Image Databases are not Databases with Images

Simone Santini and Ramesh Jain

Center for Information Engineering, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093-0407. ssantini@cs.ucsd.edu, jain@ece.ucsd.edu

Abstract. In this paper, we discuss a number of new problems that arise in image databases, and that set them apart from traditional databases. The fact that image databases are based on similarity, rather than matching, creates a whose set of new issues.Most noticeably, while matching is, by and large, a well defined concept, there are many possible types of similarities. In this paper, we consider the problem of simulating human similarity perception. We argue that a satisfactory solution is possible for *preattentive similarity*, and we present a general and comprehensive geometric similarity model.

1 Introduction

Searching information in large images repositories is a problem of great importance for the development of visual information systems and has received considerable attention in the last few years [5, 3, 10]. While consolidated techniques for accessing structured information have been studied for many years, accessing inherently unstructured and sparse information is a problem whose surface we are just beginning to scratch.

The solution of this problem required a fresh rethinking of many concepts taken for granted in traditional databases. The most relevant of these is possibly the concept of matching. In traditional databases, matching is a binary operation[1]: every item either matches the query, or doesn't. On the other hand, when searching an image in a repository, typically we don't have a specific target in mind: we have some generic image (like "the image of a red Ferrari"), and try to retrieve something similar to that. Because of this difference, search in image databases should abandon the matching paradigm, and rely instead on *similarity* searches. In a similarity search we order the images with respect to similarity with the query, given a fixed similarity criterion. A similarity measure must behave "well" (in the sense of matching human similarity) for any pair of images, no matter how different they are. This is quite a different requirement from what we ask to a matching technique, which only has to behave well when there is relatively little difference between a database image and the query.

[1] An exception to this rule is represented by image retrieval systems that, although based on text searches, are closer in spirit to the approach presented here.

2 Similarity Theories

Computer vision research has been influenced by psychology studies on human perception in many ways. A number of researchers were influenced by the Gestalt school. This influence has had the unfortunate effect of concentrating the attention of the researchers on the *observer* (the human they were trying to emulate) rather than on the relation between the observer and its environment. This is sadly ironic since, at the same time, psychologists were recognizing the validity of an *ecological* approach [4], admitting that "*we will have to study the environment as carefully as we do the organism*[2] [9]". If we take the environment into account, we notice that—at least for an application like image databases—trying to replicate certain functions of human perception is pointless.

At a superficial analysis, the images upon which a database operates appear as a reasonable approximation of the stimuli experienced by people. After all, if we take one of those images and print it, we can easily see its contents. There are, however, important differences: the world of human beings is never static, and is subject to an incessant exploration made possible by motion; the world of the database is composed of a number of still images that are available immediately in their entirety. These differences are complemented by the different use that the humans and the machine make of visual information: humans and other mammals use it to identify places to run from (dangers) and objects to grasp (food, weapons...); the database uses the information to select some stimuli over others.

Several concepts that are familiar to humans must be revised or abandoned when the environment and the goals of the observer are changed. For instance, the concept of *object* is connected to the fact that certain stimuli in the visual field of animals correspond to entities on which the animal can operate: eat them, or grasp them, or seat on them, etc. It is ecologically valuable to consider these stimuli as representations of independent units (viz. the objects). Since a search engine in a database does not participate actively in its world, and its visual stimuli are static and hopelessly two-dimensional, the concept of "object" has no correspondent in databases. At least, not with the generality the concept has for human observers.

In spite of these differences, natural (human and animal) similarity perception is interesting to us for two reasons. First, we have to replicate—albeit in a different environment—certain characteristics of human similarity perception. Second, there are animals whose perceptual worls is at least as remote from the humans as that of a database. Studying the differences and commonalities in the solutions found by these animals, we can derive useful indications on general principles, valid in a wide range of environments. We argue that the only constructs that can, at least approximately, translate to the database domain are

[2] This should not be intended as an endorsement *in toto* of the ecological approach. Several of its methodological aspect are still vividly debated inside the psychological community; the main idea of a careful study of the interactions of an organism with the environment, though, has been generally accepted, even by critics

preattentive and, therefore, the similarity concept to be used in image databases should be preattentive. Preattentive similarity judgment is done without focusing attention on any part of the image. The higher processes responsible for recognition also cannot operate and therefore preattentive perception is in general based on different features than attentive perception [6].

