A method for comparing content based image retrieval methods

Kobus Barnard[a] and Nikhil V. Shirahatti[b]

[a]Department of Computer Science, University of Arizona
kobus@cs.arizona.edu

[b]Department of Electrical and Computer Engineering, University of Arizona
nvs@ece.arizona.edu

ABSTRACT

We assume that the goal of content based image retrieval is to find images which are both semantically and visually relevant to users based on image descriptors. These descriptors are often provided by an example image--the query by example paradigm. In this work we develop a very simple method for evaluating such systems based on large collections of images with associated text. Examples of such collections include the Corel image collection, annotated museum collections, news photos with captions, and web images with associated text based on heuristic reasoning on the structure of typical web pages (such as used by Google(tm)). The advantage of using such data is that it is plentiful, and the method we propose can be automatically applied to hundreds of thousands of queries. However, it is critical that such a method be verified against human usage, and to do this we evaluate over 6000 query/result pairs. Our results strongly suggest that at least in the case of the Corel image collection, the automated measure is a good proxy for human evaluation. Importantly, our human evaluation data can be reused for the evaluation of any content based image retrieval system and/or the verification of additional proxy measures.

Keywords: content based image retrieval, Benchathlon, performance, benchmarking, image semantics

1. INTRODUCTION

The field of content based image retrieval (CBIR) has generated much interest in the past decade[1 2-8], and a large scale benchmarking effort is underway[9,10] (see also http://www.benchathlon.net). The approach here generally revolves around creating a dataset of images with appropriate semantic labels which then can be used to test CBIR systems. A slightly different performance evaluation approach[11] uses images with manually classified regions and thus a region sensitive CBIR system can evaluated based on these labels.

These approaches have two problems which we address here. The first is that effective manual labeling and annotation of the images is difficult and time consuming and therefore there will always be a paucity of data despite the commendable efforts of those who offer such data to the community. Furthermore, even specifying the vocabulary or semantic classes is intellectually difficult[12].

This hints at the second problem which is that these approaches assume a sound relationship between the activity of retrieving images that match these annotations/labels and the activity of retrieving images which satisfy the user. We argue that this relationship must be considered, and we demonstrate incorporating it into retrieval benchmarking.

Some have argued that benchmarking image retrieval is premature[13]. Bluntly, it could be argued the current systems do not even come close to serving real user needs, so time spent measuring them is better spent doing something else. In this work we take an intermediate stance. It is important to realize how far off the mark we are[14-17], but this should encourage us to set up tests which move the systems in the direction of real utility. One thing which is clear is that semantics count[14-18]. For example, a user who would like an image of a tiger will not be satisfied with an image whose histogram matches a tiger; the pieces need to be arranged in the shape of a tiger. This intuition, is, of course, what leads to the idea of semantic intermediates for benchmarking in the approaches mentioned above.

In this work we consider whether text associated with images (but not produced as part of a benchmarking effort) can be an adequate intermediary for performance characterization. Such text is available in significant quantities. Examples include the Corel image collection (4-5 keywords for 40,000 images), a data set that we have been studying courtesy of the Fine Arts Museum of San Francisco (meta data as some descriptive text for 83,000 images), news photos with captions (of the order of 1,000 a day available on line), and, of course, the web at large, with millions of images which have associated text, a workable part of which can be extracted using a variety of heuristics [3,19].

A natural objection to using this kind of text is that it is typically incomplete and inaccurate. A case can also be made that an exact description of what is pertinent about an image depends on the seeker. However, we argue that these considerations may not be important for method comparison because all algorithms will have the same handicap. If that is the case, we may learn more by evaluating our queries on 40,000 images with adequate keywords rather than 400 images where the semantics has been carefully specified. Developing measures based on the this approach is discussed in §3.

To test the efficacy of this approach we investigate the relationship between the above score and human evaluation of retrieval results. In §4 we discuss our approach to collecting this data. We are careful to collect it in a form so that it can be used to measure any CBIR method. In fact it is our intention to collect, and eventually make available, data embodying human evaluation of CBIR relative to several data sets. For this paper, however, the focus is to gauge the automated measure based on associated words and to determine whether it can be used as a proxy for the human evaluation.

