Minimum Probability of Error Image Retrieval: From Visual Features to Image Semantics

Minimum Probability of Error Image Retrieval: From Visual Features to Image Semantics

Nuno Vasconcelos

University of California, San Diego
La Jolla, CA 92093
USA
nuno@ece.ucsd.edu

Manuela Vasconcelos

University of California, San Diego
La Jolla, CA 92093
USA
maspcv@gmail.com

the essence of knowledge

Boston – Delft

Foundations and Trends® in Signal Processing

Published, sold and distributed by:
now Publishers Inc.
PO Box 1024
Hanover, MA 02339
USA
Tel. +1-781-985-4510
www.nowpublishers.com
sales@nowpublishers.com

Outside North America:
now Publishers Inc.
PO Box 179
2600 AD Delft
The Netherlands
Tel. +31-6-51115274

The preferred citation for this publication is N. Vasconcelos and M. Vasconcelos, Minimum Probability of Error Image Retrieval: From Visual Features to Image Semantics, Foundations and Trends® in Signal Processing, vol 5, no 4, pp 265–389, 2011

ISBN: 978-1-60198-608-5

Foundations and Trends® in Signal Processing
Volume 5 Issue 4, 2011

Editorial Scope

Foundations and Trends® in Signal Processing will publish survey and tutorial articles on the foundations, algorithms, methods, and applications of signal processing including the following topics:

- Adaptive signal processing
- Audio signal processing
- Biological and biomedical signal processing
- Complexity in signal processing
- Digital and multirate signal processing
- Distributed and network signal processing
- Image and video processing
- Linear and nonlinear filtering
- Multidimensional signal processing
- Multimodal signal processing
- Multiresolution signal processing
- Nonlinear signal processing
- Randomized algorithms in signal processing
- Sensor and multiple source signal processing, source separation
- Signal decompositions, subband and transform methods, sparse representations
- Signal processing for communications
- Signal processing for security and forensic analysis, biometric signal processing
- Signal quantization, sampling, analog-to-digital conversion, coding and compression
- Signal reconstruction, digital-to-analog conversion, enhancement, decoding and inverse problems
- Speech/audio/image/video compression
- Speech and spoken language processing
- Statistical/machine learning
- Statistical signal processing
  - classification and detection
  - estimation and regression
  - tree-structured methods

Information for Librarians

Foundations and Trends® in Signal Processing, 2011, Volume 5, 4 issues. ISSN paper version 1932-8346. ISSN online version 1932-8354. Also available as a combined paper and online subscription.

Foundations and Trends® in
Signal Processing
Vol. 5, No. 4 (2011) 265–389

DOI: 10.1561/2000000015

Minimum Probability of Error Image Retrieval: From Visual Features to Image Semantics

Nuno Vasconcelos[1] and Manuela Vasconcelos[2]

[1] *University of California, San Diego, 9500 Gilman Drive, MC 0407, La Jolla, CA 92093, USA, nuno@ece.ucsd.edu*
[2] *University of California, San Diego, 9500 Gilman Drive, MC 0407, La Jolla, CA 92093, USA, maspcv@gmail.com*

Abstract

The recent availability of massive amounts of imagery, both at home and on the Internet, has generated substantial interest in systems for automated image search and retrieval. In this work, we review a principle for the design of such systems, which formulates the retrieval problem as one of decision-theory. Under this principle, a retrieval system searches the images that are likely to satisfy the query with *minimum probability of error* (MPE). It is shown how the MPE principle can be used to design optimal solutions for practical retrieval problems. This involves a characterization of the fundamental performance bounds of the MPE retrieval architecture, and the use of these bounds to derive optimal components for retrieval systems. These components include a feature space where images are represented, density estimation methods to produce this representation, and the similarity function to be used for image matching. It is also

shown that many alternative formulations of the retrieval problem are closely related to the MPE principle, typically resulting from simplifications or approximations to the MPE architecture. The MPE principle is then applied to the design of retrieval systems that work at different levels of abstraction. *Query-by-visual-example* (QBVE) systems are strictly visual, matching images by similarity of low-level features, such as texture or color. This is usually insufficient to produce perceptually satisfying results, since human users tend to make similarity judgments on the basis of image semantics, not visual attributes. This problem is addressed by the introduction of MPE labeling techniques, which associate descriptive keywords with images, enabling their search with text queries. This involves computing the probabilities with which different concepts explain each image. The query by example paradigm is then combined with these probabilities, by performing MPE image matching in the associated probability simplex. This is denoted *query-by-semantic-example* (QBSE), and enables example-based retrieval by similarity of semantics.

Contents

1

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

We are currently living through a confluence of three technological revolutions – the advent of digital imaging, broadband networking, and inexpensive storage – that allow millions of people to communicate and express themselves by sharing media. It could be argued, however, that a few pieces are still missing. While it is now trivial to acquire, store, and transmit images, it is significantly harder to manipulate, index, sort, filter, summarize, or search through them. Significant progress has, without doubt, happened in domains where the visual content is tagged with text descriptions, due to the advent of modern search engines and their image/video search off-springs. Nevertheless, because they only analyze *metadata*, not the images per se, these are of limited use in many practical scenarios. For example the reader can, at this moment, use one of the major image search engines to download $7,860,000$ pictures of "kids playing soccer", most served from Internet sites across the world. Yet, these are all useless, to *the reader*, when he/she is looking for pictures of *his/her* kids playing soccer. Although the latter are stored in the reader's hard-drive, literally at "hand's reach", they are completely inaccessible in any organized manner. The reader could, of course, take the time to manually label them, enabling the computer to perform more effective searches, but this somehow feels wrong. After all, the machine should be working for the user, not the other way around.

The field of *content-based image search* aims to develop systems capable of retrieving images because they understand them and are able to represent their content in a form that is intuitive to humans. It draws strongly on computer vision and machine learning, and encompasses many sub-problems in image representation and intelligent system design. These include the evaluation of image similarity, the automatic annotation of images with descriptive captions, the ability to understand user feedback during image search, and support for indexing structures that can be searched efficiently. In this monograph, we review the progress accomplished in this field with a formulation of the problem as one of decision theory. We note that the decision theoretic view is not the only possible solution to the retrieval problem and that many alternatives have been proposed in the literature. These alternatives are covered by recent extensive literature reviews [24, 68, 105, 115] and will not be discussed in what follows, other than in context of highlighting possible similarities or differences to MPE retrieval.

