing a large collection by hand involves a substantial volume of work. Furthermore, there is the prospect of having to reindex sections of the collection; for example, if a news event makes a previously unknown person famous, it would be nice to know if the collection contained pictures of that person.

Any technology that helps manage collections of pictures has a tremendous range of practical applications. One important tool is search — find me a picture matching these criteria — but this is by no manner of means the only need. In my opinion, search has been overemphasized by the content based image retrieval literature, and a number of other interesting activities — browsing, organising and image data mining — have not been sufficiently well studied. As we shall see, there is some evidence of user need for good tools for these activities.

We might wish to organise the pictures in a way that supports browsing, so that pictures with similar content are near to one another. We might wish to search for trends, or to have tools that identify important changes.

1.1. Some Application Areas

Planning and government: there is a lot of satellite imagery of the earth, which can be used to inform important political debates. For example, how far does urban sprawl extend? what acreage is under crops? how large will the maize crop be? how much rainforest is left?, etc. (e.g.[7]).

Military intelligence: satellite imagery can contain important military information. Typical queries involve finding militarily interesting changes — for example, is there a concentration of force? how much damage was caused by the last bombing raid? what happened today? etc. — occuring at particular places on the earth (e.g.[8–10]).

Stock photo and stock footage: commercial libraries — which often have extremely large and very diverse collections — survive by selling the rights to use particular images (e.g.[1–3]). Effective tools may unlock value in these collections by making it possible for relatively unsophisticated users to obtain images that are useful to them at acceptable expense in time and money.

Access to museums: museums are increasingly creating web views of their collections, typically at restricted resolutions, to entice viewers into visiting the museum (e.g.[11–13]). Ideally, one would want viewers to get a sense of what is at the museum, why it is worth visiting and the particular virtues of the museum's gift store.

Trademark and copyright enforcement: as electronic commerce grows, so does the opportunity for automatic searches to find violations of trademark or of copyright (e.g.[14–17]). For example, at time of writing, the owner of rights to a picture could register it with an organisation called BayTSP, who would then search for stolen copies of the picture on the web; recent changes in copyright law make it relatively easy to recover fines from violators (see `http://www.baytsp.com/index.asp`).

Managing the web: indexing web pages appears to be a profitable activity; the images present on a web page should give cues to the content of the page. Users may also wish to have tools that allow them to avoid offensive images or advertising. A number of tools have been built to support searches for images on the web using CBIR techniques (e.g.[18–20]). There are tools that check images for potentially offensive content, both in the academic and commercial domains (academic, see[21–23]; commercial include a product called "PORNSWEEPER", `http://www.mimesweeper.com/products/pornsweeper/default.asp`).

Medical information systems: recovering medical images "similar" to a given query example might give more information on which to base a diagnosis or to conduct epidemiological studies (e.g.[24–26]). Furthermore, one might be able to cluster medical images in ways that suggest interesting and novel hypotheses to experts.

Image data mining: the attraction of data mining is that one can go on "fishing expeditions" using large data sets. Sometimes a data mining method will suggest a genuinely useful or novel hypothesis that can be verified by domain experts. Many image collections could support a similar activity. For example, there is a large collection of digitised images of Buddhist art, which is collected together with geolocation data (where was the object found?) and various expert comments, etc. If we could recover representations from the images, we could look for, say, trends in the depiction of the human figure across both space and time.

The core issue for all of these applications is the nature of the underlying representation of the images. Once we have decided on a representation, it is relatively easy to search (by finding images with a representation like this); to organise (by putting images with similar representations near to one another); or to search for trends (by looking for relationships between components of representations).

1.2. Why study this subject

If one could extract useful representations of image content, one could probably build quite helpful image retrieval systems. However, one could build much more — it would be possible to organise collections containing images more effectively, to browse such collections more effectively, and to mine them for new information. This means that there are two reasons to study the subject: firstly, even quite small advances may lead to genuinely useful practical artefacts; and secondly, very little is known about how to obtain the kinds of representations one wants to extract.

We will deal with the first point below; the second has not received enough attention. Computer vision is the study of computer programs that interpret images. It has been astonishingly successful in some important matters; for example, it is now easy to build convincing models of complicated objects from small numbers of photographs (e.g.[27,28]). However, object recognition remains very poorly understood. People can name many thousands of different kinds of object. This facility is not affected by superficial changes in individual objects — for example, disrupting the spot pattern on a cheetah, or changing the upholstery on, or the design of, a chair. Furthermore, people need to see only very few examples of a new object to "get it", and be able to recognize other instances of this object at some later date.

