Evaluating Image Retrieval

Nikhil V Shirahatti
Electrical and Computer Engineering
University of Arizona
shirahatti@gmail.com

Kobus Barnard
Department of Computer Science
University of Arizona
kobus@cs.arizona.edu

Abstract

We present a comprehensive strategy for evaluating image retrieval algorithms. Because automated image retrieval is only meaningful in its service to people, performance characterization must be grounded in human evaluation. Thus we have collected a large data set of human evaluations of retrieval results, both for query by image example and query by text. The data is independent of any particular image retrieval algorithm and can be used to evaluate and compare many such algorithms without further data collection. The data and calibration software are available on-line (http://kobus.ca/research/data).

We develop and validate methods for generating sensible evaluation data, calibrating for disparate evaluators, mapping image retrieval system scores to the human evaluation results, and comparing retrieval systems. We demonstrate the process by providing grounded comparison results for several algorithms.

1. Introduction

The problem of automatically retrieving desired images from a large, often unstructured data set has attracted much attention in the research community [14, 26, 20, 7, 31, 8, 27, 21, 23]. The task is difficult and tightly connected to computer vision because users are interested in the semantics of the retrieved images [13, 12, 3, 22, 15].

These studies confirm that current image retrieval methods are well off the required mark. We argue that moving forward will require quantifying real performance, and that the image retrieval community will be well served by an appropriate evaluation process and reference data set. Thus we have made our data and software available on-line (http://kobus.ca/research/data).

Automated image retrieval is only meaningful in its service to human users, and thus performance must be grounded in direct human evaluations. Our approach is to evaluate query-result pairs for both query by image example and query by text. By focusing only on the input and output, such data is applicable to any image retrieval method.

Previous work. Often evaluation of image retrieval is focused on results obtained with a specific instance of a specific algorithm. With this approach, changes to the algorithm requires additional human evaluation, which is expensive. More automatic methods typically involve having sets of images tagged with high level concepts (e.g., *sky*, *grass*), and retrieval is evaluated based on those labels [28, 29, 30], making performance evaluation similar to that in text retrieval [25]. The Benchatholon project proposes providing much more detailed and publicly available keywords of images using a controlled vocabulary [16, 24, 1]. A problem with both these approaches is that they are only indirectly connected to the task that they are trying to measure. For example, there is an implicit assumption that a person seeking an image like one labeled *grass* will be content with all the images labeled *grass* and none of the ones not labeled *grass*. While we do not reject this hypothesis outright, image retrieval evaluations need to be grounded on tasks closer to what end-users do, hence this work. Our results can be used to calibrate these less expensive measures. We also acknowledge work on observer variance [19], especially in the case of judging medical data [9].

2. Developing a reference data set

For this study we use the Corel image data set which is arguably the most used data in image retrieval research. This fact alone suggests that a suite of comparison data sets should include data for those images. However, we remind the reader that the Corel data has significant problems. It is a particularly easy one for content based image retrieval because many of the CD's contain many images which are semantically similar, but are not too different in terms of the kinds of descriptors which current retrieval systems exploit. The Corel image data set also has copyright problems, and purchasing the same data as one's colleagues is becoming difficult.

We set up human retrieval evaluation experiments to

gather grounded data for two tasks: query by image and query by text paradigms. For the query by image paradigm we present the user with one query image and four result images (see Figure 1). The selection of the result images is discussed in detail in the subsequent paragraph. The participant was asked to score each of the four result images on a scale of 1 to 5, with 1 being a poor match and 5 being a good match. We provided an additional choice of *undecided* (ignored) so that participants could move onto the next example without spending too much time on ones they find hard to evaluate. Participants were given very little in the way of guidelines for making their selection. For the second interface, we presented the participant with a text query and a corresponding result image. They rated the match by selecting a score from 1-9 or *undecided*.

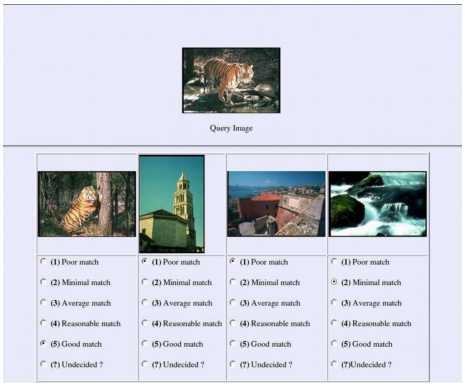

Figure 1. Screen shot showing the interface for gathering human image retrieval evaluation data for query by image example.