1.1 Query by Visual Example

Query by visual example (QBVE) is the classical paradigm for content-based image search. It is based on strict visual matching, ranking database images by similarity to a user-provided query image. The steps are as follows: user provides query, retrieval system extracts a signature from it, this signature is compared to those previously computed for the images in the database, and the closest matches are returned to the user. There are, of course, many possibilities for composing image signatures or evaluating their similarity, and a rich literature has evolved on this topic [105]. While early solutions, such as the pioneering *query-by-image-content* system [80], were based on very simple image processing (e.g., matching of histograms of image colors), modern systems (1) rely on more sophisticated representations, and (2) aim for provably optimal retrieval performance.

In what follows, we review one such approach, usually denoted as *minimum probability of error* (MPE) retrieval. The retrieval problem is

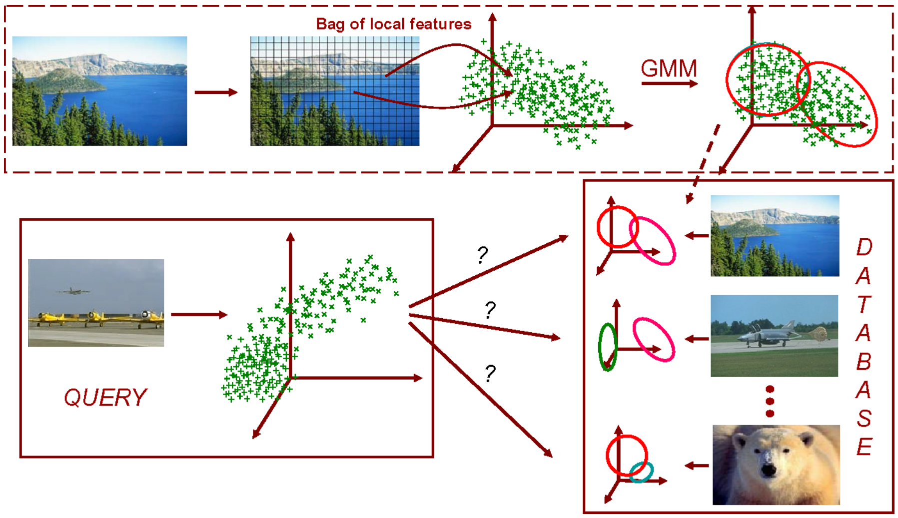

Fig. 1.1 MPE retrieval architecture. Images are decomposed into bags of local features, and characterized by their distributions on feature space. Database images are ranked by posterior probability of having generated the query features.

formulated as one of classification, and all components of the retrieval system are designed to achieve optimality in the MPE sense. This leads to the retrieval architecture depicted in Figure 1.1. Images are first represented as bags of local features (that measure properties such as texture, edginess, color, etc.), and a probabilistic model (in the figure a Gaussian mixture) is learned from the bag extracted from each image. The image signature is, therefore, a compact probabilistic representation of how it populates the feature space. When faced with a query, the retrieval system extracts a bag of features from it, and computes how well this bag is explained by each of the probabilistic models in the database. In particular, it ranks the database models according to their posterior probability, given the query. As we will see later on, this is optimal in the MPE sense.

Note that, besides finding the closest matches, the system assigns a probability of match to all images in the database. This allows the combination of visual matching with other sources of information that may impact the relevance of each database image. For example, the text in an accompanying web page [92], how well the image matches previous

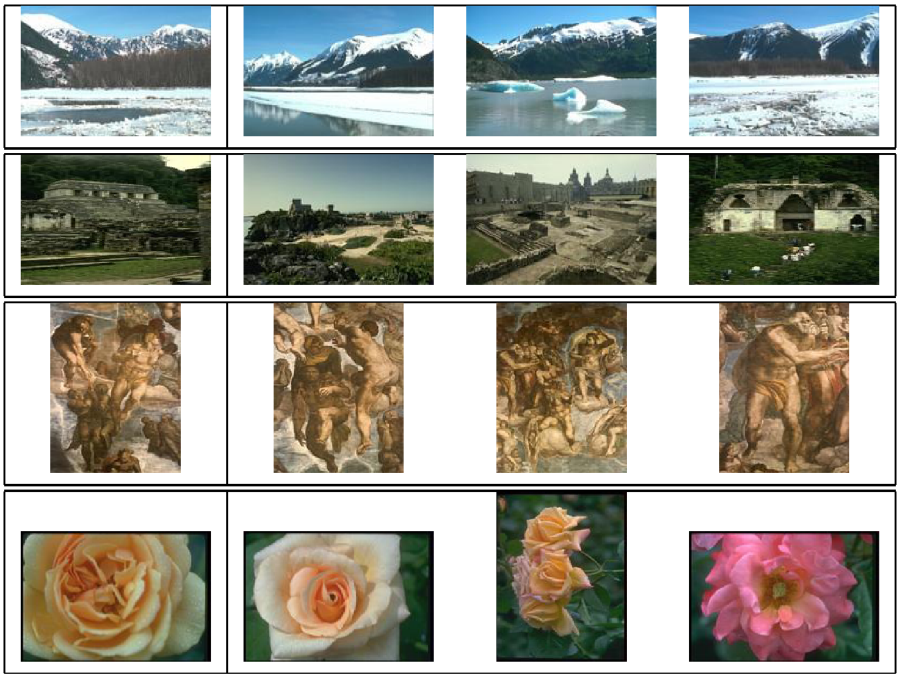

Fig. 1.2 MPE retrieval results. Each row shows the top three matches (among 1,500) to the query on the left.

queries [127, 128], external events that could increase the relevance of certain images on certain days (e.g., high demand for football images on Sunday night), etc.

