CONTENT-BASED RETRIEVAL FROM IMAGE DATABASES: CURRENT SOLUTIONS AND FUTURE DIRECTIONS

Nuno Vasconcelos

Compaq Cambridge Research Laboratory

Murat Kunt

Ecole Polytechnique Federale de Lausanne

ABSTRACT

We review recent advances in image retrieval. The two fundamental components of a retrieval system, representation and learning, are analyzed. Each component is decomposed into its constituent building blocks: features, feature representation, and similarity function for the representation; short- and long-term procedures for learning. We identify a series of requirements for each of the sub-areas, e.g. optimality, invariance, perceptual relevance, computational tractability, and point out various approaches proposed to satisfy them. Several open problems are also identified.

1. INTRODUCTION

Research in the analysis, classification, and retrieval of images from large visual repositories is, at the present time, one of the most active topics in image processing. There are a few reasons for this. First, the retrieval problem is of great practical interest: while digital cameras make picture taking inexpensive and large amounts of new imagery become available on the web every day, there is still a shortage of effective tools for searching/manipulating visual content. Second, because it touches a significant number of unsolved challenging questions in image understanding (e.g. image similarity, segmentation, shape, invariance, etc.), the retrieval problem is also interesting at a deeper theoretical level. Finally, visual databases provide a new testing ground to evaluate image processing ideas, where it is not acceptable to make strict assumptions about scene or imaging conditions or test an algorithm on a few images alone. In particular, the introduction of several large databases has lead to the reevaluation of old ideas, allowing a better understanding of what works and what does not.

The goal of the current special session is two-fold: to assess 1) how much of the problem has been solved, and 2) what are the most challenging directions to address in the next few years. This paper addresses the first point by identifying the main components of a retrieval system, and briefly reviewing common solutions to the problems posed by each component. The review is not meant to be exhaustive, but simply to provide a unifying context to the remaining papers of the session, where the second point is addressed through the presentation of various exciting research directions.

2. THE RETRIEVAL COMPONENTS

At the coarsest level, one can identify two major components of the retrieval problem: representation and learning. The representation establishes a computational basis for the retrieval operation, e.g. by defining a set of features and a similarity function. Learning relies on the representation to address the dynamic aspects of a retrieval system, namely how to adapt to time-varying user requests. While learning is not mandatory, it leads to more effective retrieval systems.

2.1. Representation

A representation for content-based image retrieval consists of three fundamental building blocks: a feature transformation, a feature representation, and a similarity function. In this section we analyze the role of each of these modules in the overall retrieval architecture, investigate what is the minimal set of requirements that they must satisfy, and check how those requirements are fulfilled by existing retrieval solutions.

2.1.1. Feature transformation

A feature transformation is a mapping from the space of image observations (usually image pixels) to a feature space that has better properties for the retrieval operation. Feature transformations have been widely studied in the texture literature, where the emphasis has been on discrimination. Under this perspective, the feature transformation is the most important component of the retrieval architecture: independently of how the observation space is populated, the various image classes that compose the database should be clearly separated in feature space. If such separation is achieved, the remaining components become fairly easy to design. In fact, small emphasis has been given to them in the texture literature, where simplistic feature representations (e.g. feature mean and covariance) and similarity functions (e.g. Euclidean distance) are fairly common [1, 2, 3].

Discrimination based on the features alone is difficult to achieve in the generic retrieval context, where there is no control over the classes of images to be processed. Because discriminant features tend to be domain specific (e.g. autoregressive models work well for texture but not for faces), the transformation that achieves clean separation in one domain may have the inverse effect in another.

2.1.2. Invariance and perceptual relevance

In addition to being generic and discriminant, the feature transformation should exhibit two important properties: invariance and perceptual relevance. Invariant transforma-

0-7803-6725-1/01/$10.00 ©2001 IEEE

tions are those robust to changes in either imaging conditions (e.g. lighting) or scene layout (e.g. object pose). Perceptually relevant transformations mimic, in some way, the properties of the human visual system. This does not mean that to be perceptually relevant a transformation has to be biologically plausible, since retrieval systems are not subject to the constraints of neural hardware.

While invariance has been extensively studied in machine vision, the majority of the proposed solutions are not directly applicable to the retrieval problem. For example, invariant object recognition techniques commonly assume a training set of cleanly segmented views of each object [4, 5]. Similarly, invariant texture features typically rely on the assumption of segmented texture patches under frontal view and subject to a limited set of transformations [3, 2]. In the retrieval context, invariance has been studied mostly for color-based representations. A possible reason for this is that, by making quite generic assumptions regarding the surfaces of objects in the world, it is possible to derive sophisticated forms of color invariance. This is exemplified by the work of Smeulders et al, as discussed in [6].