This work is essentially independent of any particular CBIR system. A CBIR system is needed to set up the experiments, but the results can be used to evaluate any other method. It is possible that there is a small bias towards the system used to set up the experiments, and therefore in future work we intend to use combine the results of several systems. This fits well with the obvious next step which is to test a number of systems. However, in this work we use two variants of a system readily available to us which we review in the next section in order to facilitate the exposition.

2. OUR CBIR SYSTEM

The CBIR system used in this work is an application of the system developed for modeling the joint probability of image region features and associated text[20]. It is not necessary to train the model on both text and image data, and we use two variants of the model—one where both text and image data is used, and one where only image data is used. We stress that the results in this paper are independent of the CBIR system(s) used to set up the test. Nonetheless, some details of the one we used are in order.

The model for the joint probability of image regions and text was inspired by one proposed for text alone by Hofmann[21,22]. It is a hierarchical combination of the assymetric clustering model which maps documents into clusters, and the symmetric clustering model which models the joint distribution of documents and features (the “aspect” model). The data is modeled as being generated by a fixed hierarchy of nodes, with the leaves of the hierarchy corresponding to clusters. Each node in the tree has some probability of generating each word, and similarly, each node has some probability of generating an image segment with given features. The documents belonging to a given cluster are modeled as being generated by the nodes along the path from the leaf corresponding to the cluster, up to the root node, with each node being weighted on a document and cluster basis. Conceptually a document belongs to a specific cluster, but given finite data we can only model the probability that a document belongs to a cluster, which essentially makes the clusters soft.

The model is illustrated further in Figure 1. To the extent that the sunset image illustrated is in the third cluster, as indicated in the figure, its words and segments are modeled by the nodes along the path shown. Taking all clusters into consideration, the document is modeled by a sum over the clusters, weighted by the probability that the document is in the cluster. Mathematically, the process for generating the set of observations D associated with a document d can be described by

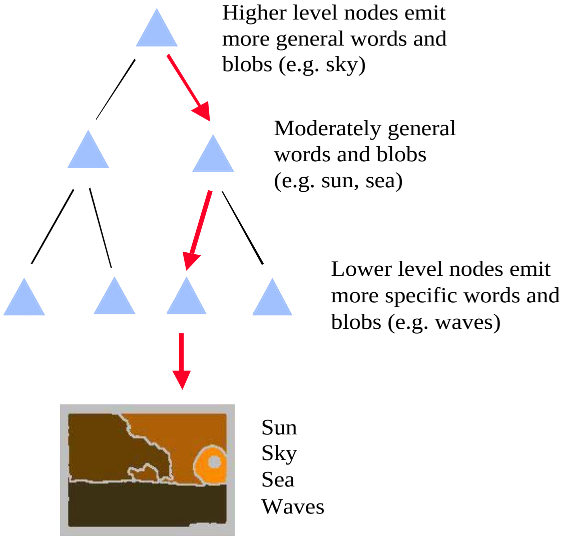

Higher level nodes emit more general words and blobs (e.g. sky)
Moderately general words and blobs (e.g. sun, sea)
Lower level nodes emit more specific words and blobs (e.g. waves)
Sun
Sky
Sea
Waves

Figure 1. Illustration of the generative process implicit in the statistical model. Each document has some probability of being in each cluster. To the extent that it is in a given cluster, its words and segments are modeled as being generated from a distribution over the nodes on the path to the root corresponding to that cluster.

$$P(D \mid d) = \sum_{c} P(c) \prod_{i \in D} \left(\sum_{l} P(i \mid l,c) P(l \mid c,d) \right)$$

where c indexes clusters, i indexes items (words or image segments), and *l* indexes levels. Notice that D is a set of observations that includes both words and image segments.

In (1) there is a separate probability distribution over the nodes for each document. Although this leads to difficulties in other applications such as attaching words to images (auto-annotate), or region labelling[23,24], is makes sense for search applications.