The retrieval architecture of Figure 1.1 is currently among the top performers in QBVE [124]. These systems work well when similarity of visual appearance correlates with human judgments of similarity. This is illustrated by Figure 1.2, which presents the top matches, from a database of 1500 images, to four queries. Note that the database is quite diverse, and the images are basically unconstrained in terms of lighting conditions, object poses, etc. (even though they are all good quality images taken by professional photographers). The system is able to identify the different visual attributes that, in each case, contribute to the perception of image similarity. For example, similar color distributions seem to be determinant in the matches of the first row, while texture appears to play a more significant role in the third, shape (of the

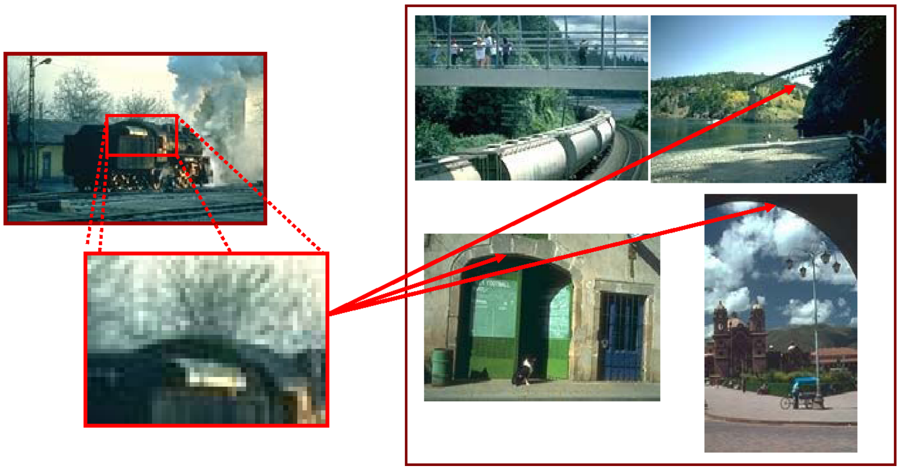

Fig. 1.3 A query image (left) and its top four matches by a QBVE system (right). Humans frequently discard strong visual cues in their similarity judgments. Failure to do this can lead to severe QBVE errors. For example, the visually distinctive arch-like structure in the train query induces the QBVE system to retrieve images of bridges or other arch-like structures.

flower petals) is probably the strongest cue for the results of the fourth, and the matches of the second row are likely due to the commonality of edge patterns in the building structures present in all images.

There are, nevertheless, many queries for which visual similarity does not correlate strongly with human similarity judgments. Figure 1.3 presents an example of how people frequently discard very strong visual cues in their similarity judgments. As can be seen from the close-up, the "train" query contains a very predominant arch-like structure. From a strictly visual standpoint, this makes it very compatible with concepts such as "bridges" or "arches". A QBVE system will fall in this trap, returning as top matches the four images also shown. Note that three of these do contain bridges or arch-like structures. Yet, the "train" interpretation of the query is completely dominant for humans, which assign very little probability to the alternative interpretations, and expect images of trains among the retrieved results.

The mismatch between the similarity judgments of user and machine can make the retrieval operation very unsatisfying. In the "train" example, most people would not be to able justify the matches returned by the retrieval system, despite the obvious similarities of the visual stimuli. This is the nightmare scenario for image retrieval, since users not only end up unhappy with the retrieval results, but also acquire the feeling that the system just "does not get it". This can be an enormous source of user frustration.

1.2 Semantic Retrieval

The discussion above reveals what is often called a *semantic gap* between user and machine. Unlike QBVE systems, people seem to first classify images as belonging to a number of semantic classes, and then make judgments of similarity in the higher level semantic space where those classes are defined. This has motivated significant interest, over the last decade, in semantic image retrieval. A semantic retrieval system aims for the two complementary goals of image *annotation* and *search*. The starting point is a training image database, where each image is annotated with a natural language caption, from which the retrieval system learns a *mapping between words and visual features*. This mapping is then used to (1) annotate unseen images with the captions that best describe them, and (2) find the database images that best satisfy a natural language query.

Usually, the training corpus is only *weakly labeled*, in the sense that (1) the absence of a label from a caption does not necessarily mean that the associated visual concept is absent from the image, and (2) it is not known which image regions are associated with each label. For example, an image containing "sky" may not be explicitly annotated with that label and, when it is, no indication is available regarding which image pixels actually depict sky. Note that the implementation of a semantic retrieval system 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 retrieval framework, a semantic retrieval system is a simple extension of a QBVE system. As shown in Figure 1.4, it can be implemented by learning probabilistic models from *image sets*, instead of single images. In particular, the set of training images labeled with

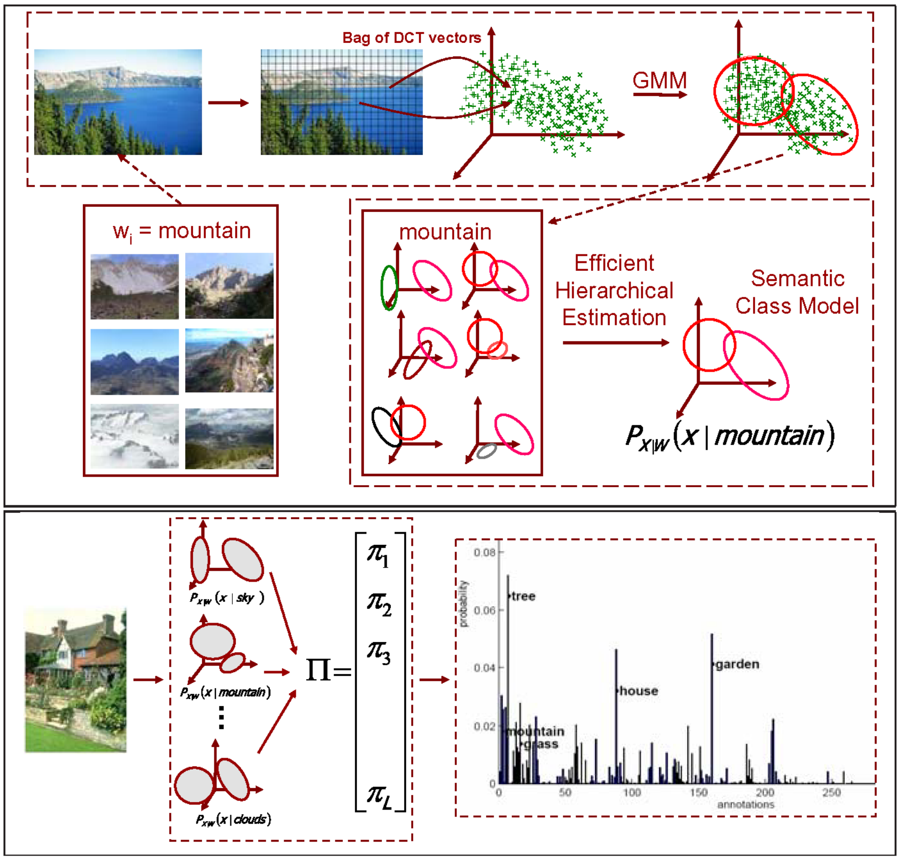

Fig. 1.4 Semantic MPE labeling. Top: images are grouped by semantic concept, and a probabilistic model learned for each concept. Bottom: each image is represented by a vector of posterior concept probabilities.

a particular keyword ("mountain", in the figure) is used to learn the model for the associated visual concept. As discussed in Section 6, this procedure converges to the true concept distribution plus a background uniform component that has small amplitude, if the set of training images is very diverse [16]. Given a set of models for different visual concepts, any image can be optimally labeled, in the MPE sense, by computing how well its features are explained by each model. In particular, the concepts are ordered by posterior probability, given the image, and the image is annotated with those of largest probability.