At the simplest levels of image representation, the mechanisms of human vision are fairly well understood. For example, various perceptual color-spaces are readily available [7] and have been widely used in the retrieval literature. For texture, a popular model consists of a space/space-frequency (e.g. wavelet) decomposition, followed by a linearity involving some form of rectification, and a pooling stage combining information from different space-frequency channels [8]. Experience in image compression has shown that, besides capturing various properties of human vision, space/space-frequency decompositions have coding performance close to optimal in terms of energy compaction. Given that a feature transformation must achieve a good balance between the amount of noisy information that is discarded (to improve invariance) and the amount of signal that is kept (to be discriminant) this is a relevant result.

Recent research in biological vision has gone one step further and actually shown that the combination of the energy compactness constraint with a sparseness constraint is sufficient to derive feature transformations that exhibit remarkable resemblance to the receptive fields of the cells in the early stages of the visual cortex [9]. It turns out that most wavelet representations are indeed sparse and they therefore provide a good approximation to the feature transformation performed by early human vision [10].

Since wavelets are generic, in fact invertible, transformations this suggests that wavelet-based representations should enable retrieval with low error probability on a wide spectrum of image domains. On the other hand, it contradicts earlier texture retrieval experience which has shown that generic frequency transformations, such as wavelets and the Fourier transform, were consistent under-achievers when compared to texture-specific transformations such as auto-regressive models [1, 2]. Even worse, these results showed that the performance loss could be significant.

Recent studies have shown that, while when combined with trivial feature representations (e.g. sample mean and covariance) and similarity functions (e.g. Mahalanobis distance) the multiresolution features can indeed perform very poorly, the differences become negligible for more sophisticated architectures [11]. I.e. the problem is not the frequency decomposition itself, but the discriminant mind set that makes feature transformation the central component of the architecture. This exemplifies how it makes little sense to find the best solution for one component of the architecture without considering the others.

2.2. Feature representation

Keeping track of all the feature vectors extracted from each image would pose a major difficulty to any retrieval system. Hence, there is a need for a feature representation to summarize the distribution of feature vectors.

As mentioned above, early texture retrieval relied on summarization by the first two sample moments. This is equivalent to a Gaussian assumption for the feature density. While computationally efficient, this assumption is unrealistic for the vast majority of real images, which are characterized by multimodal densities. The lack of expressiveness of the Gaussian (or, for that matter, of any of the parametric density models in common use) was realized early on in color retrieval where the histogram rapidly emerged as the standard representation [12, 1, 13, 14]. Histograms produce significantly more accurate estimates than the Gaussian, and are also fairly easy to compute.

These two attributes, expressiveness and computational tractability, are in fact the two main requirements for an effective feature representation. Notice that there are two aspects to tractability: the complexity of density estimation and the complexity of evaluating similarity. While the former is an off-line process that typically does not have great impact on the performance of retrieval systems, the latter must be performed thousands, or millions, of times for every retrieval operation and should be fast.

While the histogram is both expressive and tractable in low dimensional spaces (such as the three dimensional color spaces used by most color retrieval algorithms) it does not retain these properties in high dimensions. On the contrary, histogram complexity (number of bins) is exponential in the dimension of the space. This limits the applicability of histograms to texture retrieval, where the need to model spatial interactions between neighboring image pixels invariably leads to high-dimensional feature vectors. To overcome this limitation one has to rely on 1) histogram extensions such as the color correlogram [15] or multimodal neighborhood signatures [16], or 2) alternative density models, such as vector quantizers [17], kernel estimators, or mixtures [18], that scale better with the dimension of the space. A good example is the paper by Neemuchwala et al [19] where the randomized decision trees of [20] provide density estimates in a 16 dimensional space. Many of these more sophisticated representations are closely related, involving different trade-offs between off-line and on-line computational complexity and expressive power [11]. The paper by Gray et al [21] introduces a new algorithm for vector quantization and provides new insights on the relationships between vector quantizers and Gaussian mixtures.

