Evaluation of Geometric & Probabilistic Distance Measures to Retrieve Electron Density Patterns for Protein Structure Determination

Kreshna Gopal[1] **Tod D. Romo**[2] **James C. Sacchettini**[2] **Thomas R. Ioerger**[1]
[1]Department of Computer Science, Texas A&M University
[2]Department of Biochemistry & Biophysics, Texas A&M University
[1]301 Harvey R. Bright Building, College Station, Texas 77843-3112, USA.
[2]103 Bio-Bio Building, College Station, Texas 77843-2128, USA.

Abstract *- Similarity between cases in pattern recognition is typically measured by computing distances between feature vectors. This paper evaluates the effectiveness of various measures of similarity in retrieving good matches in TEXTAL™, a system that uses nearest neighbor learning to retrieve matching 3D patterns of electron density to incrementally determine the structure of proteins by X-ray crystallography. We investigate various geometric measures of similarity, including Euclidean, Manhattan (city-block, or L_1), the generalized Minkowsky metric (L_m) and the Cosine measure. We also experiment with a probabilistic distance metric – a likelihood measure based on the Bayesian classifier. Our experiments in the protein crystallography domain show that the probabilistic measure of similarity outperforms geometric ones significantly. We present a general framework for efficient pattern retrieval from a large database using feature-based matching, and argue that probabilistic and statistical measures of similarity are more robust in noisy, high-dimensional feature spaces representing visual patterns.*

Keywords: pattern recognition, nearest neighbor learning, case-based reasoning, distance measure.

1.0 Introduction

Similarity between images or visual patterns is typically determined by measuring the distance between feature vectors representing these patterns. Since feature spaces can have very large dimensions, non-parametric methods are usually employed to determine the similarity between patterns. The nearest neighbor rule, for instance, is based on the assumption that patterns that are close together (for some appropriate distance measure) in the feature space are similar. Geometric measures, especially the Euclidean distance, are very widely used. But such measures may not effectively simulate human perception of visual patterns [9,8], and do not provide any *a priori* guarantee that they really reflect the similarities and dissimilarities between cases [6]. Thus, other similarity measures need to be explored [1,2].

In this paper, we investigate the effectiveness of various geometric and probabilistic measures of similarity in the context of TEXTAL™, a case-based reasoning system that retrieves matching electron density patterns from a database to determine protein structures by X-ray crystallography. The rest of this paper is organized as follows. In the next section, we describe the protein crystallography domain and give a brief overview of the relevant parts of TEXTAL™. Next, we provide a general framework for efficient pattern recognition. In section 4, we define the various similarity measures used in these experiments, and we empirically compare the effectiveness of different measures of similarity in the ensuing section. These results are discussed in the concluding section.

2.0 X-ray protein crystallography

Proteins are very important macromolecules that perform a wide variety of biological functions. Knowledge of their structures is essential, and X-ray crystallography is the most widely used method for protein structure determination. One of the most difficult steps in protein crystallography is the interpretation of the electron density cloud surrounding the protein i.e. inferring the atomic coordinates given the distribution of density of electrons around the protein (Fig. 1). This is done by crystallographers, and is usually very time-consuming and error-prone, especially since the electron density data collected is typically noisy and blurred (poor resolution). TEXTAL™ uses pattern recognition and case-based reasoning approaches to automate this process i.e. given the electron density data for a protein, the coordinates of all atoms are determined. TEXTAL™ employs a divide-and-conquer approach, where an unsolved structure is decomposed into smaller spherical regions (with a diameter of 5Å, where 1Å = 10^{-10}m) around special carbon atoms known as Cα. To model each region, a database of ~50,000 regions (which are already interpreted) are searched to find one with similar density pattern. Thus, the structure of the protein is incrementally determined by solving each region and combining the solutions. TEXTAL™ is a deployed application used mostly by crystallographers; a full description of the system is beyond the scope of this paper. For more details, refer to [4].

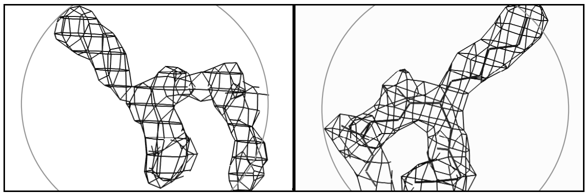

Fig. 1. The pattern on the left represents a spherical region of electron density. An expert crystallographer would recognize the shape (with the help of a 3D visualization program) and model this region of the protein i.e. determine the positions of the atoms, and how they bond together (as shown in the pattern on the right). These two patterns are, in fact, identical but oriented differently in 3D space. Thus, rotation-invariant features are required to represent them.