This is shown in Figure 1.4 where, 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”.

It turns out that, under the MPE framework, it is possible to learn semantic models very efficiently, when individual image models are already available, i.e., when QBVE is also supported. In fact, it can be shown that the design of a semantic MPE retrieval system has complexity equivalent to that of an MPE system that only supports QBVE [16, 17]. Some examples of retrieval and annotation are shown in Figures 1.5 and 1.6. Note that the system recognizes concepts as diverse as “blooms”, “mountains”, “swimming pools”, “smoke”, or “woman”. In fact, the system has learned that these classes can exhibit a wide diversity of patterns of visual appearance, e.g., that smoke can be both

Fig. 1.5 Semantic retrieval results. Each row shows the top four matches to a semantic query. From first to fifth row: *‘blooms’*, *‘mountain’*, *‘pool’*, *‘smoke’*, and *‘woman’*.

			
Human Annotation	sky jet plane smoke	snow fox arctic	sky buildings street cars
Automated Annotation	plane jet smoke flight prop	arctic snow polar fox ice	street buildings bridge sky arch
			
Human Annotation	grass forest cat tiger	bear polar snow tundra	coral fish ocean reefs
Automated Annotation	cat tiger plants leaf grass	polar tundra bear snow ice	reefs coral ocean fan fish
			
Human Annotation	water bridge train railroad	buildings clothes shops street	mountain sky clouds tree
Automated Annotation	sky bridge locomotive water train	buildings street shops people skyline	mountain valley sky clouds tree

Fig. 1.6 Comparison of the annotations produced by the system with those of a human subject.

white or very dark, that both blooms and humans can come in multiple colors, multiple sizes (depending on image scale), and multiple poses, or that pools can be mostly about water, mostly about people (swimmers), or both. This type of *generalization* is impossible for QBVE systems, where each image is modeled independently of the others.

The annotation results of Figure 1.6 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 that depicts an animal in the wild. This is also the case for the MPE semantic retrieval system, whose errors tend to be (in significant part) due to this type of 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 forth. Of course, there are also many situations in which these associations are highly beneficial and allow the correct identification of concepts that would otherwise be difficult to detect (due to occlusion, poor imaging conditions, etc.).

The ability to make such contextual generalizations stems from the weakly supervised nature of the training of the labeling system. Because concept models are learned from unsegmented images, most positive examples of "shop" are also part of the positive set for "people" (even though the latter will include many non-shopping related images as well). Hence, an image of a shop will originate some response from the "people" model, even when it does not contain people. That 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 on a non-shopping context, e.g., fishing in a lake. These asymmetries are routine in human reasoning and, therefore, appear natural to users, making the errors of a semantic retrieval system less annoying than those of its QBVE counterpart. In fact, informal surveys conducted in our lab have shown that (1) humans frequently miss the labeling errors, and (2) even when the error is noted, the user can frequently find an explanation for it (e.g., "it confused a jet for a propeller plane"). This creates the sense that, even in making errors, the semantic retrieval system "gets it".

1.3 Exploring Semantic Feature Spaces

Despite all its advantages, semantic retrieval is not free of limitations. An obvious difficulty is that most images have multiple semantic interpretations. Since training images are usually labeled with a short caption, some concepts may never be identified as present. This reduces the number of training examples and can impair the learning of concepts that (1) have high variability of visual appearance, or (2) are relatively rare. Furthermore, the semantic retrieval system is limited by the size of its vocabulary. Since it is still difficult to learn massive vocabularies, this can severely compromise generalization. It is, in fact, important to distinguish two types of generalization. The first is with respect to the concepts on which the system is trained, or *within the semantic space*. The second is with respect to all other concepts, or *outside the semantic space*.

While, as discussed in the previous section, semantic retrieval generalizes better (than QBVE) inside the semantic space, this is usually not true outside of it. One possibility, to address this problem, is to return to the query-by-example paradigm, but now at the semantic level, i.e., to adopt *query by semantic example* (QBSE) [91]. The idea is to represent each image by its vector of posterior concept probabilities (the π vector of Figure 1.4), and perform query by example in the simplex of these probabilities. Because the probability vectors are multinomial distributions over the space of semantic concepts, we refer to them as *semantic multinomials*. A similarity function between these objects is defined, the user provides a query image, and the images in the database are ranked by the distance of their semantic multinomials to that of the query. The process is illustrated in Figure 1.7.

When compared to semantic retrieval, a QBSE system is significantly less affected by the problems of (1) multiple semantic interpretations, and (2) difficult generalization outside of the semantic space. This follows from the fact that the system is not faced with a definitive natural language query, but 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 a number of alternative concepts, such as "water", "boat", "people", and "nets", in its vocabulary. This is likely to produce high scores for other images of fishing.

When compared to QBVE, QBSE has the advantage of a feature space where it is much easier to generalize. This is illustrated by Figure 1.8, which shows the QBSE matches to the query image of Figure 1.3. Note how these correlate much better with human

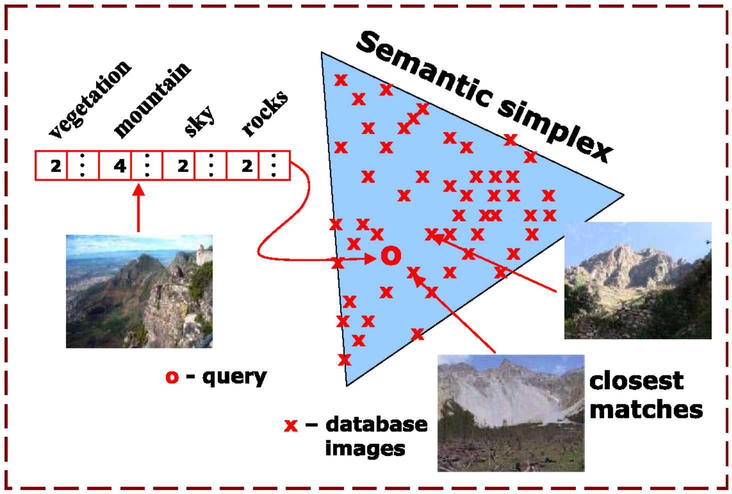

Fig. 1.7 Query by semantic example. Images are represented as vectors of concept probabilities, i.e., points on the semantic probability simplex. The vector computed from a query image is compared to those extracted from the images in the database, using a suitable similarity function. The closest matches are returned by the retrieval system.