Avoiding too many negative matches. The main difficulty in setting up such an experiment is choosing query-result pairs. If they were randomly generated then nearly all the matches would be judged *poor match*. Ideally, we would like roughly a uniform distribution of the responses of the evaluations (excluding *undecided* where fewer is always better).

The main idea is to use existing image retrieval systems to help bias the selection process to get more uniform responses. Doing so may put us at risk for introducing unwanted biases in the test set due to some poorly characterized property of the retrieval system. While we do not expect significant problems, we guard against this by using four very different image retrieval process. Each one is used for the selection of one of the four result images, randomly permuted for each query.

The second issue is how to use the retrieval systems to improve the sampling. Initially, we know very little about the relationship between retrieval results and human evaluation results. However, trial and error revealed that choosing images with probability proportional to the negative fifth power of the rank gave a serviceable starting point. This can be improved once some data has been collected, as our approach revolves around estimating the mapping from computer scores to human scores. In §5.2 we present results which suggests that this iterative process is helpful.

It is critical to understand that the query-result pairs are evaluated completely independently of the retrieval systems used to help select the images. Ideally, the only effect of the selection process is that the responses are more uniformly distributed. Using four different allows us to address the whether the process introduces significant bias into the measurement of retrieval systems (§5.3).

Evaluation experimental protocol. We asked many people to evaluate query-result pairs. This achieves two goals. First, we are interested in the range of results due to human subjectivity. Second, we wanted to collect as much data as possible. We collected data for two experiments two paradigms: query by image and query by text. When the participant began the experiment they were first asked to provide a login string. If it was their first time for a particular experiment, they began by evaluating a common set of query-result pairs for that experiment. The query-result pairs in the common set for the each of the two interfaces was the same. After evaluating the common set, the query-result pairs evaluated by each participant was unique. It was convenient for the participants to stop at any point. If they logged on to the system with the same login string that they used initially, then the experiment proceeded from that point.

Due to practical considerations, roughly half of the data was produced by a single person. In total, 20,000 query-result pairs were evaluated for query by image example and 5,000 pairs were evaluated for query by text example. The evaluation was performed by 32 participants, out of which 3 participants evaluated both the paradigms. The data domain of this work is 16,000 images from the Corel data set.

Calibrating for participant variance. We used the data from the common sets to reduce the biases due to the different participants. To do so, we mapped the results of each participant in a given experiment by a single linear transformation so that their mean and variance of their results on the common set was the global mean and variance on this set. The effect of this is studied in (§5.1).

3. Image retrieval systems

Keyword retrieval. The Corel images have keywords, and these can be used as a pseudo query by example method. Here, we score the match of two images by:

$$score = \frac{|\, W_Q \cap W_R \,|}{min(|\, W_Q \,|, |\, W_R \,|)} \qquad (1)$$

where W_Q is the set of words associated with the query, W_R is and the set of words associated with the retrieved image,

and $|X|$ is the number of elements in a set X. We denote this retrieval method as "Keywords".

Regions based mixture models. Recent work proposes modeling image data as being generated by hidden factors which are responsible for jointly generating image region features and associated text [6, 4, 10, 5]. Here we model the joint probability of a particular blob, *b*, and a word *w*, as

$$P(w,b) = \sum_l P(w|l)P(b|l)P(l) \quad (2)$$

where *l* indexes over the concepts, $P(l)$ is the concept prior, $P(w|l)$ is a frequency table, and $P(b|l)$ is a Gaussian distribution (diagonal covariance) over features. Image clustering (in addition to the region clustering) can be integrated into this model, but we do not use that capability here. There are several options for training such models in the case of loosely labeled data where we don't have the correspondence between image words and image regions. To implement image retrieval, we compute the probability that the model parameters for a database document can generate the observed regions of the query document.

The model can be trained with both image region features and words (labeled "RWMM"), or using regions only ("ROMM"). For image retrieval, we only use the image feature part of the model. Thus, if words are used at all, it is only during training. We further consider two retrieval scenarios. The first assumes complete access to all data, and thus we are able to match images in our training set. While in most situations using the training set model is not interesting, in the retrieval context it can make sense. In this case we affix the suffix "ALL" to the method label.