An expert crystallographer can model an unknown region, drawing from experience on how to visualize 3D density patterns (usually with the help of 3D visualization and model building programs). To automate this process of modeling, we can find the *density correlation* of the unknown region with known ones in the database and return the best match; this objective metric is computationally very expensive, since it involves searching for the optimal rotation between two regions [5]. To make this search more tractable, we filter *k* (500, for instance) cases using a feature-based comparison method, and use the more expensive metric to make the final selection. 76 numeric features were chosen, based on domain knowledge [5]. Feature identification was particularly challenging for various reasons: (1) features have to be rotation-invariant since the regions may occur in any orientation in the database – statistical features like mean, standard deviation, skewness and kurtosis of the density distribution are examples of features independent of 3D orientation; (2) features may be noisy or irrelevant; in TEXTAL™, we use an algorithm called SLIDER [5] to weigh the relevance of features in describing patterns of electron density by comparing how similar features are for pairs of matching regions relative to pairs of mismatching regions; (3) relevance of features can be context sensitive [3] i.e. the relevance depend on where we are in the feature space; (4) features may interact, where the relevance of a feature is weak on an individual basis, but their relationship to the pattern is evident only when looked at in combination [4]. These issues make retrieval a challenging task, and the effectiveness of the database search hinges on an appropriate choice of the feature-based distance metric.

3.0 Efficient pattern retrieval

TEXTAL™ is based on the following general model, which is applicable to and potentially useful for many real applications: we are given a pattern (like an image), and we are asked to classify it or, more generally, find similar patterns using cases from a (typically) large database. Given a query case q, and a database of N cases, our goal can be met if we could rank the N cases on their distances to q, by using an appropriate distance function. But the optimal distance metric is often computationally too expensive, especially if the database is large. The approach that we propose is to quickly filter *k* approximate matches (relative to the objective metric) by representing all cases by a vector of features, and using a fast method of finding similarity (like the Euclidean measure) that we can afford to run over the whole database. These *k* cases can be further examined by computationally expensive methods to make the final selection. The effectiveness of the method depends on the choice of good (and appropriately weighted) features, a good understanding of how approximate the feature-based distance method is (which will help choose a suitable *k*) and the database size. In this paper, we focus of this issue of assessing various similarity metrics in this framework.

Very often, we may not require the retrieval of the very best match, but be complacent with reasonably good ones. We formalize this notion of "good enough" matches as follows. Given an objective (correct) measure of similarity (called *sim*), a query case q, and a tolerance δ, the objectively best λ matches from the database are deemed to be good enough if:

$$[sim(q,m_1) - sim(q,m_\lambda)]/\ sim(q,m_1) \ < \ \delta$$

where m_i is the i^{th} best match according to *sim* (*sim* > 0). Therefore, we essentially need to choose a distance measure that is most effective in getting any one of the λ matches in the top *k*.

4.0 Definitions of similarity metrics

4.1 Geometric measures of similarity

Let $L_n(x,y) = [\Sigma w_j|x_j - y_j|^n]^{1/n}$, $1 \leq j \leq F$, where x, y are F-dimensional feature vectors, and w_j is a measure of the relevance of feature j. If n = 1, we get the Manhattan distance (also known as city-block or histogram intersection). If n = 2, the distance is known as Euclidean. The generalized form, L_n is known as the Minkowsky distance. The cosine distance between vectors x and y = 1 - x.y/|x||y|.

4.2 Probabilistic measure of similarity

This measure proposed by [1] essentially evaluates the likelihood that two patterns are similar, and use this likelihood to rank all patterns in the database in terms of how similar they are to a query pattern. First a class S of similar pairs of patterns, and a class D of different pairs of patterns are defined. In TEXTAL™, some pairs of regions which have a density correlation greater (less) than a threshold are randomly chosen to represent S (D). Given patterns x and y, we compute the difference vector, d = x – y. The posterior probabilities that x & y are similar and different are respectively given by:

$$P(S|d) = P(d|S)\, P(S)/P(d)$$
$$\& \; P(D|d) = P(d|D)\, P(D)/P(d)$$

The likelihood that x & y are similar can be defined as $r(d) = P(S|d)/P(D|d)$. Assuming that is it equally probable for x & y to be different and similar [i.e. $P(S) = P(D)$], we get:

$$r(d) = P(d|S)/P(d|D)$$

We also assume that the feature differences in S and D have a multivariate normal density with mean μ_S and μ_D respectively, and covariance matrices Σ_S and Σ_D respectively i.e.

$$f(d|\mu_S,\Sigma_S) = (2\pi)^{-F/2} |\Sigma_S|^{-1/2} \exp[-(d-\mu_S)^T \Sigma_S^{-1}(d-\mu_S)/2]$$

$$f(d|\mu_D,\Sigma_D) = (2\pi)^{-F/2} |\Sigma_D|^{-1/2} \exp[-(d-\mu_D)^T \Sigma_D^{-1}(d-\mu_D)/2]$$

$$\text{Now, } r(d) = P(d|S)/P(d|D)$$
$$= f(d|\mu_S,\Sigma_S) / f(d|\mu_D,\Sigma_D)$$

By taking log and eliminating constants, r(d) can be re-written as:

$$r(d) = (d-\mu_D)^T \Sigma_D^{-1}(d-\mu_D) - (d-\mu_S)^T \Sigma_S^{-1}(d-\mu_S)$$

Given a query pattern q, the likelihood ratio can be computed for all $d_i = q - x_i$, and used to rank all the patterns in the database in terms of how similar they are to q, where the higher the ratio, the more similar the pattern.

