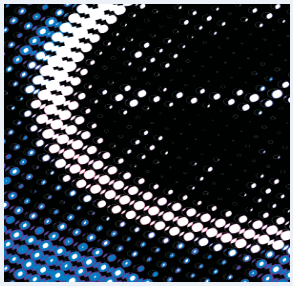
From Pixels to Semantic Spaces: Advances in Content-Based Image Retrieval

The paradigm for image retrieval has evolved from low-level image representations to semantic concept models to higher-level semantic inferences. UCSD's Statistical Visual Computing Laboratory has developed effective techniques for each paradigm that equate retrieval with classification and strive for minimum-probability-of-error optimality.



Nuno Vasconcelos
University of California, San Diego

In August 2006, Nielsen//NetRatings announced that five of the 10 fastest growing Web brands were user-generated content sites—platforms for photo or video sharing and blogs (www.nielsen-netratings.com/pr/PR_060810.PDF). Earlier statistics revealed that in April 2006 alone, the top five photo-sharing sites received close to 34 million unique US users (http://pic.photobucket.com/press/2006-06-PopPhoto.pdf).

These numbers illustrate a well-known corollary of the information revolution: the shift from passive users content with tuning in to rigidly formatted broadcast services to active users who demand ownership of the medium and become publishers. Technological advances in digital imaging, broadband networking, and data storage are motivating millions of ordinary people to communicate with one another and express themselves by sharing images, video, and other forms of media online.

However, certain capabilities are still lacking. In the context of image retrieval, acquiring, storing, and transmitting photos is now trivial, but it is significantly harder to manipulate, index, sort, filter, summarize, or search through them. Modern search engines and their image/video search offsprings have enabled significant progress in domains where visual content is tagged with text descriptions, but they only analyze metadata, not the images per se, and thus are of limited use in many practical scenarios.

For example, I can use one of the major image search engines to download 17,700 images of "kids playing soccer," most served from Internet sites across the world. Yet these are all useless to me when I am looking for pictures of *my* kids playing soccer. Although the latter are stored in my computer's hard drive, literally at hand's reach, they are inaccessible in any organized manner. I could, of course, manually label them, enabling my computer to perform more effective searches, but this feels wrong. After all, the machine should be working for me, not the other way around.