Attentive and high-level recognition processes are in many cases domain dependent and representative of learned ability. We will not deal, in this paper, with this type of domain-dependent knowledge.

2.1 Is similarity a Distance?

A number of models [12] assume that human similarity assessment is based on the measurement of a suitable *distance* in a psychological space. Stimuli are translated into points in a *perceptual space*, and the similarity between stimuli p_1 and p_2 is a function of $p_1 - p_2$. Similarity is then a function of a distance d that satisfies the metric axioms:

Constancy of self-similarity: for all stimuli p it is $d(p,p) = Const.$
Minimality: for all stimuli p_1 and p_2, it is $d(p_1,p_1) \leq d(p_1,p_2)$.
Symmetry: for all stimuli p_1 and p_2, it is $d(p_1,p_2) = d(p_2,p_1)$.
Triangle inequality: for all stimuli p_1, p_2, p_3, it is $d(p_1,p_3) \leq d(p_1,p_2) + d(p_2,p_3)$

This is also the common assumption in vision applications. On the other hand, there is convincing evidence that human similarity does not satisfy the metric axioms [7, 13]. Some relatively recent models in psychology make the assumption that, since so many metric axioms are violated, similarity assessment in human is not based on a distance function after all. One successful approach is based on set-theoretic considerations.

In a 1977 paper [13], Amos Tversky proposed his famous *Feature Contrast Model.* Instead of considering stimuli as points in a metric space, Tversky characterized them as sets of features. Let p_1, p_2 be two stimuli, and P_1, P_2 the respective sets of features. Also, let $s(p_1,p_2)$ be a measure of the similarity between p_1 and p_2.

Theorem 1. *Let s be a similarity function*[3]*. Then there are a similarity function S and a non-negative function f such that, for all p_1,p_2,p_3,p_4:*

- $S(p_1,p_2) \geq S(p_3,p_4) \iff s(p_1,p_2) \geq s(p_3,p_4)$
- $S(p_1,p_2) = f(P_1 \cap P_2) - \alpha f(P_1 - P_2) - \beta f(P_2 - P_1)$

This result implies that any similarity ordering can be obtained using a linear combination (contrast) of a function of the common features ($P_1 \cap P_2$) and the distinctive features ($P_1 - P_2$ and $P_2 - P_1$). This representation is called the

[3] There are some technical hypotheses about this function that we omit here for the sake of brevity. For details, the reader should refer to [13].

contrast model. Note that Tversky assumes sets of "binary" features, which can also be seen as sets of predicates which are true for a stimulus.

To this date, this model has given some of the best explanations of experimental data, and it can account for all the violations of the distance axioms observed in experiments. In particular, $S(p_1, p_2)$ is asymmetric if $\alpha \neq \beta$, and the self-similarity, $S(p_1, p_1) = f(P_1)$ is not constant, depending on the saliency of the stimulus P_1.

3 Application of Psychological Ideas

The ideas introduced in the previous section are not immediately applicable to artificial observers. To obtain a good definition of similarity we must consider the relation between the artificial observer (the search engine) and its environment (the images). The stimuli that the search engine receives in a general database lack much of the liveness and coherence of the world in which humans live. The only aspect of human (and animal) similarity perception that can survive in this environment is preattentive. This means that the database should not try to recognize objects or interpret the image in any way, but will operate acritically on the data considered as simple and meaningless patches of color. The similarity measures between two images will derive from this interpretation. They will not be based on high level constructs like three-dimensional models of objects, or even on recognition of objects. The similarity measures will be based only on the two-dimensional structure if the image matrix.