Fig. 1.8 Top four matches to the QBSE query derived from the image shown on the left. Because good matches require agreement along various dimensions of the semantic space, QBSE is significantly less prone to the errors made by QBVE. This can be seen by comparing this set of image matches to those of Figure 1.3.

judgments of similarity that the QBVE matches of that figure. Inspection of the semantic multinomials associated with all images shown reveals that, although the query image receives a fair amount of probability for the concept “bridge”, it receives only slightly inferior amounts of 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. Hence, while the erroneous “bridge” label is individually dominant, it looses this dominance when the semantic multinomials are matched as a whole.

1.4 Organization of the Manuscript and Acknowledgments

In the following sections, we study in greater detail the fundamental properties of MPE retrieval. We start by laying out its theoretical foundations in Section 2. The sources of error of a retrieval system are identified, and upper and lower bounds on the resulting probability of error are derived. In Section 3, MPE retrieval architectures are related to a number of other approaches in literature. It is shown that many of the latter are special cases of the former, under simplifying assumptions that are not always sensible. In Section 4, we start to address the practical design of retrieval systems, by proposing a particular MPE implementation. This architecture is shown to have a number of interesting properties, and perform well in QBVE retrieval experiments. In Section 5, we consider the problem of semantic retrieval, by introducing MPE techniques for image annotation, and showing how they can be used to retrieve images with keyword-based queries. Some core technical issues in automated image annotation are then discussed in Section 6, where we study the possibility of learning image labels from weakly annotated training sets. The issue of generalization beyond the semantic space is introduced in Section 7, where we discuss QBSE. Finally, some conclusions are drawn in Section 8.

At this point, we would like to acknowledge the contributions of a number of colleagues that, over the last 10 years, have helped shape the research effort from which this work has resulted. Gustavo Carneiro has played an instrumental role in the development of the early ideas, from the design of multiple feature representations, to the first generation of our image annotation system. This work was then pursued by Antoni Chan, in a collaboration that also involved Pedro Moreno at Google. This allowed us to evaluate the experimental performance of the theoretical ideas, at a scale that would not be possible in an academic laboratory. Nikhil Rasiwasia then took over, and developed most of the QBSE framework, as well as a number of more recent contributions that are not discussed here, mostly for lack of space. Since this manuscript follows closely a number of papers that we have co-written with all these colleagues, we will not include a more extensive discussion of who-did-what here. If interested, please refer to [16, 91, 119, 120, 124, 125]. Instead, we would like to thank a number of other people who were instrumental in the development of many of the ideas discussed here, including Andrew Lippman at MIT, and several students at the Statistical Visual Computing Laboratory at UCSD. These include Dashan Gao, Hamed Masnadi-Shirazi, Sunhyoung Han, and Vijay Mahadevan, among others. The many discussions that we have had over the years, about retrieval and related topics, have made our ideas much more clear and effective.

References

[1] A.Kalai and A.Blum., "A note on learning from multiple instance examples," *Artificial Intelligence*, vol. 30, pp. 23–30, 1998.

[2] M. Artin, *Algebra*. Prentice Hall, 1991.

[3] P. Auer, "On learning from multi-instance examples: Empirical evaluation of a theoretical approach," in *Proceedings of the International Conference on Machine Learning*, 1997.

[4] J. Bach, "The virage image search engine: An open framework for image management," in *SPIE Storage and Retrieval for Image and Video Databases*, San Jose, California, 1996.

[5] K. Barnard and D. Forsyth, "Learning the semantics of words and pictures," in *International Conference on Computer Vision,* vol. 2, pp. 408–415, Vancouver, 2001.

[6] A. Bell and T. Sejnowski, "The independent components of natural scenes are edge filters," *Vision Research*, vol. 37, no. 23, pp. 3327–3328, December 1997.

[7] S. Belongie, C. Carson, H. Greenspan, and J. Malik, "Color-and texture-based image segmentation using EM and its application to content-based image retrieval," in *International Conference on Computer Vision*, pp. 675–682, Bombay, India, 1998.

[8] J. Bergen and E. Adelson, "Early vision and texture perception," *Nature*, vol. 333, no. 6171, pp. 363–364, 1988.

[9] J. Bergen and M. Landy, "Computational modeling of visual texture segregation," in *Computational Models of Visual Processing*, (M. Landy and J. Movshon, eds.), MIT Press, 1991.

[10] D. Bertsekas, *Nonlinear Programming*. Athena Scientific, 1995.

[11] C. M. Bishop, *Pattern Recognition and Machine Learning,* vol. 4. New York: Springer, 2006.

[12] R. Blahut, *Principles and Practice of Information Theory.* Addison Wesley, 1991.

[13] D. Blei and M. Jordan, "Modeling Annotated Data," in *Proceedings of ACM SIGIR Conference on Research and Development in Information Retrieval*, 2003.

[14] J. Boreczky and L. Rowe, "Comparison of video shot boundary detection techniques," in *Proceedings of SPIE Conference on Visual Communication and Image Processing*, 1996.

[15] J. Cardoso, "Blind signal separation: Statistical principles," *Proceedings of the IEEE*, vol. 90, no. 8, pp. 2009–2026, October 1998.

[16] G. Carneiro, A. Chan, P. Moreno, and N. Vasconcelos, "Supervised learning of semantic classes for image annotation and retrieval," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 3, pp. 394–410, March 2007.

[17] G. Carneiro and N. Vasconcelos, "A database centric view of semantic image annotation and retrieval," in *Proceedings of ACM SIGIR*, 2005.

[18] S. Chang, W. Chen, H. Meng, H. Sundaram, and D. Zhong, "A fully automated content-based video search engine supporting spatiotemporal queries," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 8, no. 5, pp. 602–615, September 1998.

[19] R. Clarke, *Transform Coding of Images.* Academic Press, 1985.

[20] D. Comaniciu, P. Meer, K. Xu, and D. Tyler, "Retrieval performance improvement through low rank corrections," in *Workshop in Content-based Access to Image and Video Libraries*, pp. 50–54, Fort Collins, Colorado, 1999.

[21] P. Comon, "Independent component analysis, a new concept?," *Signal Processing*, vol. 36, pp. 287–314, 1994.

[22] T. Cover and J. Thomas, *Elements of Information Theory.* John Wiley, 1991.