The probability for an item, $P(i \mid l,c)$, is conditionally independent, given a node in the tree. A node is uniquely specified by cluster and level. In the case of a word, $P(i \mid l,c)$ is simply tabulated, being determined by the appropriate word counts during training. For image segments, we use Gaussian distributions over a number of features capturing some aspects of size, position, colour, texture, and shape. These features taken together form a feature vector X. Each node, subscripted by cluster c, and level *l*, specifies a probability distribution over image segments by the usual formula. In this work we assume independence of the features, as learning the full covariance matrix leads to precision problems. A reasonable compromise would be to enforce a block diagonal structure for the covariance matrix to capture the most important dependencies.

To train the model we use the Expectation-Maximization algorithm[25]. This involves introducing hidden variables $H_{d,c}$ indicating that training document d is in cluster c, and $V_{d,i,l}$ indicating that item i of document d was generated at level l. Additional details on the EM equations can be found in [21].

Both the vertical and horizontal structure of the nodes are important. The vertical structure allows the generation of images of a variety of combinations of components (aspects, topics) without encoding all possibilities. The horizontal structure provides the clustering and modeling economy, which also facilitates training. The hierarchical structure provides further economy, but more importantly, can exploit the expected hierarchical nature of data, and similarly can learn some of that structure. Specifically, more general terms and more generic image segment descriptions will occur in the higher level nodes because they occur more often.

3.1 Image segmentation and feature extraction

For segmentation we use Normalized Cuts[26]. For this work we use a modest set of features, specifically region color and standard deviation, region average orientation energy (12 filters), and region size, location, convexity, first moment, and ratio of region area to boundary length squared.

3.2. Retrieval

Our approach to searching is to compute the probability of each candidate image of emitting the query items. Defining search by the computation of probabilities very naturally yields an appealing soft query system. Given a set of query items, Q, and a candidate document d, we can express the probability that the document produces the query by:

$$P(Q\,|\,D,d)=\frac{P(Q,D)}{P(D)}=\sum_{c}P(Q\,|\,c,d)\,\frac{P(D\,|\,c,d)P(c)}{P(D)}=\sum_{c}\left\{\prod_{i\in Q}\left(\sum_{l}P(i\,|\,l,c)P(l\,|\,c,d)\right)P(c\,|\,D,d)\right\}$$

Documents are ranked according to the above score. A second approach is to find the probabilities that the query is generated by each cluster, and then sample from the clusters, weighted by the probability that the cluster emits the query. This often works reasonably well because cluster membership plays the dominant role in generating the documents, which simply reflects the fact that the clusters are coherent. Nonetheless, we have found that the results using the above equation are sometimes significantly better, and rarely worse, and this is the method used in this work.

We have implemented search for arbitrary combinations of words and image features. To use the system for CBIR, and more specifically for "query by image example", the query image is processed in the same manner as all the training data, and then a query is made based on all the regions of the query image. Specifically, Q above is a list of the segments in the query image specified by their features.

3. HUMAN EVALUATION OF CBIR OUTPUT

To evaluate CBIR systems we present a subject with query and corresponding result image pairs. The subject evaluates each pair as either "undecided", "poor match", "faint match", or "good match". Thus we are evaluating the "query by image example" paradigm. (We are currently extending the method to text queries). In our implementation, the user evaluates 20 pairs for each query image. The subject was given very little in the way of guidelines for making their selection. Figures 2 and 3 show the interface.

The main difficulty in setting up such an experiment is that the query pairs cannot be randomly generated. If they were randomly generated, then nearly all the matches would be judged "poor match". Ideally, we would like roughly even numbers of the 3 choices. To get more balance among the results we use the CBIR system to score all possible matches, and then select evenly from the result. This still leads to many poor matches, so we put our score through a nonlinear function (specifically we raise it to the one third power). This makes the of match distribution more usable, but we are considering providing even better normalization of this space through an iterative process.

Our scoring scheme is very simple. "undecided" is ignored, "poor match" scores 0, "faint match" scores 1/2, and "good match" scores 1. Importantly, the data set can be used to rate any CBIR system, not just the one used to select the pairs presented to the subject. Furthermore, such data contains information which can be used to model CBIR users.

4. AUTOMATED EVALUATION OF CBIR OUTPUT

We propose considering evaluating "query by example" CBIR methods by simply comparing the word sets of the query image to that of the retrieved image. If we denote the set of words associated with the query image as W_Q, and the set of words associated with the retrieved image as W_R, and the number of elements in a set W by $|W|$, then our score is given by:

$$score=\frac{|W_Q \cap W_R|}{\min(|W_Q|,\ |W_R|)}$$

To evaluate a system, one could either weight the scores by a function of rank, or simply average the results of a typical display set such as the top 20 matches.