3.1 Fuzzy Feature Contrast Model

Consider a typical task for a restricted domain database: assessing the similarity between faces. A face is characterized by a number of different features, like the width of the mouth, the darkness of the skin, the distance between the eyes. One problem we have to solve is how to go from these features—which usually are expressed as a real-valued measurement on the image—to the predicate-like features we need for the set-theoretic model.

A predicate like *the mouth of this person is wide* can be modeled as a fuzzy predicate whose truth, in the first approximation, is based only on the measure of the width of the mouth.

We have an image I on which we do a number of measurements ϕ_i, which we use to assess the truth value, $\mu_i(\phi)$, of p *fuzzy predicates.* We collect the truth values in a vector

$$\mu(\phi) = \{\mu_1(\phi), \ldots \mu_p(\phi)\} \tag{1}$$

and call $\mu(\phi)$ the (fuzzy) set of *true predicates* on the measurements ϕ. We use this fuzzy set as a basis to extend Tversky's theory.

In order to apply the feature contrast model to the fuzzy sets $\mu(\phi)$ and $\mu(\psi)$ (of the predicates true for the measurements ϕ and ψ) we need to compute the fuzzy sets $\mu(\phi) \cap \mu(\psi)$ and $\mu(\phi) - \mu(\psi)$ (and, by the same definition, $\mu(\psi) - \mu(\phi)$), and to choose a suitable salience function f.

The saliency of the fuzzy set $\mu = \{\mu_1 \dots \mu_p\}$ is given by its cardinality:

$$f(\mu) = \sum_{i=1}^{p} \mu_i \tag{2}$$

The intersection of the sets $\mu(\phi)$ and $\mu(\psi)$ is defined as:

$$\mu_{\cap}(\phi, \psi) = \{\min\{\mu_1(\phi), \mu_1(\psi)\}, \dots \min\{\mu_p(\phi), \mu_p(\psi)\}\}, \tag{3}$$

and the difference between two sets is defined as:

$$\mu_{-}(\phi, \psi) = \{\max\{\mu_1(\phi) - \mu_1(\psi), 0\}, \dots \max\{\mu_p(\phi) - \mu_p(\psi), 0\}\}. \tag{4}$$

With these definitions, we can write the similarity function between two fuzzy sets $\mu(\phi)$ and $\mu(\psi)$—corresponding to measurements made on two images—as:

$$S(\phi, \psi) = \sum_{i=1}^{p} \min\{\mu_i(\phi), \mu_i(\psi)\} - \alpha \max\{\mu_i(\phi) - \mu_i(\psi), 0\} - \beta \max\{\mu_i(\phi) - \mu_i(\psi), 0\} \tag{5}$$

We refer to the model defined by eq. (5) as the *Features Contrast* (FC) model.

As an example of application of these ideas to a restricted domain, consider the determination of similarity of texture images. The samples we used are derived from the Brodatz album [2]. All 112 samples of the album were scanned at 72 dots/inch, and a patch of size 128 × 128 pixels was extracted at a random position from each sample. We based our features on the measurement of the energy in every band of a multi-scale decomposition of the texture image which, in this case, we obtain by a wavelet transform. The most outstanding features used for the recognition of textures are their brightness, their coarseness, and the preferred orientation of the harmonic component [8]. In order to apply our similarity theory, we computed a limited number of predicate-like features from the energy values:

1. The *brightness* (used to support the predicate "the texture is bright") was determined as the average gray level in the lowest frequency band.
2. The *scale* (used to support the predicate "the feature is finely grained") was defined as the band at which the energy of the transform is maximal.
3. The *verticality* (used to support the predicate "the texture has strong vertical component") was determined by taking the energy in the vertical bands and dividing it by the energy in the horizontal bands.
4. The *horizontality* is the reciprocal of the verticality.