We would very much like to know how to build computer programs that, even partially, shared these skills. People probably posess them because they have practical value (knowing what to eat, who owes you food, when to fight, when to flee, what is going to eat you, etc.). The key matter seems to be one of building object representations that behave well when there are large numbers of different objects to be recognised.

We know how to recognise objects whose geometry is tightly controlled (e.g.[28]) and objects whose appearance is strongly constrained (the best examples are the successful face-finding literature,[29–32] and the literature on handwritten digit recognition[33]). We don't know how to mimic human recognition. The very little evidence we possess suggests that the key issues are extracting appropriate representations from images and organising the process of representation to yield good object hypotheses efficiently (e.g. [28]). This means that organising image collections poses some valuable model problems that require us to study issues we don't fully understand, but should.

All this means that a piece of work could be good because it may lead to a practical system, or because it may advance our understanding of object recognition. Each criterion leads to somewhat different methods of evaluation.

2. EVALUATING PRACTICAL SYSTEMS

The need for practical systems is clear, and an evaluation mechanism is clear, too — does it do what people need? In my opinion, our discipline does poorly by this criterion.

2.1. Mechanisms of Evaluation

2.1.1. Qualitative evaluation

It used to be common to build a system, and then show some responses to some queries, while claiming that these responses demonstrated the system works. This practice is a response to the (genuine) difficulty of evaluating practical systems. It is fortunately in decline.

2.1.2. Recall and precision

It is increasingly common to publish claims about recall and precision; these may appear in the form of statistics or of curves, typically with precision plotted against recall. Such experiments are very difficult to evaluate, because their meaning depends both on the method used to determine relevance, and on the quality of the survey one performs on the collection to score recall. As an example of how things can go wrong, assume that an experimenter wishes to determine how well a system for obtaining images that meet a broad semantic criterion performs. One strategy for doing this is to mark, in advance, every image in the collection that meets this criterion. The experimenter would then run the system in various configurations and evaluate recall and precision by determining how many marked images had been recovered, etc. The problem is that it is difficult to mark the images accurately, because there are lots of them and the semantic criterion is broad. Furthermore, the experimenter gets a reward — recall will tend to look better at fixed precision — if too few images are marked.

This difficulty is universal, and it is a constant nuisance in evaluating the claims of Internet filter manufacturers. In the case of filters, it is very difficult to measure the percentage of material that should not have been blocked by the filter, but was — this is because, to do so, one would need to look at an intractably large set of information. Furthermore, manufacturers who conduct this test poorly can make more impressive claims for the performance of their product.

2.1.3. Reference collections

One way to sidestep this difficulty is to construct a large reference collection with images annotated to indicate their content. Different research groups can then compare systems for queries where the annotations can be used to determine relevance. There are several efforts of this form; good start points for this literature are.[34, 35]

I don't see this form of activity as being particularly helpful, though it does allow a comparison between systems. The first difficulty is that determining relevance from annotations can be tricky; furthermore, the images need to be fairly comprehensively annotated to yield an helpful estimates. A more significant difficulty is that there seems to be little point in comparing systems that really don't meet user needs — which is true of pretty much all that is available at present. Do we really need to know how much worse system a is than system b? Yet more significant is the difficulty in constructing these reference collections — and the scoring process — to reflect practical application needs. Finally, there is a significant prospect of biasing research toward good performance on reference collections, as opposed to meeting application needs.

2.1.4. User studies

What users want to do with a system has a significant effect on how we should evaluate it. Recall and precision don't usually tell us all that much about whether a system is useful or not. However, high recall or precision numbers do not mean that a system works well. This is best illustrated by example. We consider systems that are built for three different purposes:

- Analysis of audit logs: a manager would like to deter employees from viewing sexually explicit pictures using the firms equipment. The manager does this by logging a copy of every picture that comes into the intranet, and running a program that analyses this log and informs the manager of any potential problems by displaying all suspect images. Such a system would not work if there were many false alarms — the manager would get bored and turn it off — but it doesnt need to find every problem image. The manager wants to be sure that if there are many problem images, there will be an alarm. In this case, one wants high precision, and low recall is not a problem (because if we mark only one in ten problem images, any large quantity of images is going to result in an alarm).
- Litigation support: a patent lawyer wishes to search the literature for prior art to overturn an existing, and valuable, patent. Only one item is required to do so, but it may be obscure. Furthermore, it is relatively cheap to hire consultants to read all items that are obtained. In this case, one wants high recall, and can tolerate low precision.
- Stock photo retrieval: a company supplies news organizations with photographs. They may posess 100, 000 pictures that are relevant to the query the late princess of Wales; typically, a client would like to receive at most a few dozen pictures. In this case, high recall could be a serious nuisance, but high precision is important

An ideal that is far more often promoted than espoused is to incorporate user studies into a design loop. There is little prospect that this will happen in the foreseeable future. However, researchers in the computer science community should have a clear understanding of the main messages of the user studies literature. This literature is not as familiar as it should be, and I'll sketch some high points below.