The second scenario uses the model as a template for matching new images. Neither the query image, nor the result image are part of the training set. Here we affix the suffix "TEST" to the method label. In particular, the method RWMM-TEST seems like an interesting retrieval paradigm. The words help ensure that the model encodes some relationship between image features and semantics, but the model is applicable to matching images without keywords and that have not been seen by the training system — of course, regions with the appropriate semantics must be in the training data.

The variant used for image selection in the query-by-example experiment, "ROMM-CALIB" is an older version of the system which was trained without words on subsets of the entire image data set. The results were then concatenated. Image selection for the query-by-text case used the analogous method, but text was included ("RWMM-CALIB").

GIFT [2] is an open framework for content-based image retrieval. In its standard implementation, it is a pixel based CBIR system based on both local and global color and texture histograms. We use the standard system as one of the four systems used for improving the uniformity of the human evaluation results. In the retrieval method comparison (§5.4) we evaluate the effect of limiting the GIFT to use only color ("GIFT-color"), and only texture ("GIFT-texture").

SIMPLIcity [31] is a region-based CBIR system which combines semantic classification methods, a wavelet based approach for feature extraction, and an integrated region matching based on image segmentation.

4. Mapping retrieval algorithm scores to human evaluation scores

The ground-truth data is composed of human scores corresponding to pairs of query-result images from the evaluation data set. We want to use this data to provide a mapping which takes the image retrieval scores into human evaluation scores. Such mappings will put all systems onto the same scale, namely human evaluation scores. They also render retrieval scores as absolute scores which is useful for negotiating with users regarding the quality of the images to be returned (e.g., "good match" versus top 10).

Because retrieval systems vary widely in what they report, the mapping functions are necessarily very different from system to system. We propose that each system use the best general mapping that can be found. We only impose one constraint on the function, specifically that it is monotonic. Because image retrieval still has a very long way to go, the data that we need to fit is very noisy (see Figure 2).

Monotonic mapping minimizing squared error. If we measure the efficacy of our mapping by squared error, then we can find the best mapping of the data by constrained least mean square, which can be conveniently implemented using the Matlab© routine *quadprog*. Since the number of constraints is too large for timely completion of the fitting, we use bootstrapping [11].

Monotonic mapping maximizing correlation. Because the data is so noisy, correlation between the mapped scores and the human scores is a more useful error measure than squared error. Thus we measure the efficacy of our maps using correlation, which suggests maximizing correlation to find the mapping function. We do this with the Matlab© routine *fmincon*. Unfortunately, we are only guaranteed a local maximum, and again, a sampling strategy is needed to deal with the scale of our problem.

Monotonic Bayesian curve fitting. A third fitting method [17, 18] uses Markov Chain Monte-Carlo (MCMC) simulation to sample from a model space of fitting functions of varying dimensions corresponding to differing numbers of *change points* or *knots*. Monotonicity is constrained during the sampling. This approach runs fine on our entire data set, and often gives us the best mapping function (§5.3).

5. Experiments

5.1. Variance across evaluators

Interface	Query by	Query by text	
	1-5	Binary	1-9
Number of participants	24	6	5
Average variance with standardized scores	1.38	0.19	2.88
Average variance with person dependent adjustment	1.15	0.036	0.937

Table 1. The effect of adjusting on human evaluation scores to reduce differences among participants. The table shows the average standard deviation for standardized scores (global mean 0 and variance 1) for the three experiments before and after adjustment using the method described in the text (§2). This adjustment significantly reduces the variance.

Table 1 shows the average variance of the results for the common test set for each of paradigms with and without the normalization described in (§2). The results show that there is variability in the participants that is worth calibrating for. Thus we apply the transformation computed on the common set to adjust all the results from that participant.

5.2. Updating evaluation pair choice based on measured mapping functions

As described in §2, once we have a reasonable amount of evaluation data, we can use the retrieval system specific mapping functions (§4) to further improve the generation of query-result pairs for subsequent data collection. Recall that our goal is to have a roughly uniform response over our evaluation responses. A simple measure of this for 5 categories is $\frac{1}{5}\sum_{i=1}^{5}\|f(i)-0.20)\|$, where $f(i)$ is the fraction of responses for category i. We computed this measure for the responses from the sampling based on the initial proposal (negative fifth power of rank), and the responses from subsequent data based on the mapping functions computed from the first part. The results in Table 2 show that the second data set induced more uniform responses.