Published by the IEEE Computer Society
0018-9162/07/$25.00 © 2007 IEEE

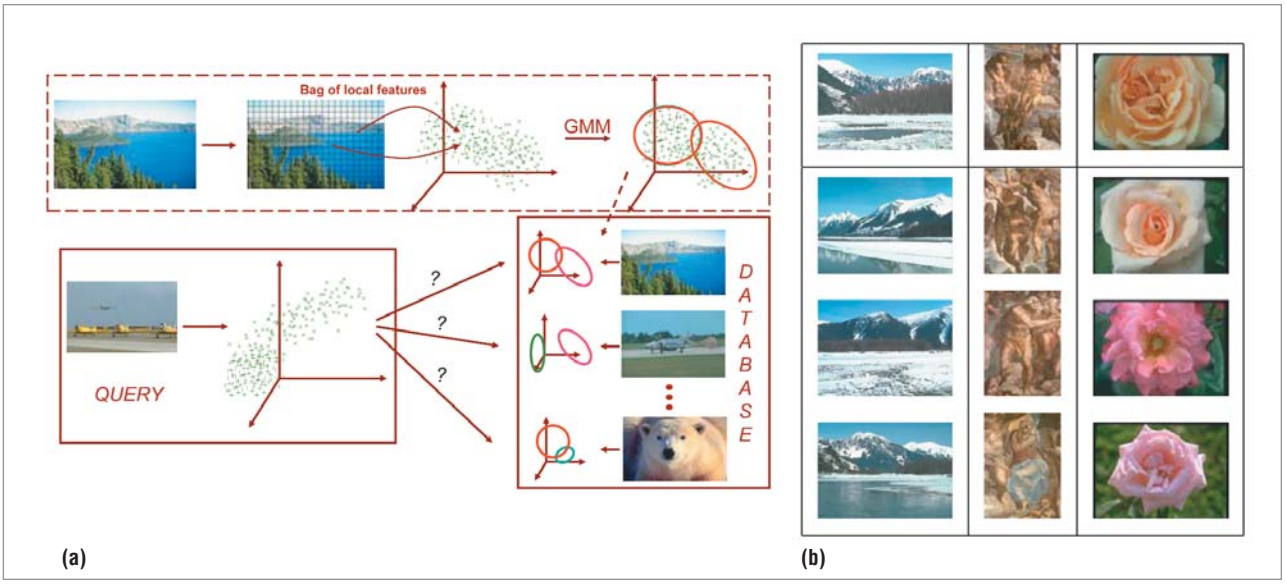


Figure 1. Minimum probability of error retrieval. (a) MPE retrieval architecture. The system decomposes images into bags of local features and characterizes them by their distributions on the feature space. Database images are ranked by posterior probability of having generated the query features. (b) Retrieval results. Each column shows the three best matches (among 1,500) to the query image shown at the top.

The Statistical Visual Computing Laboratory at the University of California, San Diego (www.svcl.ucsd.edu), has been considering the problem of content-based image retrieval for several years. One of SVCL's goals is to develop systems capable of retrieving images because they understand them and are thus able to represent their content in a form intuitive to humans. Drawing strongly on computer-vision and machine-learning research, this effort explores many issues in image representation and intelligent system design including the evaluation of image similarity, the automatic annotation of images with descriptive captions, the ability to understand user feedback during image search, and the design of indexing structures that can be searched efficiently.

QUERY BY VISUAL EXAMPLE

The classical paradigm for content-based image retrieval is *query by visual example*. QBVE retrieves images using strict visual matching, ranking database images by similarity to a user-provided query image. The system extracts a signature from the query, compares this signature to those previously computed for the images in the database, and returns the closest matches.

There are many ways to compose image signatures or evaluate their similarity.[1] While early solutions, including the pioneering *query by image content* system,[2] relied on very simple image-processing techniques, such as matching histograms of image colors, modern systems rely on more sophisticated representations and aim for provably optimal retrieval performance.

Minimum probability of error retrieval

SVCL's *minimum probability of error* retrieval system illustrates this evolution. In developing this system, we formulated the retrieval problem as one of classification and designed all system components to achieve optimality in the MPE sense.

As Figure 1a shows, the system decomposes images into bags of local features that measure properties such as texture, edginess, and color and then learns a Gaussian mixture model (GMM) from each bag. An image signature is thus a compact probabilistic representation of how the image populates the feature space. When faced with a query, the system extracts a bag of features from it and computes how well each GMM in the database explains this bag. In particular, the system ranks the database models according to their posterior probability of having generated the query features. This can be shown to be MPE optimal.[3]

In addition to finding the closest matches, the system assigns a match probability to all images in the database. This lets the system combine visual matches with other sources of information that might impact the relevance of each database image—for example, the text in an accompanying Web page, how well the image matches previous queries, and external events that could increase the relevance of certain images on certain days, such as high demand for football photos on Sunday night. By supporting probabilistic information fusion, the retrieval system is automatically compatible with most state-of-the-art techniques for intelligent system design.

The MPE retrieval system is currently among the top QBVE performers. Like most QBVE systems, it is most