2.3. Similarity function

Given a feature representation for each database image, retrieval consists of extracting a set of feature vectors from

a query image and relying on a similarity function to evaluate which feature representation best explains those features. Once again, early texture efforts used simple metrics that are only appropriate in the Gaussian context, e.g. the Euclidean or Mahalanobis distances. More sophisticated feature representations require the ability to match entire densities. This has been accomplished in at least three different ways: through L^p norms, maximum likelihood (ML), or information theoretic criteria.

The L^p norm of the distance between two densities $p(x)$ and $q(x)$ is defined by

$$\left[\int |p(x) - q(x)|^p dx\right]^{\frac{1}{p}}. \quad (1)$$

L^p norms have been quite popular in color retrieval. When $p = 1$, they reduce to the histogram intersection metric [12]. ML retrieval methods evaluate the likelihood of the query vectors according to each database density and pick the density that maximizes this quantity. Information-theoretic similarity functions include statistical criteria such as the $\chi^2 distance$, the Itakura-Saito distortion metric, commonly found in the speech literature, the Euclidean, and Mahalanobis distances. All of these are particular cases of the relative entropy or Kulback-Leibler divergence (KLD)

$$KL(p||q) = \int p(x) \log \frac{p(x)}{q(x)} dx \quad (2)$$

that make various assumptions or approximations to the underlying densities $p(x)$ and $q(x)$ [22]. The minimization of the KLD can be interpreted as the solution to the classification problem known as minimum discrimination information (MDI). It can be shown that MDI is equivalent to ML when the cardinality of the set of query feature vectors grows to infinity (a relationship exploited in the paper by Gray et al [21]). ML is, in turn, a particular case of the maximum a-posteriori probability (MAP) criteria that is well known to minimize the probability of classification error [23]. A closely related information theoretic similarity function, the α-divergence, is introduced in the paper by Neemuchwala et al [19]. Like MAP, it is optimal in the decision theoretic sense, but for a slightly different problem: that of deciding if two random variables (the query and the one from the database) are independent or not. In summary, ML and information-theoretic similarity functions formulate retrieval as a classification problem, leading to discriminant solutions where the burden of discrimination does not rest solely on the feature transformation.

2.4. Shape

Most of what has been covered so far emphasizes texture and color retrieval. There is however a third component of retrieval representations that is crucial for the success of retrieval systems: shape. Shape is at the core of object-based representations and plays a central role in perceptual judgments of similarity. Unfortunately, it is not clear that shape can be used for generic retrieval without requiring the solution of the segmentation problem, and this is well known to be very hard. Nevertheless, significant effort has been devoted to segmentation in the last few years and encouraging progress reported in areas like probabilistic [24] and graph-theoretical [25] segmentation methods.

In the mean time, shape retrieval finds application in scenarios where it is realistic to assume that cleanly segmented images are available, e.g. databases of trademarks or image silhouettes. Several representations have been proposed, including simple histograms of edge direction [26], more sophisticated forms of contour parameterization [27], or combinations of a local shape description to achieve coarse correspondence and global splines to account for deformation [28]. An elegant decomposition of shape similarity into its structural and metric components is possible with the shock graph representation introduced by Kimia and colleagues [29]. Structural similarity is based on a coarse description of the geometric relationships between the parts that compose each shape, metric similarity captures the cost of finely aligning two shapes. The paper by Sebastian and Kimia [30] presents a comparison between retrieval based on shock graphs and curve matching.

3. LEARNING

Image retrieval is usually an interactive process where 1) system makes suggestions, 2) user provides feedback, 3) system updates suggestions, and the process is iterated. This can be tedious, in particular if the system does not appear to make smart use of the previous interaction, and there is a need for systems that learn from user feedback. Learning should take place both within a retrieval session (short-term) and across retrieval sessions (long-term).

3.1. Short-term learning

The goal of short-term learning, also known as relevance feedback, is to minimize the average number of iterations required for convergence within a retrieval session. Typically, user feedback is integrated throughout the retrieval session and used as guidance for tuning the free parameters of the underlying retrieval algorithm. For example, a system that relies on different representations for color, texture, and shape might adapt the weights given to the three components according to the user selections [31]. Other possibilities include adapting the weight of the different features within a given representation [32], or the similarity function [33].

Most principled short-term learning algorithms can be grouped into two main classes: geometric and statistic. Geometric methods rely on the Euclidean distance, or variations of it, and strive to find the query vector that minimizes the distance to the examples provided by the user. An optimal joint solution for the query vector, the feature transformation, and the similarity function was presented in [32]. Statistical methods can be further subdivided into generative and discriminant, according to the nature of the underlying representation. Generative methods are based on the MAP criteria and the feature representations discussed in section 2.2. Integration of user feedback is achieved by searching for decisions that are optimal with respect the entire retrieval session, not just the current iteration [34, 35]. This can be done very efficiently through the use of be-

lief propagation algorithms for updating the probabilities of different hypothesis [35].