2.2. What do users want?

The literature on the behaviour of users of image collections is quite small. The most comprehensive study of the behaviour of users of image collections is Enser's work on the then Hulton-Deutsch collection[1, 2, 4] (the collection has been acquired by a new owner since these papers were written, and is now known as the Hulton-Getty collection). This is a collection of prints, negatives, slides and the like, used mainly by media professionals. Enser studied the request forms on which client requests are logged. Enser and Armitage extended these studies to cover seven libraries which collect both stills and video.[3]

Other user studies include the work of
Ornager,[36] who studied practice at a manually operated newspaper photo archive and Markkula and Sormunen[6] who study practice at a Finnish newspaper's digital photo archive. In the case of the digital archive, archivists index images based on captions and their experience, and users have the option of text searches or of contacting an archivist. Keister studied requests received by the National Library of Medicine's Archive.[37] Frost *et al* study the behaviour of users of a collection of art images; users could engage in keyword search or browse.[5] Typically, knowledgeable users searched and casual users browsed, but both classes found both types of interaction useful.

2.2.1. Semantics

Typically, users query images on semantics. For example, Enser classified requests into four semantic categories, depending on whether a unique instance of an object class is required or not and whether that instance is refined. Significant points include the fact that the specialised indexing language used gives only a "blunt pointer to regions of the Hulton collections" (,[1] p. 35) and the broad and abstract semantics used to describe images. For example, users requested images of hangovers, physicists and the smoking of kippers. All these concepts are well beyond the reach of current image analysis techniques. A recent paper of Enser's deals with the disparity between user needs and what technology supplies. The paper makes hair-raising reading[38]; for example, he cites a request to a stock photo library for "Pretty girl doing something active, sporty in a summery setting, beach - not wearing lycra, exercise clothes - more relaxed in tee-shirt. Feature is about deodorant so girl should look active - not sweaty but happy, healthy, carefree - nothing too posed or set up - nice and natural looking".

In the user studies, authors break out the semantics of the images requested in different ways, but from our perspective the important points are:

- that users request images both by object kinds (i.e. a princess) and identities (i.e. the princess of Wales);
- that users request images both by what they depict (i.e. things visible in the picture) and by what they are about (i.e. concepts evoked by what is visible in the picture;
- that queries based on image histograms, texture, overall appearance, etc. are vanishingly uncommon;
- and that text associated with images is extremely useful in practice — for example, newspaper archivists index largely on captions.[6]

2.2.2. Browsing

One interaction technique that hasn't received as much attention as it deserves from our community is browsing, with the important exception of.[39] There are number of interesting technical problems associated with browsing. Firstly, there are the mechanics: one must transmit and display images efficiently; one must lay out the display — multidimensional scaling seems to be well liked here — and one must choose what to display. Secondly, there is the question of how to organize a large collection of images to support browsing. One might, as Frost *et al.* did, use metadata[5]; one might use textual data; or one might attempt to cluster on a combination of image and text data, as Barnard *et al.* do.[40] No option is clearly best, but it is clear that one can't simply scroll through hundreds of thousands of images — some form of structure is required.

2.2.3. Simple collection management tools

As Peter Enser pointed out at the recent NSF-INRIA meeting, shot boundary detection tools are extremely helpful to video archivists and are beginning to be quite widely adopted. Because we all know lots of shot boundary detection algorithms, we tend not to regard research on such tools as interesting. However, the tools themselves are useful in real applications; one might expect that a closer study of the user studies literature would yield some more examples of simple technologies with high potential impact.

2.2.4. Scale

Real archives have collections of the order of millions to tens of millions of pictures; our research collections have of the order of tens of thousands of images *or fewer.* This disparity of scale — which should be a source of real anxiety about the practical significance of much research in the area — is so large that there is no reason to believe that results transfer across the scales.