Fig. 1 shows the results of a similarity query three textures in the Brodatz album.

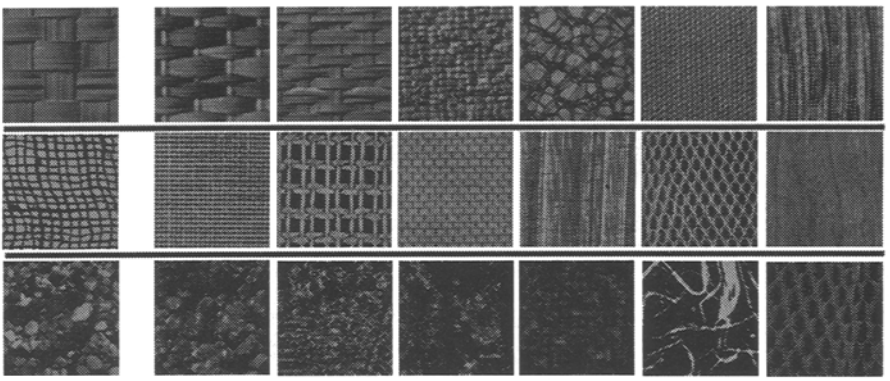

Fig. 1. Similarity results for the textures D64 (hand-woven oriental rattan), D103 (loose burlap), and D54 (beach pebbles) of the Brodatz album.

3.2 Nonlinear Geometric Models

Tversky's approach refuses the concept of similarity as function of a distance and relies instead on set-theoretic considerations. By using fuzzy sets, we have brought the similarity measure back to a function of real-valued features, just like postulated by the metric approach.

We can carry our considerations one step further and actually unify the set-theoretic and the metric approaches [11]. Our first step will be to make a smooth approximation of the max and min operators that we use to compute the Feature Contrast similarity. Notice that we can write

$$\min(x, y) = xu(y - x) + yu(x - y) \tag{6}$$

where u is the step function.

We can substitute the step function with the analytic approximation:

$$u_\omega(x) = \frac{1}{1 + \exp(-\omega x)} \tag{7}$$

obtaining an infinitely derivable function. Note that for $x = 0$ the approximation error is $1/2$ independently of ω, but it is easy to prove the following property:

Lemma 2. *For all $\epsilon > 0$ and $\eta > 0$ there exist an $\omega > 0$ such that, for all $x : |x| > \eta$ it is $|u(x) - u_\omega(x)| < \epsilon$.*

Thus, the approximation error can be made as small as desired outside a region as small as desired around the point 0. Defining the function

$$p_\omega(x, y) = xu_\omega(y - x) + yu_\omega(x - y) \tag{8}$$

we have our approximation. In the same way, we approximate the minimum function with:

$$m_\omega(x,y) = xu_\omega(y-x) + yu_\omega(y-x) \tag{9}$$

With the usual definition of the cardinality of a fuzzy set, we can define the *dissimilarity* between two stimuli as:

$$D(\psi,\phi) = \sum_\lambda \alpha p_\omega(\mu^\lambda(\psi)-\mu^\lambda(\phi),0)+\beta p_\omega(\mu^\lambda(\phi)-\mu^\lambda(\psi),0)-\theta m_\omega(\mu^\lambda(\psi),\mu^\lambda(\phi)) \tag{10}$$

This is a function of the measurements (features) ψ and ϕ on a pair of stimuli and can be considered as a *distance function* for some particular nonlinear Riemann space.

Fig. 2 shows a two dimensional example with $A = \text{diag}\,\{2,2\}$ and $b = [0,0]^T$. The first case is that in which one of the two predicates is "strongly true"

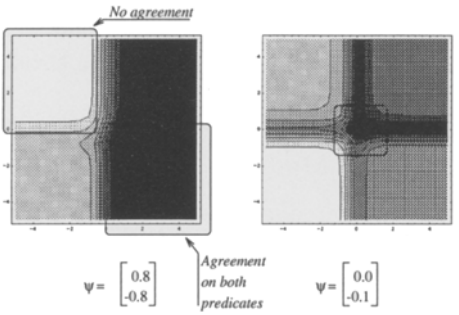

Fig. 2.