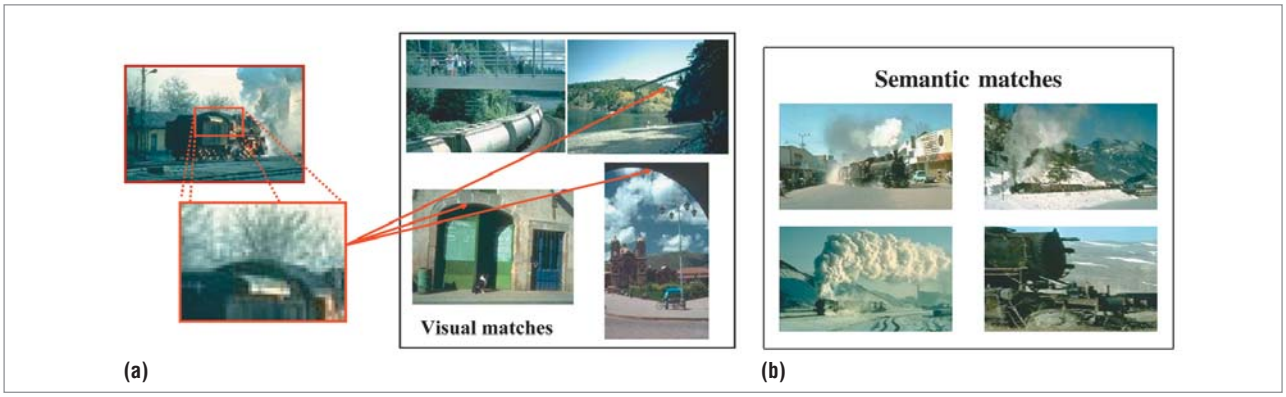


Figure 2. Closing the semantic gap. (a) People frequently discard strong visual cues in their similarity judgments, which can lead to severe query-by-visual-example errors such as retrieving bridges in response to a "train" query. (b) Because good matches require agreement along various dimensions of the semantic space, query by semantic example (QBSE) is significantly less prone to errors than QBVE.

accurate when similarity of visual appearance correlates with human judgments of similarity.[3] This is illustrated by Figure 1b, which presents the top matches from a database of 1,500 images to three queries. Note that the database is quite diverse, and the images are basically unconstrained in terms of lighting conditions, object poses, and so on (although all are good-quality images taken by professional photographers). The system can identify the different visual attributes that, in each case, contribute to the percept of image similarity. For example, similar color distributions seem to determine the matches of the first column, while texture appears to play a more significant role in the second column, and shape (of the flower petals) is probably the strongest cue for the third column's results.

Semantic gap

There are, nevertheless, many queries for which visual similarity does not correlate strongly with human similarity judgments. This can lead to a *semantic gap* between user and machine.

Figure 2a presents a subtle example of how people frequently discard strong visual cues in their similarity judgments. The "train" query contains a predominant arch-like structure that, from a strictly visual standpoint, makes the query highly compatible with concepts such as "bridge" or "arch." A QBVE system will return as top matches images like the four shown, three of which indeed contain bridges or archlike structures. However, people expect images of trains among the retrieved results and assign little probability to alternative interpretations, such as "bridge" or "arch." They seem to decide first that the image is about trains, and then use "train-ness" as the dimension that determines image similarity. Whether other trains are visually similar to what the query depicts—for example, in terms of colors, shape, or size—is relatively unimportant.

This mismatch between the similarity judgments can make user interaction with a QBVE system extremely frustrating. Most people would not be able to justify the matches returned in Figure 2a, despite the obvious similarities of the visual stimuli. This is the nightmare scenario for image retrieval, leaving users both unhappy with the retrieval results and convinced that the system "doesn't get it."

IMAGE ANNOTATION AND SEARCH

In recent years, the semantic gap between user and machine has motivated significant interest in semantic image retrieval. A semantic retrieval system aims for two complementary goals: image annotation and search.

Semantic labeling

The starting point for a semantic retrieval system is a training database of images, each annotated with a natural-language caption. From this database, the system learns to create a mapping between words and visual features. The system then uses this mapping to

- annotate unseen images with the captions that best describe them, and
- find the database images that best satisfy a natural-language query.

Usually, the training corpus is *weakly labeled*, meaning that

- the absence of a label from a caption does not necessarily mean that the associated visual concept is absent from the image, and
- it is not known which image regions are associated with each label.