Discriminant methods strive to directly design the classifier that best separates the positive from negative user examples. This is accomplished by explicitly finding the boundaries in feature space that best separate the two classes. These approaches are based on discriminant classifiers, such as perceptrons and support vector machines [36], that are widely used in machine learning. Statistical procedures such as boosting [37] and linear discriminant analysis have been proposed to guide the boundary updates from iteration to iteration. A comparative analysis of various discriminant techniques is presented in the paper by Huang and Zhou [36].

3.2. Long-term learning

While short-term learning is confined to a given retrieval session, concepts acquired through long-term learning persist across retrieval sessions. Typically, long-term learning involves asking the user to label some examples that are then processed off-line. Learning techniques can be used to classify the remaining database images according to each of the concepts defined by the user. The problem fits in the framework of weakly supervised learning that has recently attracted attention in the vision literature [38, 20].

While various success stories have developed in the last few years in areas such as face detection [39] and recognition [40], the resulting systems typically require very large training sets and careful performance tuning. When compared to the amount of resources that a typical user is willing to spend training a retrieval system, these solutions can be seen as taking "infinite-time" and having "infinite training complexity". Consequently, they are unlikely to be deployed for all visual concepts that a user may be interested in searching for. In fact, the set of such concepts is not even well defined since it depends on the user and the particular query. Furthermore, features of predominant interest are those of a semantic nature, as discussed in the paper by Kittler et al [41].

The goal of weakly supervised learning is to extend the capabilities of current recognition architectures by making them able to learn visual concepts from a few, non-segmented, examples. If successful, such architectures will play a crucial role in the personalization of retrieval systems, by allowing users to effortlessly define the set of visual concepts that are most relevant to them.

4. REFERENCES

[1] W. Ma and H. Zhang, "Benchmarking of Image Features for Content-based Retrieval," in 32^{nd} Asilomar Conference on Signals, Systems, and Computers, Asilomar, California, 1998.

[2] R. Picard, T. Kabir, and F. Liu, "Real-time Recognition with the entire Brodatz Texture Database," in Proc. IEEE Conf. on Computer Vision, New York, 1993.

[3] J. Mao and A. Jain, "Texture Classification and Segmentation Using Multiresolution Simultaneous Autoregressive Models," Pattern Recognition, vol. 25, no. 2, pp. 173–188, 1992.

[4] 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.

[5] A. Georghiades, D. Kriegman, and P. Belhumeur, "Illumination Cones for Recognition Under Variable Lighting: Faces," in IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Santa Barbara, California, 1998.

[6] A. Smeulders, J. Geusebroek, and T. Gevers, "The Use of Invariant Representations in Image Retrieval," in IEEE Int. Conf. on Image Processing, Thessaloniki, Greece, 2001.

[7] A. Byrne and D. Hilbert, Eds., Readings on Color, MIT Press, 1997.

[8] D. Sagi, "The Psychophysics of Texture Segmentation," in Early Vision and Beyond, T. Papathomas, Ed., chapter 7. MIT Press, 1996.

[9] 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.

[10] J. Portilla and E. Simoncelli, "Texture Modeling and Synthesis using Joint Statistics of Complex Wavelet Coefficients," in IEEE Workshop on Statistical and Computational Theories of Vision, Fort Collins, Colorado, 1999.

[11] N. Vasconcelos, Bayesian Models for Visual Information Retrieval, Ph.D. thesis, Massachusetts Institute of Technology, 2000.

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

[13] J. Puzicha, Y. Rubner, C. Tomasi, and J. Buhmann, "Empirical Evaluation of Dissimilarity Measures for Color and Texture," in International Conference on Computer Vision, Korfu, Greece, 1999, pp. 1165–1173.

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

[15] 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.

[16] J. Matas, D. Koubaroulis, and J. Kittler, "Colour Image Retrieval and Object Recognition Using the Multimodal Neighborhood Signature," in European Conference on Computer Vision, Dublin, Ireland, 2000, pp. 48–66.

[17] K. Popat and R. Picard, "Cluster-based Probability Model and its Application to Image and Texture Processing," IEEE Trans. on Image Processing, vol. 6, no. 2, pp. 268–284, 1997.