($\phi = 0.8$) and the other "strongly false" ($\psi = -0.8$). We see the presence of 4 plateaus, corresponding to agreement in both predicates, agreement in one predicate, and agreement in no predicate. The second figure corresponds to an almost neutral reference stimulus ($\psi = [0, 0.1]^T$). The distance is quite regular in the area inside the central rectangle, then saturates.

Given this distance function, we can study the differential geometric properties of the nonlinear space. Then we can apply standard learning techniques to determine the metric of the space based on the statistics of the images in the database.

Although the geometric formulation was derived from the Feature Contrast Model, there is no reason why we should limit to that. We can be guided by the following observation [7]: not all the similarity judgments follow the same law. Some judgments follow simple metric laws, and these are those associated with global, undecomposable properties of the images. In other words, similarity follows simpler metrics for the lowest frequencies of an image.

This suggests to base our similarity measure on a suitable multi-resolution decomposition of the image. If $I(x)$ is the image, u are the parameters of an image decomposition (e.g. $u = (x, \nu, \theta)$, where ν is the spatial frequency and θ the direction, for the Gabor transform), and $\tilde{I}(u)$ is the transformed image, then a generalized geometric distance between $I_1(x)$ and $I_2(x)$ is

$$\int_u du \int_{\tilde{I}_1(u)}^{\tilde{I}_2(u)} dp \; g(u,p), \tag{11}$$

g being the metric tensor of the decomposition space. This is tantamount to the use of the coefficients of a suitable decomposition (which can be generated by a frame, like the Gabor transform or by a basis like the wavelet transform [1]) as features for similarity measurement.

References

1. J. J. Benedetto. Gabor representations and wavelets. *Contemp. Math.*, 19:9–27, 1989.
2. Phil Brodatz. *Textures. A Photographic Album for Artists & Designers.* Dover, New York, 1966.
3. Myron Flickner, Harpreet Sawhney, Wayne Niblack, Jonathan Ashley, Qian Huang, Byron Dom, Monika Gorkani, Jim Hafner, Denis Lee, Dragutin Petkovic, David Steele, and Peter Yanker. Query by image and video content: the QBIC system. *IEEE Computer*, 1995.
4. J. J. Gibson. *The Ecological Approach to Visual Perception.* Houghton Mifflin, 1979.
5. Ramesh Jain and Amarnath Gupta. Computer vision and visual information retrieval. In *Festschrift for Prof. Azriel Rosenfeld.* IEEE Computer Soc., 1996.
6. B. Julesz. Experiments in the visual perception of texture. *Scientific American*, 232:34–43, April 1975.
7. Carol L Krumhansl. Concerning the applicability of geometric models to similarity data: The interrelationship between similarity and spatial density. *Psychological Review*, 85:445–463., 1978.
8. F. Liu and R. W. Picard. Periodicity, directionality and randomness: Wold features for image modeling and retrieval. Technical Report 320, MIT Media Laboratory Perceptual Computing Section, 1995.
9. Lewis Petrinovich. Probabilistic functionalism; a concept of research method. *American Psychologist*, 34(5):373–390, May 1979.
10. H. Samet and A. Soffer. MARCO: MAp Retrieval by COntent. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(8):783–798, 1996.
11. Simone Santini and Ramesh Jain. Gabor space and the development of preattentive similarity. In *International Conference on Pattern Recognition, Vienna*, 1996. available at http://www-cse.ucsd.edu/users/ssantini.
12. Roger N. Shepard. Toward a universal law of generalization for physical science. *Science*, 237:1317–1323, 1987.
13. Amos Tversky. Features of similarity. *Psychological review*, 84(4):327–352, July 1977.