	1	2	3	4
initial data	0.20	0.19	0.14	0.08
mapped data	0.15	0.14	0.12	0.05

Table 2. Deviation from uniformity of human evaluation results for the four calibration retrieval systems: (1) GIFT; (2) SIMPLIcity; (3) ROMM-CALIB; and (4) Keywords.

5.3. Mapping CBIR system scores to human evaluation results

Figure 2 shows the data from the four systems used for calibration, and the mapping function found for each using the best curve fitting method. Table 3 provides the correlations between the mapped score and the adjusted human score for all three fitting methods. In order to investigate sources of bias, we computed results for each of the four calibration system evaluated using only the images selected by each of the four. We found no significant consistent trend that using the same algorithm for selection and testing is an advantage to that algorithm. For example, if we used the maximum of each of the three fitting methods, and allow each algorithm to be paired with its own selection results, then the rank order does not change compared to using the mean, or the value from all data.

In general, we find that the Bayesian fitting method gave consistently good results. The constrained correlation maximization method also gave serviceable results. In contrast, least squares fitting did not work very well, which is perhaps not surprising given that we settled for correlation as our main measure of interest.

5.4. Comparing image retrieval algorithms

To compare image retrieval algorithms we first find a good mapping of the scores of that algorithm on the evaluation set to the adjusted human scores as described above. We then compute the correlation of the mapped scores to the human scores. The results are in Table 4.

	Correlation of the calibrated human to the mapped system scores
ROMM-ALL	0.24
ROMM-TEST	0.17
RWMM-ALL	0.35
RWMM-TEST	0.23
GIFT	0.17
GIFT-color	0.15
GIFT-texture	0.07
SIMPLIcity	0.19
Keywords	0.51

Table 4. Grounded comparison of content based retrieval methods. We report the correlation of mapped computer scores with human scores. Each method uses its own, most favorable, monotonic mapping.

Estimated precision recall curves. We consider the correlation results to be the best single indicator of performance under our methodology. However, we can use our results to estimate other performance characterizations such