[23] I. Csiszár, "I-divergence geometry of probability distributions and minimization problems," *Annals of Probability*, vol. 3, pp. 146–158, 1975.

[24] R. Datta, D. Joshi, J. Li, and J. Z. Wang, "Image retrieval: Ideas, influences, and trends of the new age," *ACM Computing Surveys*, vol. 40, pp. 1–60, 2008.

[25] J. De Bonet and P. Viola, "Structure driven image database retrieval," in *Neural Information Processing Systems,* vol. 10, Denver, Colorado, 1997.

[26] J. De Bonet, P. Viola, and J. Fisher, "Flexible histograms: A multiresolution target discrimination model," in *Proceedings of SPIE,* vol. 3370-12, (E. G. Zelnio, ed.), 1998.

[27] A. Dempster, N. Laird, and D. Rubin, "Maximum-likelihood from incomplete data via the EM Algorithm," *Journals of the Royal Statistical Society*, vol. B-39, 1977.

[28] L. Devroye, L. Gyorfi, and G. Lugosi, *A Probabilistic Theory of Pattern Recognition.* Springer-Verlag, 1996.

[29] T. Dietterich, R. Lathrop, and T. Lozano-Pere, "Solving the multiple-instance problem with axis-parallel rectangles," *Artificial Intelligence*, vol. 89, no. 1, pp. 31–71, 1997.

[30] R. Duda, P. Hart, and D. Stork, *Pattern Classification.* John Wiley and Sons, 2001.

[31] P. Duygulu, K. Barnard, D. Forsyth, and N. Freitas, "Object recognition as machine translation: Learning a Lexicon for a fixed image vocabulary," in *European Conference on Computer Vision*, Copenhagen, Denmark, 2002.

[32] Y. Ephraim, A. Denbo, and L. Rabiner, "A minimum discrimination information approach for hidden Markov modeling," *IEEE Transactions on Information Theory*, vol. 35, no. 5, pp. 1001–1013, September 1989.

[33] Y. Ephraim, H. Lev-Ari, and R. M. Gray, "Asymptotic minimum discrimination information measure for asymptotically weakly stationary processes," *IEEE Trans. on Information Theory*, vol. 34, no. 5, pp. 1033–1040, September 1988.

[34] C. Fellbaum, *Wordnet: An Electronic Lexical Database.* MIT Press, 1998.

[35] S. Feng, R. Manmatha, and V. Lavrenko, "Multiple Bernoulli relevance models for image and video annotation," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Washington DC, 2004.

[36] R. Fergus, P. Perona, and A. Zisserman, "Object class recognition by unsupervised scale-invariant learning," *IEEE Conference in Computer Vision and Pattern Recognition*, 2003.

[37] D. Field, "What is the goal of sensory coding?," *Neural Computation*, vol. 6, no. 4, pp. 559–601, January 1989.

[38] I. Fogel and D. Sagi, "Gabor filters as texture discriminators," *Biological Cybernitics*, vol. 61, pp. 103–113, 1989.

[39] K. Fukunaga, *Introduction to Statistical Pattern Recognition.* Academic Press, 1990.

[40] W. Gardner and B. Rao, "Theoretical analysis of the high-rate vector quantization of LPC parameters," *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 5, pp. 367–376, September 1995.

[41] A. Gelman, J. Carlin, H. Stern, and D. Rubin, *Bayesian Data Analysis.* Chapman Hall, 1995.

[42] A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression.* Kluwer Academic Press, 1992.

[43] R. M. Gray, "Vector quantization," *Signal Processing Magazine*, vol. 1, April 1984.

[44] R. M. Gray, A. Gray, G. Rebolledo, and J. Shore, "Rate-distortion speech coding with a minimum discrimination information distortion measure," *IEEE Transactions on Information Theory*, vol. IT-27, pp. 708–721, November 1981.

[45] V. Guillemin, *Differential Topology.* Pearson Education, 1974.

[46] M. Gupta and Y. Chen, "Theory and use of the EM method," *Foundations and Trends in Signal Processing, NOW Publishers*, vol. 4, pp. 223–296, 2010.

[47] N. Howe, "Percentile blobs for image similarity," in *Workshop in Content-based Access to Image and Video Libraries*, pp. 78–83, Santa Barbara, California, 1998.

[48] J. Huang, S. Kumar, M. Mitra, W. Zhu, and R. Zabih, "Spatial color indexing and applications," *International Journal of Computer Vision*, vol. 35, no. 3, pp. 245–268, December 1999.

[49] D. Hubel and T. Wiesel, "Brain mechanisms of vision," *Scientific American*, September 1979.

[50] A. Hyvarinen, J. Karhunen, and E. Oja, *Independent Component Analysis*. New York: Wiley, 2001.

[51] A. Hyvarinen and E. Oja, "Independent component analysis: Algorithms and applications," *Neural Networks*, vol. 13, pp. 411–430, 2000.

[52] F. Idris and S. Panchanathan, "Storage and retrieval of compressed sequences," *IEEE Transactions on Consumer Electronics*, vol. 41, no. 3, pp. 937–941, August 1995.

[53] G. Iyengar and A. Lippman, "Clustering images using relative entropy for efficient retrieval," in *International workshop on Very Low Bitrate Video Coding*, Urbana, Illinois, 1998.

[54] A. Jain and A. Vailaya, "Image retrieval using color and shape," *Pattern Recognition Journal*, vol. 29, pp. 1233–1244, August 1996.

[55] N. Jayant and P. Noll, *Digital Coding of Waveforms: Principles and Applications to Speech and Video*. Prentice Hall, 1984.

[56] T. Kailath, "The divergence and bhattacharyya distance measures in signal selection," *IEEE Transactions on Communication Technology*, vol. 15, pp. 52–60, February 1967.

[57] V. Kotel'nikov, *The Theory of Optimum Noise Immunity*. New York: McGraw-Hill, 1959.

[58] S. Kullback, *Information Theory and Statistics*. New York: Dover Publications, 1968.

[59] M. Kupperman, "Probabilities of hypothesis and information-statistics in sampling from exponential-class populations," *Annals of Mathematical Statistics*, vol. 29, pp. 571–574, 1958.

[60] V. Lavrenko, R. Manmatha, and J. Jeon, "A model for learning the semantics of pictures," in *Neural Information Processing Systems*, Denver, Colorado, 2003.

[61] V. Lavrenko, R. Manmatha, and J. Jeon, "A model for learning the semantics of pictures," in *Neural Information Processing Systems*, 2003.