For example, an image containing "sky" might not be explicitly annotated with that label and, when it is, no

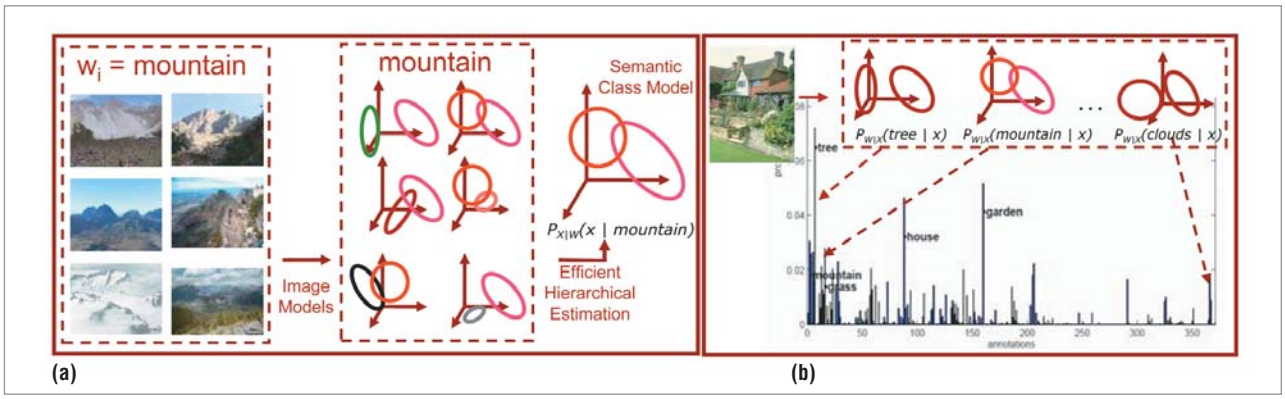


Figure 3. Semantic labeling. (a) An MPE semantic retrieval system groups images by semantic concept and learns a probabilistic model for each concept. (b) The system represents each image by a vector of posterior concept probabilities.

indication is available regarding which image pixels actually depict sky. A semantic retrieval system also does not require individual users to label training images. While this can certainly be supported to personalize the vocabulary, the default is to rely on generic vocabularies shared by many systems.

Under the MPE framework, a semantic retrieval system is a simple extension of a QBVE system. As Figure 3a shows, an MPE semantic retrieval system learns probabilistic models from image sets instead of single images. In particular, the system uses the set of training images labeled with a particular keyword—in this case, "mountain"—to learn the model for the associated visual concept. This procedure can be shown to converge to the true concept distribution, plus a background uniform component with small amplitude, if the set of training images is diverse.[4]

Given a set of models for different visual concepts, the system can optimally label any image in the MPE sense by computing how well each model explains its features. In particular, the system orders concepts by posterior probability, given the image, and annotates the image with the concepts of largest probability. Figure 3b shows how, among a vocabulary of more than 350 semantic concepts, an image of a country house receives as most likely the labels "tree," "garden," and "house."

Generalization

An MPE semantic retrieval system can learn semantic models very efficiently when individual image models are already available—that is, when the system also supports QBVE. In fact, an MPE semantic retrieval system's design has complexity equivalent to that of an MPE system that only supports QBVE. While simple, this semantic retrieval architecture currently achieves the best published results for both retrieval and annotation on a collection of standard retrieval benchmarks.[4,5]

Figure 4a shows some examples of MPE semantic retrieval. Note that the system recognizes concepts as diverse as "blooms," "mountain," "pool," "smoke," and "woman." In fact, it has learned that these classes can exhibit varying visual patterns—for example, that smoke can be white, black, or gray; that both blooms and humans can come in multiple colors, sizes (depending on image scale), and poses; or that "pool" can be about water, people (swimmers), or both. This type of generalization is impossible for QBVE systems, which model each image independently of the others.

The annotation results of Figure 4b illustrate a second form of generalization, based on contextual relationships, that humans also regularly exploit. For example, the fact that stores usually contain people makes us more prone to label an image of a store where no people are visible with the "people" keyword than an image depicting an animal in the wild. An MPE semantic retrieval system's errors likewise tend to involve improper contextual associations. Note, for example, that the system erroneously associates the concept "prop" with a jet fighter, the concept "leaf" with grass, the concepts "people" and "skyline" with a store display, and so on. Of course, in many situations such associations enable the system to correctly identify concepts that would otherwise be difficult to detect due to occlusion, poor imaging, and other factors.