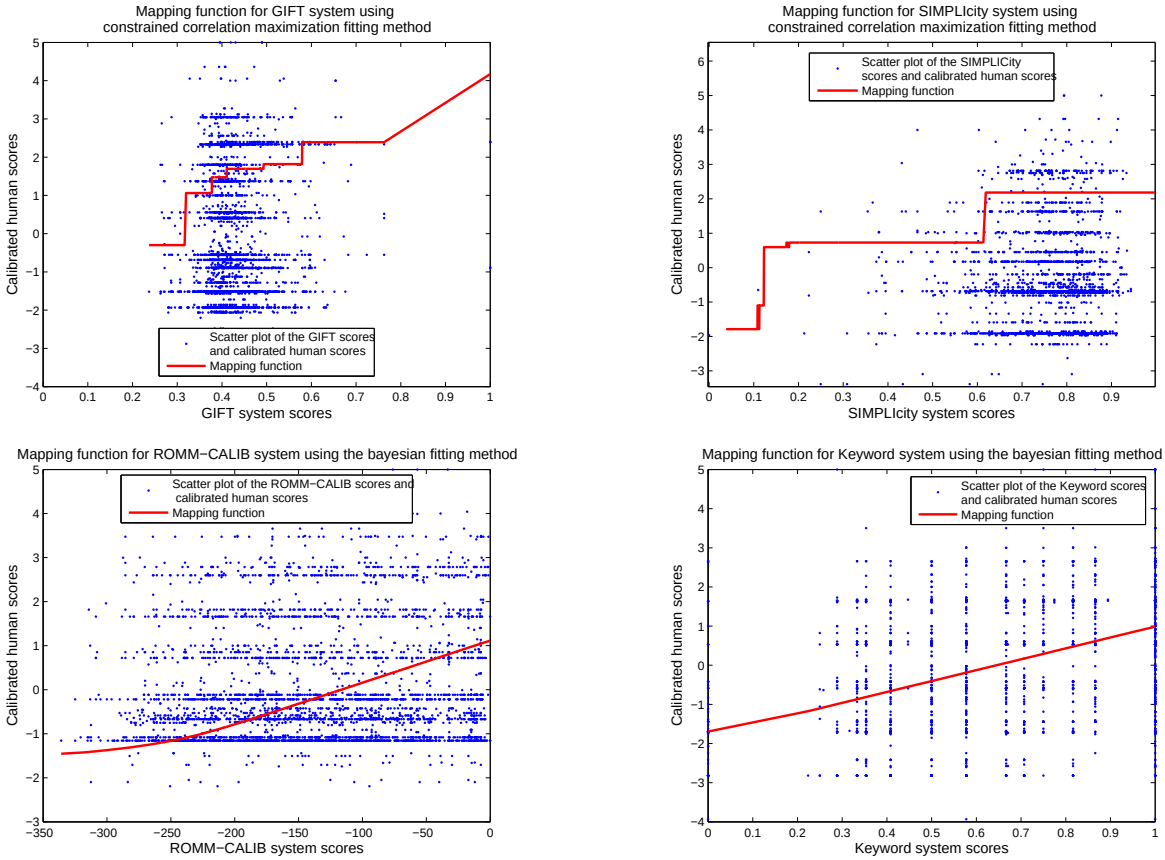

Figure 2. Scatter plots for the image retrieval scores and the adjusted human scores. We show the mapping function found using best curve fitting method for that system. Top left: GIFT; top right SIMPLIcity; bottom left ROMM-CALIB; and bottom right Keywords. The Keyword data appears sparse because the measure (1) admits only a limited number of values when the number of keywords per image is small (roughly 5 in our case). The values for ROMM-CALIB are the logarithms of non normalized probabilities.

Fitting methods	*Average correlation between human scores and mapped GIFT scores on data selected by different systems*					
	1	*2*	*3*	*4*	*Mean*	*All*
a	0.18	0.10	0.13	0.10	0.13(0.04)	0.10
b	0.13	0.16	0.26	0.23	**0.20**(0.03)	**0.17**
c	0.13	0.18	0.22	0.21	0.19(0.04)	0.10

Fitting methods	*Average correlation between human scores and mapped SIMPLIcity scores on data selected by different systems*					
	1	*2*	*3*	*4*	*Mean*	*All*
a	0.13	0.20	0.14	0.20	0.17(0.04)	0.18
b	0.19	0.23	0.24	0.31	**0.24**(0.05)	0.18
c	0.17	0.25	0.23	0.25	0.23(0.04)	**0.19**

Fitting methods	*Average correlation between human scores and mapped ROMM scores on data selected by different systems*					
	1	*2*	*3*	*4*	*Mean*	*All*
a	0.17	0.18	0.18	0.20	0.18(0.01)	0.21
b	0.22	0.26	0.29	0.37	0.29(0.06)	0.23
c	0.31	0.33	0.43	0.34	**0.35**(0.05)	**0.24**

Fitting methods	*Average correlation between human scores and mapped Keywords scores on data selected by by different systems*					
	1	*2*	*3*	*4*	*Mean*	*All*
a	0.17	0.28	0.51	0.41	0.34(0.14)	0.27
b	0.25	0.32	0.61	0.57	0.44(0.17)	0.38
c	0.53	0.58	0.62	0.56	**0.57**(0.04)	**0.51**

Table 3. The correlation between the mapped scores and the human evaluation scores. The tabulated values are the correlation measures for each of the four calibration systems, as computed based on the samples provided from each of the four systems, the average of those results, and based on all the data combined. The systems are: 1) GIFT; 2) SIMPLIcity; 3) ROMM-CALIB; and 4) Keywords. Results are provided for each of the three methods for fitting monotonic curves: (a) constrained least squares;(b) maximum constrained correlation; (c) Bayesian curve fitting. We used the best combined result for a given algorithm to map computer scores to human scores.

as precision recall curves using the usual definition:

$$\text{Precision} = \frac{\text{Number of relevant documents retrieved}}{\text{Total number of documents retrieved}} \quad (3)$$

$$\text{Recall} = \frac{\text{Number of relevant documents retrieved}}{\text{Number of relevant documents in database}} \quad (4)$$

Typically one plots the average values of precision versus recall over a threshold modulating the number of images returned. We emphasize that the form of our data is *different* than the form suggested by the formulas, and thus producing estimated PR curves requires care. We have a large number of query-result pairs which, by design, are a non-uniform sampling of the space of such pairs. Since we have many such pairs we can weight our averages to correct for the sampling. To compute the curves we essentially treat the top M CBIR responses as a single query for which we can compute the three quantities in the above two formulas. However, in order to estimate the ratios in the case of uniform sampling, which, in turn, estimates the ratios if we had all the data, we weight the computation of the quantities in (3) and (4) by the reciprocal of the sampling function. The estimated PR curves are in Figure 3.