5. THE KEY EXPERIMENT

We have discussed two ways of evaluating CBIR systems. The human evaluation is more important, but it is difficult to collect in large quantities. The automated word based method is painless, but less accurate. We would like to use the automated method in place of, or in conjunction with, the human evaluation method. Thus we ask how the two methods relate.

We report the results of two experiments. In the first, query image / retrieval image pairs generated by our CBIR system were trained on both words and image region features, and in the second experiment only image features were used. The results are based on examining 2860 pairs for the first experiment, and 4000 pairs for the second. The simple scoring method described above yields only a handful of possible automated scores because the number of words in our data set is at most 5 and typically less. For each such score, we tabulate the average of the human score on all images receiving that automated score. The results shown in Table 1 clearly indicated that we have a solid relationship between the two methods, at least for this data set. Note that we do not require a linear relationship, only one which is distinctly monotonic.

	Experiment One		Experiment Two	
Automated word based score	Average human evaluation score	Fraction of pairs falling in this range	Average human evaluation score	Fraction of pairs falling in this range
0.0000	0.109	0.721	0.080	0.713
0.2500	0.287	0.040	0.258	0.046
0.3333	0.392	0.117	0.352	0.110
0.5000	0.412	0.048	0.382	0.061
0.6667	0.648	0.023	0.555	0.025
0.7500	0.864	0.004	0.900	0.004

Table 1. The average human evaluation score as a function of the common word score which is naturally quantized into the 6 values above due to the nature of the measure and the limited number of words for each image in the Corel data set. The results show clearly that there is a strong relationship between human evaluation and our simple score based on the common words associated with the test and query images.

6. CONCLUSION

This preliminary study strongly suggests that associated text can be used to evaluate CBIR programs, at least in the case of image collections like the Corel data set. Work is underway to expand the boundaries of the work, both with respect to the variety of data sets, and also to text queries. Our human based evaluation approach, while tedious, provides a tractable way to gather lots of data on user evaluate query results. We note that relevance feedback systems typically have access to this kind of information and could collect it for the kind of use outlined in this paper. We suggest that such data, regardless of how it is obtained can be used to calibrate other CBIR evaluation methods, or simply evaluate other CBIR systems using the same data set. The combination of shared repository of human CBIR evaluation data, and calibrated proxy measurements such as the one proposed in this paper would provide a significant tool for progress in image retrieval work.