[62] H. Lev-Ari, S. Parker, and T. Kailath, "Multidimensional maximum-entropy covariance extension," *IEEE Transactions on Information Theory*, vol. 35, no. 3, pp. 497–508, May 1988.

[63] J. Li, N. Chadda, and R. M. Gray, "Asymptotic performance of vector quantizers with a perceptual distortion measure," *IEEE Transactions on Information Theory*, vol. 45, no. 4, pp. 1082–1091, May 1999.

[64] J. Li and J. Wang, "Automatic linguistic indexing of pictures by a statistical modeling approach," *IEEE Transactions on Pattern. Analysis and Machine Intelligence*, vol. 25, no. 10, 2003.

[65] Q. Li, "Estimation of mixture models," PhD thesis, Yale University, 1999.

[66] T. Linder and R. Zamir, "High-resolution source coding for non-difference distortion measures: The rate-distortion function," *IEEE Transactions on Information Theory*, vol. 45, no. 2, pp. 533–547, March 1999.

[67] F. Liu and R. Picard, "Periodicity, directionality, and randomness: Wold features for image modeling and retrieval," *IEEE Transactions on Pattern. Analysis and Machine Intelligence*, vol. 18, no. 3, pp. 722–733, July 1996.

[68] Y. Liu, D. Zhang, G. Lu, and W. Ma, "A survey of content-based image retrieval with high-level semantics," *Pattern Recognition*, vol. 40, no. 1, pp. 262–282, 2007.

[69] W. Ma and H. Zhang, "Benchmarking of image features for content-based retrieval," in *Asilomar Conference on Signals, Systems, and Computers*, Asilomar, California, 1998.

[70] J. Malik and P. Perona, "Preattentive texture discrimination with early vision mechanisms," *Journal of the Optical Society of America*, vol. 7, no. 5, pp. 923–932, May 1990.

[71] S. Mallat, "A theory for multiresolution signal decomposition: The wavelet representation," *IEEE Transactions on Pattern. Analysis and Machine Intelligence*, vol. 11, pp. 674–693, July 1989.

[72] S. Mallat, *A Wavelet Tour of Signal Processing*. Academic Press, 1999.

[73] B. Manjunath and W. Ma, "Texture features for browsing and retrieval of image data," *IEEE Transactions on Pattern. Analysis and Machine Intelligence*, vol. 18, no. 8, pp. 837–842, August 1996.

[74] J. Mao and A. Jain, "Texture classification and segmentation using multiresolution simultaneous autoregressive models," *Pattern Recognition*, vol. 25, no. 2, pp. 173–188, 1992.

[75] O. Maron and T. Lozano-Perez, "A framework for multiple-instance learning," in *Neural Information Processing Systems 10*, Denver, Colorado, 1998.

[76] O. Maron and A. Ratan, "Multiple-instance learning for natural scene classification," in *Proceedings of 15th International Conference on Machine Learning*, 1998.

[77] B. Moghaddam, H. Bierman, and D. Margaritis, "Defining image content with multiple regions-of-interest," in *Workshop in Content-based Access to Image and Video Libraries*, pp. 89–93, Fort Collins, Colorado, 1999.

[78] Y. Mori, H. Takahashi, and R. Oka, "Image-to-word transformation based on dividing and vector quantizing images with words," in *First International Workshop on Multimedia Intelligent Storage and Retrieval Management*, 1999.

[79] H. Murase and S. Nayar, "Visual learning and recognition of 3-D objects from appearence," *International Journal of Computer Vision*, vol. 14, pp. 5–24, 1995.

[80] W. Niblack, R. Barber, W. Equitz, M. Flickner, E. Glasman, D. Pektovic, P. Yanker, C. Faloutsos, and G. Taubin, "The QBIC project: Querying images by content using color, texture, and shape," in *SPIE Storage and Retrieval for Image and Video Databases*, pp. 173–181, San Jose, California, 1993.

[81] N. Nill, "A visual model weighted cosine transform for image compression and quality assessment," *IEEE Transactions on Communications*, vol. 33, pp. 551–557, June 1985.

[82] B. Olshausen and D. Field, "Emergence of simple-cell receptive field properties by learning a sparse code for natural images," *Nature*, vol. 381, pp. 607–609, 1996.

[83] M. Ortega, Y. Rui, K. Chakrabarti, K. Porkaew, S. Mehrotra, and T. Huang, "Supporting ranked Boolean similarity queries in MARS," *IEEE Transactions*

on Knowledge and Data Engineering, vol. 10, no. 6, pp. 905–925, December 1998.
[84] A. Papoulis, *Probability, Random Variables, and Stochastic Processes*. McGraw-Hill, 1991.
[85] G. Pass and R. Zabih, "Comparing images using joint histograms," *ACM Journal of Multimedia Systems*, vol. 7, no. 3, pp. 234–240, May 1999.
[86] A. Pentland, R. Picard, and S. Sclaroff, "Photobook: Content-based manipulation of image databases," *International Journal of Computer Vision*, vol. 18, no. 3, pp. 233–254, June 1996.
[87] R. Picard, T. Kabir, and F. Liu, "Real-time recognition with the entire brodatz texture database," in *Proceedings of IEEE Conference on Computer Vision*, New York, 1993.
[88] M. Pinsker, *Information and Information Stability of Random Variables and Processes*. San Francisco: Holden-Day, 1964.
[89] J. Puzicha, Y. Rubner, C. Tomasi, and J. Buhmann, "Empirical evaluation of dissimilarity measures for color and texture," in *International Conference on Computer Vision*, pp. 1165–1173, Korfu, Greece, 1999.
[90] L. Rabiner and B. Juang, *Fundamentals of Speech Recognition*. Prentice Hall, 1993.
[91] N. Rasiwasia, P. Moreno, and N. Vasconcelos, "Bridging the gap: Query by semantic example," *IEEE Transactions on Multimedia*, vol. 9, no. 5, pp. 923–938, August 2007.
[92] N. Rasiwasia, J. C. Pereira, E. Coviello, G. Doyle, G. Lanckriet, R. Levy, and N. Vasconcelos, "A new approach to cross-modal multimedia retrieval," in *ACM Proceedings of the International Conference on Multimedia*, 2010.
[93] N. Rasiwasia and N. Vasconcelos, "Scene classification with low-dimensional semantic spaces and weak supervision," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2008.
[94] N. Rasiwasia and N. Vasconcelos, "Holistic context modeling using semantic co-occurrences," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009.
[95] R. Redner and H. Walker, "Mixture densities, maximum likelihood and the EM algorithm," *SIAM Review*, vol. 26, no. 2, pp. 195–239, April 1984.
[96] Y. Rui, T. Huang, and S.-F. Chang, "Image retrieval: Current techniques, promising directions and open issues," *Journal of Visual Communication and Image Representation*, vol. 10, pp. 39–62, March 1999.
[97] Y. Rui, T. Huang, M. Ortega, and S. Mehrotra, "Relevance feedback: A power tool for interactive content-based image retrieval," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 8, no. 5, pp. 644–655, September 1998.
[98] D. Sagi, "The Psychophysics of Texture Segmentation," in *Early Vision and Beyond* , chapter 7, (T. Papathomas, ed.), MIT Press, 1996.
[99] H. Sakamoto, H. Suzuki, and A. Uemori, "Flexible montage retrieval for image data," in *SPIE Storage and Retrieval for Image and Video Databases*, San Jose, California, 1994.