The system's ability to make contextual generalizations stems from the weak labeling of its training corpus. Because the system learns concept models from unsegmented images, most positive examples of "shop" are also part of the positive set for "people," although the latter will include many non-shopping-related images as well. Thus, an image of a shop will elicit some response from the "people" model even if it does not contain people. This response will be weaker than that of an image of a shop that contains people but stronger than the response of the "shop" model to a picture of people in a nonshopping context, such as fishing at a lake.

These asymmetries are routine in human reasoning and thus appear natural to users, making an MPE semantic retrieval system's errors less annoying than those of its

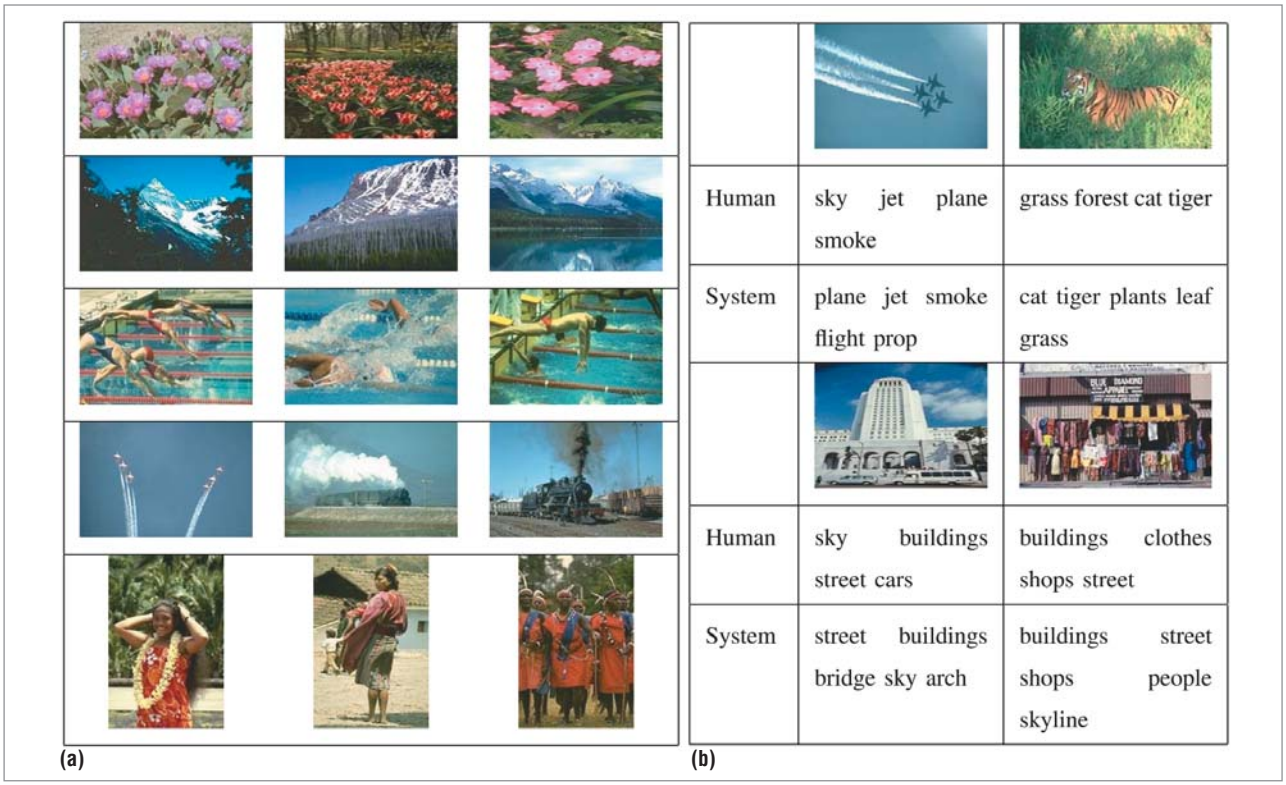


Figure 4. Generalization. (a) An MPE semantic retrieval system can recognize concepts with varying visual patterns. Each row shows the top three matches to a semantic query (from top to bottom): "blooms," "mountain," "pool," "smoke," and "woman." (b) Comparison of the annotations produced by the system with those of a human subject.