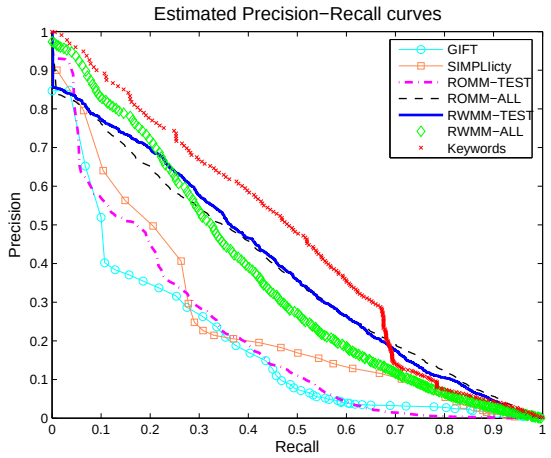

Figure 3. Precision recall curves for a number image retrieval methods. A relevant retrieved image corresponds to an adjusted human evaluation score greater than 3. Because the evaluation set is biased towards good matches, we have to estimate the PR curves by reversing the bias in rank. See text for details.

Our results show, not surprisingly, that keyword retrieval outshines the image based methods. This simply reflects the fact that semantics play a dominant role in what users consider a match, and that we are not very good at determining image semantics from features. The results also corroborate the notion that annotation oriented evaluation can serve as a proxy for grounded evaluation. However, the results also suggest that the scope of such a proxy is limited. Since the keyword results were far from perfect, a significant portion of what our participants expressed through their choices is not captured, and thus not measurable, using the keyword proxy.

Using words in training helps capture some relation between semantics and features, and the methods RWMM-ALL and RWMM-TEST do relatively well as a result. Without words, but still encoding the entire training set, the performance drops but is still respectable. We see this method (ROMM-ALL) to be a alternative method to SIMPLIcity in that it reports a match over several image regions. However, while ROMM-ALL models the statistics of the data set, SIMPLIcity computes the matches on the fly. We found that ROMM-ALL performs a bit better than SIMPLIcity. When forced to model images in general, but not in the training set, the the mixture model approach becomes a simple feature match method, and the performance results reflect this (worse than SIMPLIcity, same as GIFT).

5.5. Evaluating text queries

While our query-by-text experiment is less applicable to most CBIR research, grounding key word indexing is also important. For example, in the Corel image data, there are many images which have the keyword “sky”, but only a few of them would be of interest to a human user searching for a“sky” photograph. What that person is seeking is a particularly canonical or interesting example, and image information is one possible way to help them. Thus we hope that our approach and our data will lead to a better understanding of the limitations of keyword search, and suggest ways on how it can be improved.

We have done one experiment with the query-by-text data. We used a similar process as in the query-by-image-example case to map the results of the “Keywords” algorithm to the human evaluation scores. We then correlated the mapped scores with the human scores, arriving at a correlation of **0.50**. While this is suggestive of good retrieval, it also leaves much room for improvement, especially given that the data set is especially friendly.

6. Summary

We have developed a system for making grounded comparisons of retrieval systems. Importantly, the data and software applies to the evaluation of any system working in a similar domain and is available on-line (http://kobus.ca/research/data). The next step is to integrate our approach with a new data set under construction which is explicitly designed for image understanding and retrieval research and free of copyright issues. In doing so, we will expand the scale of data collection to include more participants, more data, and more data selection methods. We will also study in more detail participant subjectivity. And, of course, we will help others evaluate their retrieval systems.

7. Acknowledgments

We are grateful to Nicholas Heard for providing an implementation of his Bayesian curve fitting method and James Wang for providing an implementation of SIMPLIcity. We also give kudos to the authors of GIFT for providing an open source image retrieval system. Finally, we acknowledge the efforts of the participants.

References

[1] The benchathlon network, http://www.benchathlon.net.

[2] The gnu image-finding tool, www.gnu.org/software/gift.

[3] L. H. Armitage and P. G. B. Enser. Analysis of user need in image archives. *Journal of Information Science*, 23(4):287–299, 1997.

[4] K. Barnard, P. Duygulu, and D. Forsyth. Clustering art. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages II:434–441, 2001.

[5] K. Barnard, P. Duygulu, N. d. Freitas, D. Forsyth, D. Blei, and M. I. Jordan. Matching words and pictures. *Journal of Machine Learning Research*, 3:1107–1135, 2003.