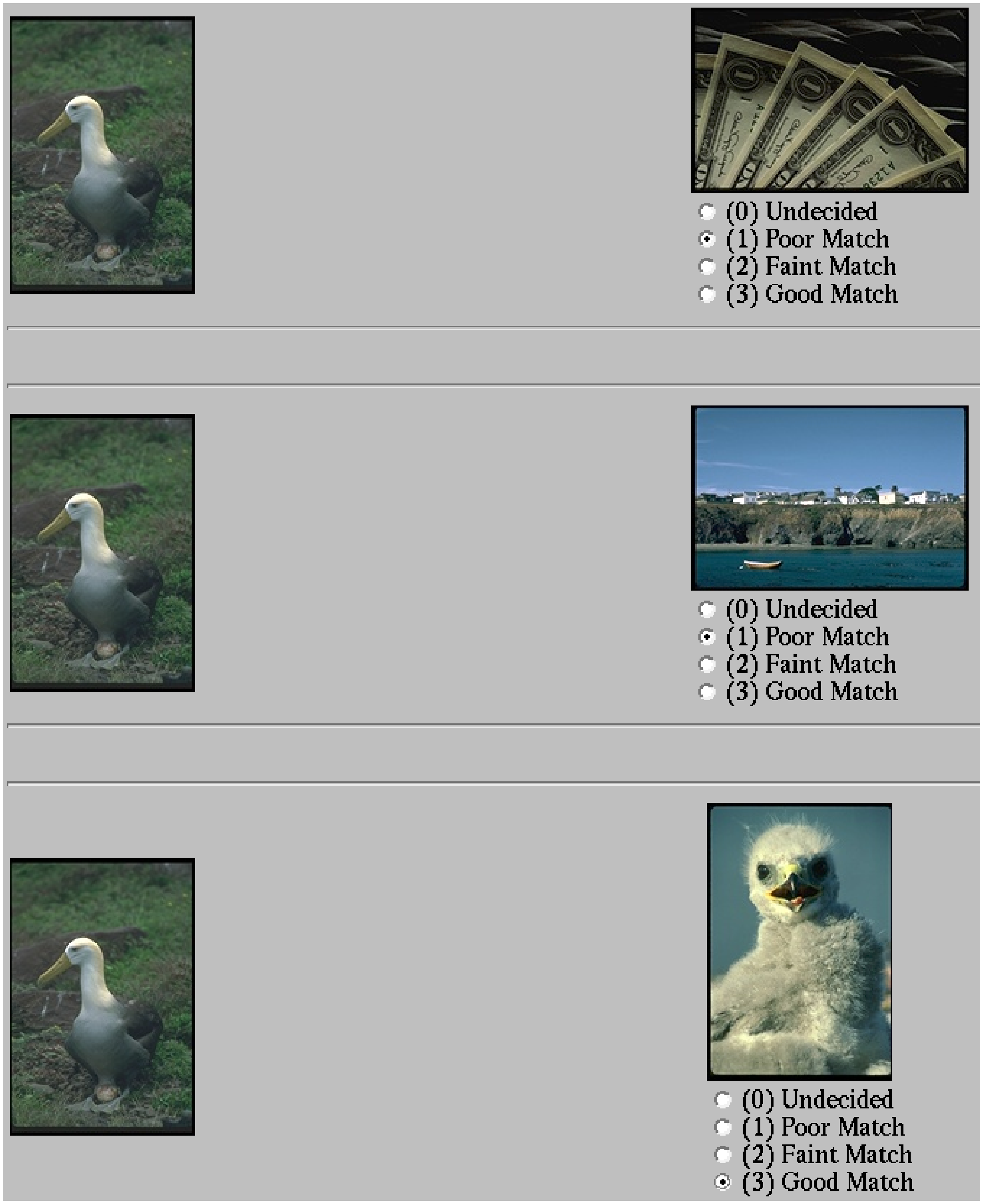
ACKNOWLEDGEMENTS

The CBIR system used for generating the query/result pair spacing was largely developed while Kobus Barnard was at the University of California at Berkeley. The UCB project was funded through the Digital Libraries Initiative sponsored by NSF and many others. While at UCB, Kobus Barnard also received funding from NSERC (Canada). We acknowledge the help of Pinar Duygulu in running the segmentation software, and we are grateful to Doron Tal and Jitendra Malik for the normalized cuts software.

REFERENCES

1. M. Flickner, H. Sawhney, W. Niblack, J. Ashley, Q. Huang, B. Dom, M. Gorkani, J. Hafner, D. Lee, D. Petkovic, D. Steele, and P. Yanker, Query by image and video content: The QBIC system, *IEEE Computer*, **28**, 22-32 (1995).