3. EVALUATING IMAGE RETRIEVAL RESEARCH AS COMPUTER VISION RESEARCH

Image retrieval research might be good as computer vision research (rather than for its practical impact). What makes good computer vision research? There seem to be a small number of ways in which computer vision research is good. **Practical impact:** for example, recent work in structure from motion[27]; but we have assumed that we are assessing work not distinguished for its practical impact. **Illuminating useful technique:** some computer vision research is significant because it helps the community understand new techniques. It is hard to see that much image retrieval work should be counted under this category now.

Illuminating computer vision problems: finally, a piece of work may be important because it shows us — or forces us — how to think about important computer vision problems. Image retrieval has great unrealised potential here. The problem forces us to think about deep issues that are not well understood — how to represent many different objects at a semantic level; how to recover object representations from complex, unsegmented images; and how to link image representations with textual representations, to name a few — and allows us to make mistakes. There is no practical alternative to a computer program in many of the applications we have described. This means that a poorly working program is better than nothing. We would spend our time better if we tried to deliver what people wanted, however poorly our solutions worked, than if we tried to rank our solutions according to some explicit but arbitrary measure.

4. SUMMARY

I have sketched some ways of evaluating image retrieval research. In my opinion, there is little to be gained in measuring current solutions with reference collections, because these solutions differ so widely from user needs that the exercise becomes empty. A study of the user needs literature suggests that we will need to solve deep problems to produce useful solutions to image retrieval problems, but that there may be a need for a number of technologies that can be built in practice.

ACKNOWLEDGMENTS

The opinions expressed in this work were inculcated during the period of two NSF digital library initiative awards to U.C. Berkeley.

REFERENCES

1. P. Enser, "Query analysis in a visual information retrieval context," *J. Document and Text Management* **1**(1), pp. 25–52, 1993.
2. P. Enser, "Pictorial information retrieval," *Journal of documentation* **51**(2), pp. 126–170, 1995.
3. L. Armitage and P. Enser, "Analysis of user need in image archives," *Journal of Information Science* **23**(4), pp. 287–299, 1997.

4. P. Enser and C. Mcgregor, “Analysis of visual information retrieval queries,” tech. rep., British Library R+D Report 6104, 1992.
5. C. O. Frost, B. Taylor, A. Noakes, S. Markel, D. Torres, and K. M. Drabenstott, “Browse and search patterns in a digital image database,” *Information retrieval* **1**, pp. 287–313, 2000.
6. M. Markkula and E. Sormunen, “End-user searching challenges indexing practices in the digital newspaper photo archive,” *Information retrieval* **1**, pp. 259–285, 2000.
7. T. Smith, “A digital library for geographically referenced materials,” *Computer* **29**(5), pp. 54–60, 1996.
8. J. Mundy, “The image understanding environment program,” *IEEE Expert* **10**(6), pp. 64–73, 1995.
9. J. Mundy, “Iu for military and intelligence applications, how automatic will it get?,” in *25’th AIPR workshop. Emerging applications of computer vision, Proc SPIE*, **2962**, pp. 162–170, 1997.
10. J. Mundy and P. Vrobel, “The role of iu technology in radius phase ii,” in *Proc. Image Understanding Workshop*, pp. 251–64, 1994.
11. B. Holt and L. Hartwick, “’quick, who painted fish?’: searching a picture database with the QBIC project at uc davis,” *Information Services and Use* **14**(2), pp. 79–90, 1994.
12. B. Holt and L. Hartwick, “Retrieving art images by image content: the uc davis QBIC project,” in *ASLIB Proceedings*, **46**, pp. 243–8, 1994.
13. A. Psarrou, V. Konstantinou, P. Morse, and P. O’Reilly, “Content based search in mediaeval manuscripts,” in *TENCON-97 - Proc. IEEE Region 10 Conf. Speech and Image technologies for computing and telecommunications*, pp. 187–190, 1997.
14. J. Eakins, J. Boardman, and M. Graham, “Similarity retrieval of trademark images,” *IEEE Multimedia* **5**(2), pp. 53–63, 1998.
15. A. Jain and A. Vailaya, “Shape-based retrieval: a case study with trademark image databases,” *Pattern Recognition* **31**(9), pp. 1369–1390, 1998.
16. T. Kato, H. Shimogaki, T. Mizutori, and K. Fujimura, “Trademark: Multimedia database with abstracted representation on knowledge base,” in *Proc. Second Int Symp on Interoperable Information Systems*, pp. 245–252, 1988.
17. T. Kato and K. Fujimura, “Trademark: Multimedia image database system with intelligent human interface,” *Denshi Joho Tsushin Gakkai Ronbunshi* **72-DII**(4), pp. 535–544, 1989. Translated in Systems and Computers in Japan, V21 n11, 33-45, 1990.
18. M. L. Cascia, S. Sethi, and S. Sclaroff, “Combining textual and visual cues for content based image retrieval on the web,” in *IEEE Workshop on Content Based Access of Image and Video Libraries*, pp. 24–28, 1998.
19. S.-F. Chang, J. Smith, M. Beigi, and A. Benitez, “Visual information retrieval from large distributed online repositories,” *Comm. ACM* **40**(12), pp. 63–71, 1997.
20. J. Smith and S.-F. Chang, “Visually searching the web for content,” *IEEE Multimedia* **4**(3), pp. 12–20, 1997.
21. D. Forsyth and M. Fleck, “Automatic detection of human nudes,” *Int. J. Computer Vision* **32**, pp. 63–77, 199.
22. J. Wang, J. Li, G. Wiederhold, and O. Firschein, “Classifying objectionable websites based on image content,” in *Interactive distributed multimedia systems and telecommunication services*, *Lecture Notes in Computer Science* **1483**, pp. 113–124, Springer-Verlag, 1998.
23. J. Wang, J. Li, G. Wiederhold, and O. Firschein, “System for screening objectionable images,” *Computer Communications* **21**, pp. 1355–1360, 1998.
24. G. Congiu, A. D. Bimbo, and E. Vicario, “Iconic retrieval by contents from databases of cardiological sequences,” in *Visual Database Systems 3: Proc third IFIP 2.6 working conference on visual database systems*, pp. 158–74, 1995.
25. S. Wong, “CBIR in medicine: still a long way to go,” in *IEEE Workshop on Content Based Access of Image and Video Libraries*, p. 114, 1998.
26. P. Kofakis and S. Orphanoudakis, “Graphical tools and retrieval strategies for medical image databases,” in *Proceedings of the International Symposium on Computer Assisted Radiology*, pp. 519–524, Springer-Verlag, 1991.