QBVE counterpart. In fact, informal surveys conducted in our lab indicate that users frequently miss the labeling errors and, even when they note an error, often find a reasonable explanation for it—for example, the system confused a jet for a propeller plane. This creates the sense in users that the system, regardless of its flaws, "gets it."

QUERY BY SEMANTIC EXAMPLE

Despite its many advantages, semantic retrieval also has limitations. An obvious difficulty is that most images have multiple semantic interpretations. Because training images are usually labeled with a short caption, some concepts might never be identified as present. This reduces the number of training examples and thereby impairs the system's ability to learn concepts that

- have a highly variable visual appearance, and
- are relatively rare.

Further, the semantic retrieval system's limited vocabulary can severely compromise generalization with respect to concepts outside the semantic space—that is, those on which the system is not trained. Although semantic retrieval generalizes better than QBVE inside the semantic space, this is usually not true outside it.

One possible solution to this problem is to adopt a *query by semantic example* paradigm.[6] The idea behind QBSE is to represent each image by its vector of posterior concept probabilities (shown in Figure 3b) and perform query by example in the simplex of these probabilities. Because the probability vectors are multinomial distributions over the space of semantic concepts, these can be referred to as *semantic multinomials*. A QBSE system defines a similarity function between these objects and, in response to a user-provided query image, ranks the images in the database by the distance of their semantic multinomials to that of the query.

Compared to semantic retrieval, QBSE is significantly less affected by multiple semantic interpretations and difficult generalization outside the semantic space. This follows from the fact that the system does not face a definitive natural-language query but rather an image that it expands into its internal semantic representation. For example, a system not trained with images of the concept "fishing" can still expand a query image of this subject into numerous alternative concepts such as "water," "boat," "people," and "nets" in its vocabulary. This is likely to produce high scores for other fishing images.

Further, it is much easier to generalize in the QBSE feature space. Figure 2b compares the matches produced by QBVE and QBSE to a common query. Inspection of the semantic multinomials associated with the images shown reveals that, although the query image receives fair probability for the concept "bridge," it receives only slightly

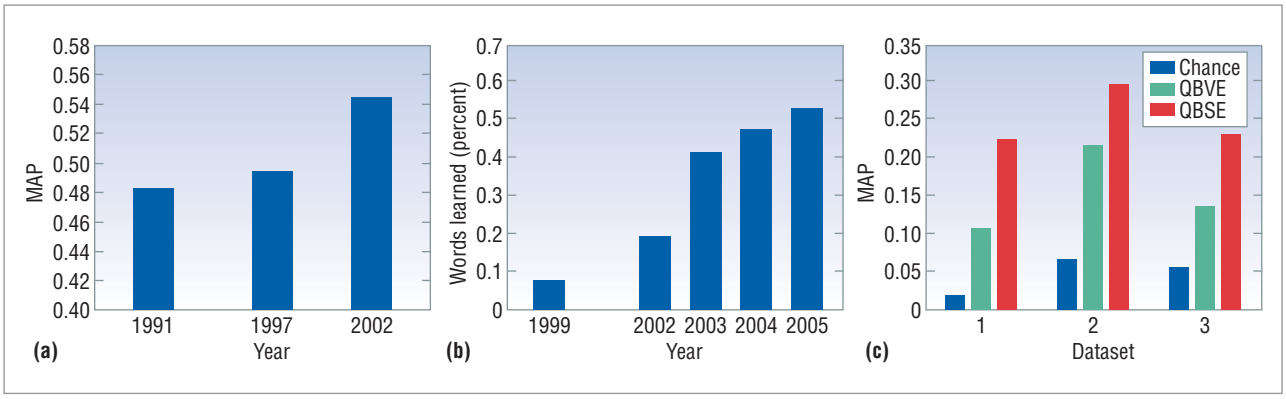


Figure 5. Historical evolution of retrieval and annotation performance. (a) Mean average precision of QBVE systems. (b) Percentage of words learned by semantic labeling systems. (c) Mean average precision of QBSE versus QBVE retrieval systems on datasets inside and outside the semantic space.

inferior probability for concepts such as "locomotive," "railroad," and "train." The latter are consistent with the semantic multinomials of other images depicting trains but not necessarily with those of images depicting bridges. Thus, although the erroneous "bridge" label is individually dominant, it loses this dominance when the semantic multinomials are matched as a whole.