[100] B. Schiele and J. Crowley, "Recognition without correspondence using multidimensional receptive field histograms," *International Journal of Computer Vision*, vol. 36, no. 1, pp. 31–50, January 2000.

[101] C. Schmid and R. Mohr, "Local greyvalue invariants for image retrieval," *IEEE Transactions on Pattern. Analysis and Machine Intelligence*, vol. 19, no. 5, pp. 530–535, May 1997.

[102] D. Scott, *Multivariate Density Estimation: Theory, Practice, and Visualization*. John Wiley, 1992.

[103] C. Shyu, C. Brodley, A. Kak, A. Kosaka, A. Aisen, and L. Broderick, "Local versus global features for content-based image retrieval," in *Workshop in Content-based Access to Image and Video Libraries*, pp. 30–34, Santa Barbara, California, 1998.

[104] J. Simonoff, *Smoothing Methods in Statistics*. Springer-Verlag, 1996.

[105] A. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-based image retrieval: The end of the early years," *IEEE Transactions on Pattern. Analysis and Machine Intelligence*, vol. 22, no. 12, pp. 1349–1380, 2000.

[106] J. Smith, "Integrated spatial and feature image systems: Retrieval, compression and analysis," PhD thesis, Columbia University, 1997.

[107] J. Smith and S. Chang, "VisualSEEk: A fully automated content-based image query system," in *ACM Multimedia*, pp. 87–98, Boston, Massachussetts, 1996.

[108] M. Stricker and A. Dimai, "Color indexing with weak spatial constraints," in *SPIE Storage and Retrieval for Image and Video Databases*, pp. 29–40, San Jose, California, 1996.

[109] M. Stricker and M. Orengo, "Similarity of color images," in *SPIE Storage and Retrieval for Image and Video Databases*, San Jose, California, 1995.

[110] M. Stricker and M. Swain, "The capacity of color histogram indexing," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 704–708, 1994.

[111] A. Sutter, J. Beck, and N. Graham, "Contrast and spatial variables in texture segregation: Testing a simple spatial-frequency channels model," *Perceptual Psychophysics*, vol. 46, pp. 312–332, 1989.

[112] M. Swain and D. Ballard, "Color Indexing," *International Journal of Computer Vision*, vol. 7, no. 1, pp. 11–32, 1991.

[113] L. Taycher, M. Cascia, and S. Sclaroff, "Image digestion and relevance feedback in the image rover WWW search engine," in *Visual*, San Diego, California, 1997.

[114] D. Titterington, A. Smith, and U. Makov, *Statistical Analysis of Finite Mixture Distributions*. John Wiley, 1985.

[115] A. Tousch, S. Herbin, and J. Audibert, "Semantic hierarchies for image annotation: A survey," *Pattern Recognition*, vol. 45, pp. 333–345, 2012.

[116] H. V. Trees, *Detection, Estimation, and Modulation Theory*. Wiley, 1968.

[117] M. Turk and A. Pentland, "Eigenfaces for recognition," *Journal of Cognitive Neuroscience*, vol. 3, 1991.

[118] V. Vapnik, *The Nature of Statistical Learning Theory*. Springer Verlag, 1995.

[119] M. Vasconcelos, G. Carneiro, and N. Vasconcelos, "Weakly supervised top-down image segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2006.

[120] M. Vasconcelos and N. Vasconcelos, "Natural image statistics and low-complexity feature selection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, pp. 228–244, 2009.

[121] N. Vasconcelos, "Bayesian models for visual information retrieval," PhD thesis, Massachusetts Institute of Technology, 2000.

[122] N. Vasconcelos, "Image indexing with mixture hierarchies," in *Proceedings of IEEE Computer Vision and Pattern Recognition Conference*, Kawai, Hawaii, 2001.

[123] N. Vasconcelos, "Exploiting group structure to improve retrieval accuracy and speed in image databases," in *Proceedings of International Conference Image Processing*, Rochester, NY, 2002.

[124] N. Vasconcelos, "Minimum probability of error image retrieval," *IEEE Transactions on Signal Processing*, vol. 52, pp. 2322–2336, 2004.

[125] N. Vasconcelos, "On the efficient evaluation of probabilistic similarity functions for image retrieval," *IEEE Transactions on Information Theory*, vol. 50, pp. 1482–1496, 2004.

[126] N. Vasconcelos and A. Lippman, "Library-based coding: A representation for efficient video compression and retrieval," in *Proceedings of Data Compression Conference*, Snowbird, Utah, 1997.

[127] N. Vasconcelos and A. Lippman, "Learning from user feedback in image retrieval systems," in *Neural Information Processing Systems*, Denver, Colorado, 1999.

[128] N. Vasconcelos and A. Lippman, "Learning over multiple temporal scales in image databases," in *Proceedings of European Conference on Computer Vision*, Dublin, Ireland, 2000.

[129] N. Vasconcelos and A. Lippman, "A probabilistic architecture for content-based image retrieval," in *Proceedings of IEEE Computer Vision and Pattern Recognition Conference*, Hilton Head, North Carolina, 2000.

[130] X. Wan and C. Kuo, "A new approach to image retrieval with hierarchical color clustering," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 8, no. 5, pp. 628–643, September 1998.

[131] J. Wang, G. Wiederhold, O. Firschein, and A. Wei, "Content-based image indexing and searching using daubechies' wavelets," *International Journal of Digital Libraries*, vol. 1, pp. 311–328, 1997.

[132] L. Xie, R. Yan, J. Tesic, A. Natsev, and J. R. Smith, "Probabilistic visual concept trees," in *ACM Multimedia*, 2010.