[6] K. Barnard and D. Forsyth. Learning the semantics of words and pictures. In *International Conference on Computer Vision*, pages II:408–415, 2001.

[7] C. Carson, S. Belongie, H. Greenspan, and J. Malik. Blobworld: Color and texture-based image segmentation using em and its application to image querying and classification. *IEEE PAMI*, 24(8):1026–1038, 2002.

[8] 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(1):20–35, 2000.

[9] L. Dodd, R. Wagner, S. r. Armato, M. McNitt-Gray, S. Beiden, H. Chan, D. Gur, G. McLennan, C. Metz, N. Petrick, B. Sahiner, and J. Sayre. Assessment methodologies and statistical issues for computer-aided diagnosis of lung nodules in computed tomography: contemporary research topics relevant to the lung image database consortium. *Acad Radiol*, 11(4):462–75, 2004.

[10] P. Duygulu, K. Barnard, J. d. Freitas, and D. Forsyth. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *The Seventh European Conference on Computer Vision*, pages IV:97–112, 2002.

[11] B. Efron and R. J. Tibshirani. *An introduction to the bootstrap*. Chapman and Hall, New York, 1993.

[12] P. G. B. Enser. Query analysis in a visual information retrieval context. *Journal of Document and Text Management*, 1(1):25–39, 1993.

[13] P. G. B. Enser. Progress in documentation pictorial information retrieval. *Journal of Documentation*, 51(2):126–170, 1995.

[14] 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(9):22–32, 1995.

[15] D. A. Forsyth. Benchmarks for storage and retrieval in multimedia databases. In *Storage and Retrieval for Media Databases III*, volume 4676. SPIE, 2002.

[16] N. J. Gunther and G. B. Beratta. Benchmark for image retrieval using distributed systems over the internet: Birds-i. In G. B. Beretta and R. Schettini, editors, *Internet Imaging III*, volume 4311, pages 252–267. SPIE, 2001.

[17] N. A. Heard and A. F. M. Smith. Bayesian piecewise polynomial modeling of ogive and unimodal curves. Technical report, Imperial College London, 2002.

[18] C. C. Holmes and N. A. Heard. Generalized monotonic regression using random change points. *Statistics in Medicine*, 22(4):623–638, 2003.

[19] H. L. Kundel and M. Polansky. Comparing observer performance with mixture distribution analysis when there is no external gold standard. In *SPIE 3340*, pages 78–84, 1998.

[20] M. La Cascia, S. Sethi, and S. Sclaroff. Combining textual and visual cues for content-based image retrieval on the world wide web. In *IEEE Workshop on Content-Based Access of Image and Video Libraries*, 1998.

[21] W. Ma and B. Manjunath. Netra: A toolbox for navigating large image databases. *Multimedia Systems*, 7:84–198, 1999.

[22] M. Markkula and E. Sormunen. End-user searching challenges indexing practices in the digital newspaper photo archive. *Information retrieval*, 1:259–285, 2000.

[23] A. Pentland, R. W. Picard, and S. Sclaroff. Photobook: content-based manipulation of image databases. *International Journal of Computer Vision*, 18(3):233–254, 1996.

[24] T. Pfund and S. Marchand-Maillet. Dynamic multimedia annotation tool. In G. B. Beretta and R. Schettini, editors, *Internet Imaging III*, volume 4672, pages 206–224. SPIE, 2002.

[25] G. Salton. The state of retrieval system evaluation. *Information Processing and Management*, 28(4):441–450, 1992.

[26] S. Sclaroff, L. Taycher, and M. La Cascia. ImageRover: A content-based image browser for the world wide web. In *IEEE Workshop on content-based access of image and video libraries*, 1997.

[27] A. 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(12):1349–1379, 2000.

[28] J. R. Smith. Image retrieval evaluation. In *IEEE Workshop on content-based access of image and video libraries (CBVAILVL)*, 1998.

[29] J. Vogel and B. Schiele. On performance characterization and optimization for image retrieval. In *7th European Conference on Computer Vision*, volume IV, pages 49–63. Springer, 2002.

[30] J. Z. Wang and J. Li. Learning-based linguistic indexing of pictures with 2-D MHMMs. In *ACM Multimedia*, pages 436–445, 2002.

[31] J. Z. Wang, J. Li, and G. Wiederhold. SIMPLIcity: Semantics-sensitive integrated matching for picture libraries. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(9):947–963, 2001.