27. R. Hartley and A. Zisserman, *Multiple View Geometry*, Cambridge University Press, 2000.
28. D. Forsyth and J. Ponce, *Computer Vision: a modern approach*, Prentice-Hall, 2001. in preparation.
29. H. Rowley, S. Baluja, and T. Kanade, "Neural network-based face detection," *IEEE T. Pattern Analysis and Machine Intelligence* **20**(1), pp. 23–38, 1998.
30. H. Rowley, S. Baluja, and T. Kanade, "Rotation invariant neural network-based face detection," in *IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 38–44, 1998.
31. H. Schneiderman and T. Kanade, "A statistical method for 3d object detection applied to faces and cars," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. I:746–751, IEEE press, 2000.
32. K.-K. Sung and T. Poggio, "Example-based learning for view-based human face detection," *IEEE T. Pattern Analysis and Machine Intelligence* **20**, pp. 39–51, 1998.
33. Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE* **86**(11), pp. 2278–2324, 1998.
34. H. Müller, W. M¨ller, D. M. Squire, S. Marchand-Maillet, and T. Pun, "Performance evaluation in content-based image retrieval: Overview and proposals," *Pattern Recognition Letters* , 2001.
35. M. Markkula, M. Tico, B. Sepponen, K. Nirkkonen, and E. Sormunen, "A test collection for the evaluation of content-based image retrieval algorithmsa user and task-based approach," *Information Retrieval* , pp. 275–293, 2001.
36. S. Ornager, "View a picture. theoretical image analysis and empirical user studies on indexing and retrieval," *Swedis Library Research* **2-3**, pp. 31–41, 1996.
37. L. Keister, *Challenges in indexing electronic text and images*, ch. User types and queries: impact on image access systems. Learned Information, 1994.
38. P. Enser, "Visual image retrieval: seeking the alliance of concept based and content based paradigms," *Journal of Information Science* **26**(4), pp. 199–210, 2000.
39. Y. Rubner, C. Tomasi, and L. J. Guibas, "A metric for distributions with applications to image databases," in *Int. Conf. on Computer Vision*, pp. 59–66, 1998.
40. K. Barnard and D. Forsyth, "Learning the semantics of words and pictures," in *Int. Conf. on Computer Vision*, pp. 408–15, 2001.