2. S. Sclaroff, L. Taycher, and M. La Cascia, ImageRover: A content-based image browser for the world wide web, *IEEE Workshop on content-based access of image and video libraries* (1997).
3. M. La Cascia, S. Sethi, and S. Sclaroff, Combining Textual and Visual Cues for Content-based Image Retrieval on the World Wide Web, *IEEE Workshop on Content-Based Access of Image and Video Libraries* (1998).
4. C. Carson, S. Belongie, H. Greenspan, and J. Malik, Blobworld: Image segmentation using Expectation-Maximization and its application to image querying, submitted for publication to IEEE PAMI. Available in the interim from http://HTTP.CS.Berkeley.EDU/~carson/papers/pami.html.
5. C. Carson, M. Thomas, S. Belongie, J. M. Hellerstein, and J. Malik, Blobworld: Image segmentation using Expectation-Maximization and its application to image querying, *Third Int. Conf. on Visual Information Systems* (1999).
6. J. Z. Wang, J. Li, and G. Wiederhold, SIMPLIcity: Semantics-Sensitive Integrated Matching for Picture Libraries, *IEEE Trans. Patt. Anal. Mach. Intell.*, **23**, 947-963 (2001).
7. I. J. Cox, M. L. Miller, T. P. Minka, T. V. Papathomas, and P. N. Yianilos, The Bayesian Image Retrieval System, PicHunter: Theory, Implementation and Psychophysical Experiments, *IEEE Transactions on Image Processing*, **9**, 20-35 (2000).
8. A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, Content-Based Image Retrieval at the End of the Early Years, *IEEE Transactions on Pattern Matching and Machine Intelligence*, **22**, 1349-1379 (2000).
9. N. J. Gunther and G. B. Beratta, Benchmark for image retrieval using distributed systems over the internet: BIRDS-I, *Internet Imaging III*, SPIE, 252-267 (2001).
10. T. Pfund and S. Marchand-Maillet, Dynamic multimedia annotation tool, *Internet Imaging III*, SPIE, 206-224 (2002).
11. J. Vogel and B. Schiele, On Performance Characterization and Optimization for Image Retrieval, *7th European Conference on Computer Vision*, Springer, 49-63 (2002).
12. C. Jörgensen and P. Jörgensen, Testing a vocabulary for image indexing and ground truthing, *Internet Imaging III*, SPIE, 207-215 (2002).
13. D. A. Forsyth, Benchmarks for storage and retrieval in multimedia databases, *Storage and Retrieval for Media Databases III*, SPIE (2002).
14. P. G. B. Enser, Query analysis in a visual information retrieval context, *Journal of Document and Text Management*, **1**, 25-39 (1993).
15. P. G. B. Enser, Progress in documentation pictorial information retrieval, *Journal of Documentation*, **51**, 126-170 (1995).
16. L. H. Armitage and P. G. B. Enser, Analysis of user need in image archives, *Journal of Information Science*, **23**, 287-299 (1997).
17. J. P. Eakins, Towards intelligent image retrieval, *Pattern Recognition*, **35**, 3-14 (2002).
18. S. Santini, Semantic Modalities in Content-Based Retrieval, *IEEE International Conference on Multimedia and Expo* (2000).
19. N. C. Rowe, Marie-4: A high-recall self-improving web crawler that finds images using captions, *IEEE Intelligent Systems*, **July/August**, 8-14 (2002).
20. K. Barnard and D. Forsyth, Learning the Semantics of Words and Pictures, *International Conference on Computer Vision*, II:408-415 (2001).
21. T. Hofmann, Learning and representing topic. A hierarchical mixture model for word occurrence in document databases, *Workshop on learning from text and the web* (1998).
22. T. Hofmann and J. Puzicha, Statistical models for co-occurrence data, Massachusetts Institute of Technology, A.I. Memo 1635, available from (1998).
23. P. Duygulu, K. Barnard, J. F. G. D. Freitas, and D. A. Forsyth, Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary, *The Seventh European Conference on Computer Vision*, IV:97-112 (2002).
24. K. Barnard, P. Duygulu, N. D. Freitas, D. Forsyth, D. Blei, and M. I. Jordan, Matching Words and Pictures, *Journal of Machine Learning Research* (in press).
25. A. P. Dempster, N. M. Laird, and D. B. Rubin, Maximum likelihood from incomplete data via the EM algorithm, *Journal of the Royal Statistical Society. Series B (Methodological)*, **39**, 1-38 (1977).
26. J. Shi and J. Malik., Normalized Cuts and Image Segmentation, *IEEE Trans. Patt. Anal. Mach. Intell.*, **22**, 888-905 (2000).