PROGRESS IN IMAGE RETRIEVAL

QBVE, semantic retrieval, and QBSE, respectively, represent steps in an evolutionary process that proceeds from modeling visual appearance, to learning semantic models, to making inferences using semantic spaces. This evolution has occurred in the image retrieval community at large, and other researchers have developed alternative approaches. The following assessment of SVCL's techniques is thus not intended to demonstrate that they are the ultimate solution for the problems they address, but simply to quantify the progress associated with each evolutionary step.

All presented results are based on recall and precision. Given a query and the top n database matches, *recall* is the percentage of all relevant images the retrieved set contains, and *precision* is the percentage of n that are relevant or belong to the query's class.

Retrieval performance is measured by the *mean average precision*. MAP is defined as the average precision, over all queries, at the ranks where recall changes—that is, where relevant items occur. It is a number between 0 and 1, with higher values reflecting more accurate retrieval systems. Annotation performance is measured by the percentage of dictionary words the labeling system effectively learns. A word is effectively learned if it has recall greater than zero within the first five labels used to annotate each image.

QBVE

Figure 5a compares the MAP of three state-of-the-art QBVE methods on a set of 1,500 images from the Corel database of stock photography: the color histograms proposed by Michael Swain and Dana Ballard in 1991,[7] the color correlogram developed by Jing Huang and colleagues in 1997,[8] and SVCL's 2002 MPE system.[3] While retrieval performance has clearly increased during the past decade, the increase has not been dramatic—rising from about 48 to slightly more than 54 percent. Several reasons could account for this, including the fact that the databases used to test the early retrieval systems were suitable for the color-matching operations on which they were based. Nevertheless, the slow rate of progress indicates that we may be close to the asymptote of QBVE.

Semantic retrieval

A significantly greater rate of progress has occurred in the area of semantic retrieval. As Figure 5b shows, early annotation methods such as the co-occurrence technique proposed in 1999 by Yasuhide Mori, Hironobu Takahashi, and Ryuichi Oka[9] performed at about chance level. Notable subsequent improvements included the translation model of Pinar Duygulu and colleagues in 2002;[10] the continuous-space relevance model of Victor Lavrenko, Raghavan Manmatha, and Jiwoon Jeon the following year;[11] the multiple Bernoulli relevance model of Shaolei Feng, R. Manmatha, and V. Lavrenko in 2004;[12] and SVCL's MPE 2005 semantic retrieval method.[5]

These results were obtained on a challenging corpus of 5,000 images and a vocabulary of 371 words, with labels produced by nontechnical subjects. The recent surmounting of the barrier of 50 percent learned words is worth noting, and a far cry from the less than 10 percent rate of the earliest system.

QBSE

While annotation performance is not directly comparable to the MAP results of Figure 5a, it is possible to quantify the gains of semantic over visual representations by considering QBSE. In fact, when QBSE and QBVE are implemented within the MPE framework,

with identical visual features and classification architectures, any performance differences can be directly attributed to the semantic versus visual nature of the associated representations.

Figure 5c compares the MAP results of QBSE versus QBVE on three datasets: the 5,000-image corpus used to learn the semantic labels, a corpus of 1,800 images collected on the Flickr Web site, and an additional set of 1,500 images from Corel. All of these sets are more challenging than that of Figure 5a, leading to lower QBVE performance. The comparison tests performance both inside (dataset 1) and outside (datasets 2 and 3) the semantic space, but the results are qualitatively similar after correcting for chance success, also shown for each dataset. In all cases, QBSE performs substantially better than QBVE. This suggests that semantic-level retrieval is a potential area of future progress, certainly more promising than the classic QBVE paradigm.

Designing systems that understand images well enough to enable effective search of large databases remains a challenging problem, and current retrieval systems are not useful for all applications. The trend is very positive, however, and the retrieval community has only just begun to explore avenues of tremendous potential, such as the use of semantic taxonomies.