[18] N. Vasconcelos and A. Lippman, "A Probabilistic Architecture for Content-based Image Retrieval," in Proc. IEEE Computer Vision and Pattern Recognition Conf., Hilton Head, North Carolina, 2000.

[19] H. Neemuchwala, A. Hero, and P. Carson, "Feature Coincidence Trees for Registration of Ultrasound Breast Images," in IEEE Int. Conf. on Image Processing, Thessaloniki, Greece, 2001.

[20] Y. Amit and D. Geman, "Shape Quantization and Recognition with Randomized Trees," Neural Computation, vol. 9, pp. 1545–1588, 1997.

[21] R. Gray, J. Young, and A. Aiyer, "Minimum Discrimination Information Clustering: Modeling and Quantization with Gauss Mixtures," in IEEE Int. Conf. on Image Processing, Thesaloniki, Greece, 2001.

[22] N. Vasconcelos, "A Unified View of Image Similarity," in Proc. Int. Conf. Pattern Recognition, Barcelona, Spain, 2000.

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

[24] Y. Weiss and E. Adelson, "A Unified Mixture Framework for Motion Segmentation: Incorporating Spatial Coherence and Estimating the Number of Models," in Proc. Computer Vision and Pattern Recognition Conf., 1996.

[25] J. Shi and J. Malik, "Normalized Cuts and Image Segmentation," IEEE Trans. on Pattern. Analysis and Machine Intelligence, vol. Vol. 22, pp. 888–905, August 2000.

[26] A. Jain and A. Vailaya, "Shape-Based Retrieval: A Case Study with Trademark Image Databases," Pattern Recognition Journal, vol. 21, no. 9, pp. 13699–1390, 1998.

[27] T. Tasdizen, "Boundary Estimation from Intensity/Color Images with Algebraic Curve Models," in International Conference on Pattern Recognition, Barcelona, Spain, 2000.

[28] S. Belongie, J. Malik, and J. Puzicha, "Matching Shapes," in International Conference on Computer Vision, Vancouver, Canada, 2001.

[29] D. Sharvit, J. Chan, H. Tek, and B. Kimia, "Symmetry-based Indexing of Image Databases," in Workshop in Content-based Access to Image and Video Libraries, 1998, Santa Barbara, California, pp. 56–62.

[30] T. Sebastian andB. Kimia, "Curves vs Skeletons in Object Recognition," in IEEE Int. Conf. on Image Processing, Thessaloniki, Greece, 2001.

[31] T. Minka and R. Picard, "Interactive learning using a "society of models"," Pattern Recognition, vol. 30, pp. 565–582, 1997.

[32] Y. Rui and T. Huang, "Optimizing Learning in Image Retrieval," in IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Hilton Head, South Carolina, 2000.

[33] L. Taycher, M. Cascia, and S. Sclaroff, "Image Digestion and Relevance Feedback in the Image Rover WWW Search Engine," in Visual, San Diego, California, 1977.

[34] I. Cox, M. Miller, S. Omohundro, and P. Yianilos, "PicHunter: Bayesian Relevance Feedback for Image Retrieval," in Int. Conf. on Pattern Recognition, Vienna, Austria, 1996.

[35] N. Vasconcelos and A. Lippman, "Learning Over Multiple Temporal Scales in Image Databases," in Proc. European Conference on Computer Vision, Dublin, Ireland, 2000.

[36] T. Huang and X. Zhou, "Image Retrieval and Relevance Feedback: From Heuristic Weight Adjustment to Optimal Learning Methods," in IEEE Int. Conf. on Image Processing, Thessaloniki, Greece, 2001.

[37] K. Tieu and P. Viola, "Boosting Image Retrieval," in IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Hilton Head, South Carolina, 2000.

[38] M. Weber, M. Welling, and P. Perona, "Unsupervised Learning of Models for Recognition," in European Conference on Computer Vision, Dublin, Ireland, 2000, pp. 18–32.

[39] H. Rowley, S. Baluja, and T. Kanade, "Neural Network-Based Face Detection," IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 20, no. 1, pp. 23–38, January 1998.

[40] M. Turk and A. Pentland, "Eigenfaces for Recognition," Journal of Cognitive Neuroscience, vol. 3, 1991.

[41] J. Kittler, W. Christmas, B. Obadia, and D. Koubaroulis, "Generation of Semantic Cues for Sports Video Annotation," in IEEE Int. Conf. on Image Processing, Thessaloniki, Greece, 2001.