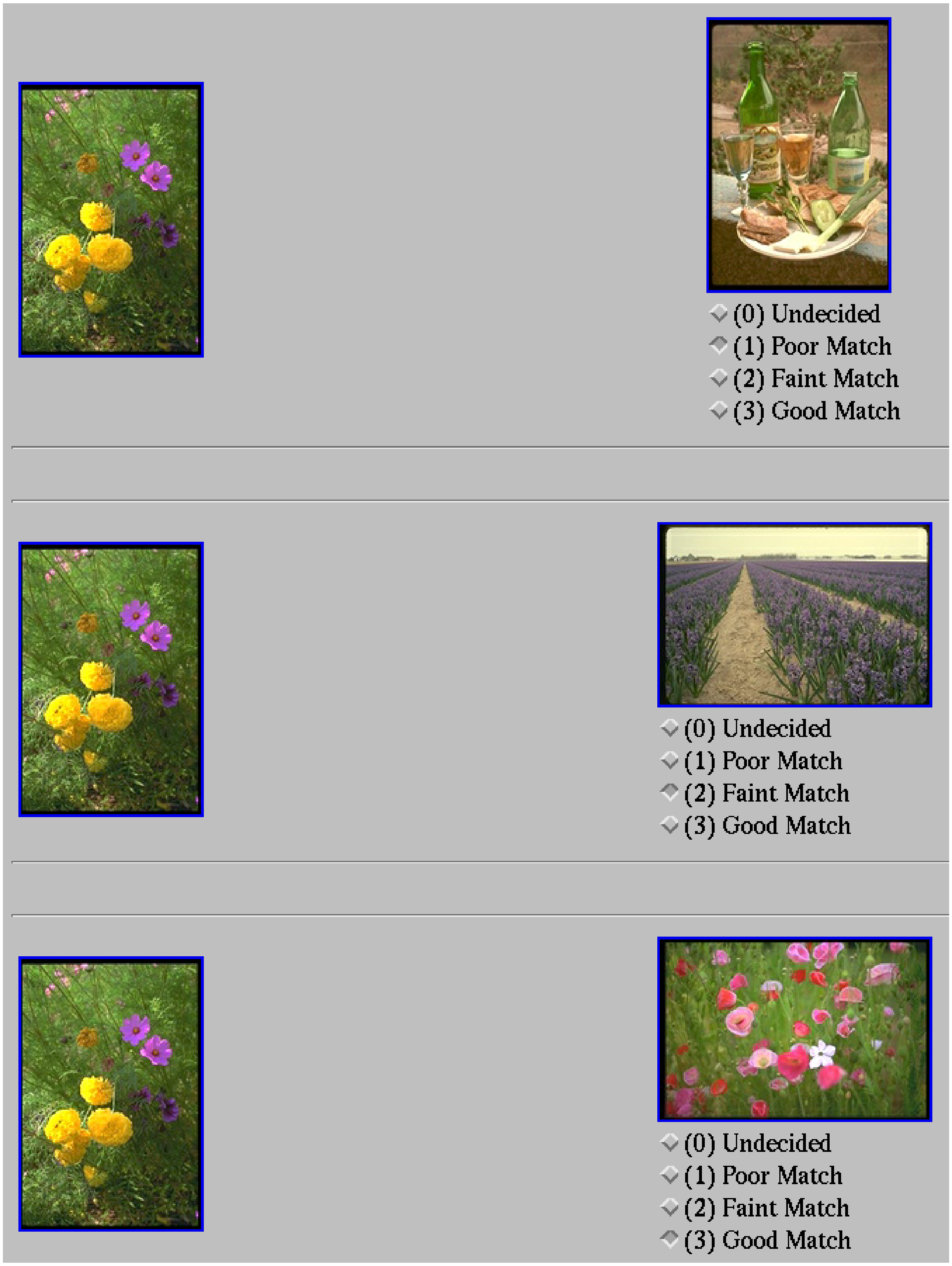


○ (0) Undecided
● (1) Poor Match
○ (2) Faint Match
○ (3) Good Match

○ (0) Undecided
● (1) Poor Match
○ (2) Faint Match
○ (3) Good Match

○ (0) Undecided
○ (1) Poor Match
○ (2) Faint Match
● (3) Good Match

Figure 2. A screen shot of the interface developed for gathering human CBIR evaluation data.



◇ (0) Undecided
◆ (1) Poor Match
◇ (2) Faint Match
◇ (3) Good Match

◇ (0) Undecided
◇ (1) Poor Match
◆ (2) Faint Match
◇ (3) Good Match

◇ (0) Undecided
◇ (1) Poor Match
◇ (2) Faint Match
◆ (3) Good Match

Figure 3. A second screen shot of the interface developed for gathering human CBIR evaluation data.