An image retrieval system is more than an image similarity engine. In addition to image matching, it should address the problems of indexing to enable fast searches; accounting for prior information, which can be used to weigh some images more strongly than others; and exploring the user's presence in the retrieval loop.

Information about the user's preferences is usually collected by relevance feedback algorithms, operating at both short and long time scales. Within a single session, the retrieval system can exploit user feedback to refine particular searches. As the user provides more information, the system becomes more confident about the user's needs, and retrieval accuracy increases. Across sessions, the system can use relevance feedback to build user profiles or improve semantic labeling of the database images.

All of these operations can be formulated under the MPE retrieval framework, and optimal solutions are available for many problems. ■

Acknowledgments

I thank Gustavo Carneiro, Antoni Chan, Andrew Lippman, Pedro Moreno, Nikhil Rasiawia, and Manuela Vasconcelos for their valuable contributions to the design of SVCL's retrieval systems. This work has been partially funded by NSF CAREER award IIS-0448609, NSF grant IIS-0534985, and a gift from Google.

References

1. A.W.M. Smeulders et al., "Content-Based Image Retrieval at the End of the Early Years," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 22, no. 12, 2000, pp. 1349-1380.
2. M. Flickner et al., "Query by Image and Video Content: The QBIC System," *Computer,* Sept. 1995, pp. 23-32.
3. N. Vasconcelos, "Minimum Probability of Error Image Retrieval," *IEEE Trans. Signal Processing,* vol. 52, no. 8, 2004, pp. 2322-2336.
4. G. Carneiro et al., "Supervised Learning of Semantic Classes for Image Annotation and Retrieval," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 29, no. 3, 2006, pp. 394-410.
5. G. Carneiro and N. Vasconcelos, "A Database-Centric View of Semantic Image Annotation and Retrieval," *Proc. 28th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval,* ACM Press, 2005, pp. 559-566.
6. N. Rasiwasia, N. Vasconcelos, and P.J. Moreno, "Query by Semantic Example," *Proc. 5th Int'l Conf. Image and Video Retrieval,* LNCS 4071, Springer, 2006, pp. 51-60.
7. M.J. Swain and D.H. Ballard, "Color Indexing," *Int'l J. Computer Vision,* vol. 7, no. 1, 1991, pp. 11-32.
8. J. Huang et al., "Image Indexing Using Color Correlograms," *Proc. 1997 Conf. Computer Vision and Pattern Recognition,* IEEE CS Press, 1997, pp. 762-768.
9. Y. Mori, H. Takahashi, and R. Oka, "Image-to-Word Transformation Based on Dividing and Vector Quantizing Images with Words," *Proc. 1st Int'l Workshop on Multimedia Intelligent Storage and Retrieval Management* (in conjunction with the 7th ACM Int'l Conf. Multimedia), ACM Press, 1999.
10. P. Duygulu et al., "Object Recognition as Machine Translation: Learning a Lexicon for a Fixed Image Vocabulary," *Proc. 7th European Conf. Computer Vision,* vol. 4, LNCS 2353, Springer, 2002, pp. 349-354.
11. V. Lavrenko, R. Manmatha, and J. Jeon, "A Model for Learning the Semantics of Pictures," *Proc. 17th Ann. Conf. Neural Information Processing Systems,* NIPS Foundation, 2003; http://books.nips.cc/papers/files/nips16/NIPS2003_AA70.pdf.
12. S.L. Feng, R. Manmatha, and V. Lavrenko, "Multiple Bernoulli Relevance Models for Image and Video Annotation," *Proc. 2004 Conf. Computer Vision and Pattern Recognition,* IEEE CS Press, 2004, pp. 1002-1009.

Nuno Vasconcelos *is an assistant professor in the Department of Electrical and Computer Engineering at the University of California, San Diego, where he also heads the Statistical Visual Computing Laboratory. His research interests include computer vision, statistical signal processing, machine learning, and multimedia. Vasconcelos received a PhD in media arts and sciences from the Massachusetts Institute of Technology. He is a member of the IEEE Computer Society. Contact him at nuno@ece.ucsd.edu.*