5.0 Results

A test set of 200 query regions of electron density patterns were generated, and for each query region, a database of ~50,000 regions were exhaustively searched and ranked according to an objective similarity measure (density correlation, which ranges from 0 to 1, the latter being the score for a perfect match). The various feature-based geometric and probabilistic similarity metrics were then used to rank all regions in the database for each query region (as an approximation to the objective ranking method). As discussed earlier, the aim of the approximate ranking systems need not be a perfect ranking of all regions in the database, but more modestly, the ability to rank a "good enough" match within the top *k*. As described in section 3, a good enough match is any one of the best λ matches (according to the objective metric), which depends on a tolerance δ. Table 1 shows how mean λ (over the 200 test cases) varies with δ. Fig. 2 displays how the first good enough match retrieved by different measures is actually ranked according to the objective metric (where rank = 1 means the retrieved match is actually the best match i.e. the lower the rank, the better the similarity measure). The average (over 200 test cases) of these ranks is plotted against varying tolerance. It can be noticed that the mean rank of the probabilistic measure is roughly half that of the Euclidean.

Table 1. λ increases exponentially with δ.

Tolerance, δ	**Mean λ**
0.01	7.2
0.02	24.5
0.03	57.5
0.04	106.9

Fig. 3 shows the percentage of times a good enough match (i.e. one which is in the first top λ matches according to the objective metric) is obtained within the top *k* of different similarity measures for various values *k* (given tolerance = 0.01, which is reasonably conservative in this domain).

6.0 Conclusion & discussion

We proposed a general framework for efficient case retrieval from a large database based on the nearest neighbor rule, where we

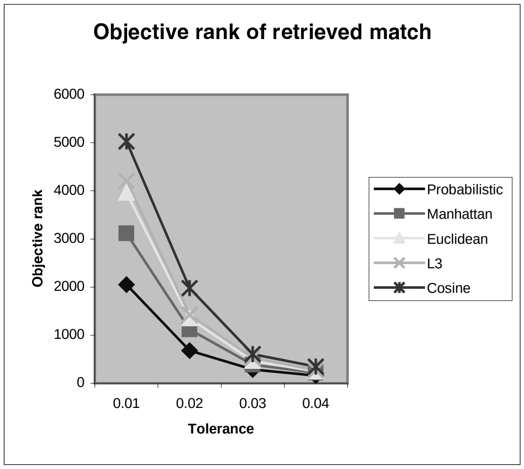

Fig. 2. The objective rank (according to the objective metric) vs tolerance.

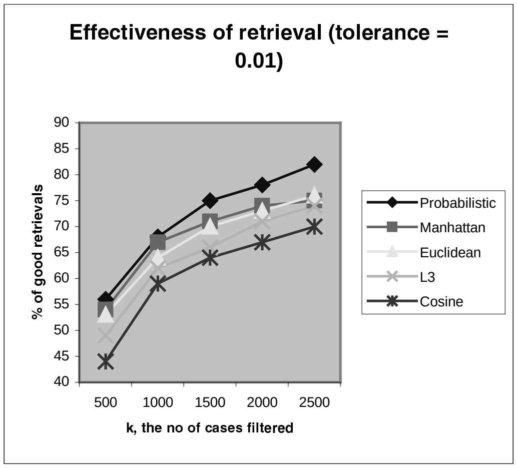

Fig 3. The percentage of times a "good enough" match is retrieved.

use an efficient but approximate feature-based measure of similarity to filter some good cases which can be further examined by a more expensive objective measure for a final selection. In this paper, we specifically analyze how effective various geometric and probabilistic measures of similarity are in filtering good enough matches. These experiments were done in the protein crystallography domain, where the 3D structures of proteins are automatically determined through the recognition and retrieval of 3D patterns of electron density.

We made the following observations: i) The probabilistic measure outperforms geometric measures significantly. The former could retrieve good matches up to 10% more often than the Euclidean measure; ii) Manhattan (L_1) performed better than the Euclidean (L_2), which performed better than L_3. The Cosine method performed relatively the worst; iii) The probabilistic metric was particularly successful in cases which are difficult to recognize (i.e. where the geometric measures performed very poorly) due to factors like poor quality of the pattern (poor resolution, noisy data) and irrelevant features. [1] reports results very similar to i) and ii) in the image retrieval domain. The superiority of probabilistic and statistical distance measures is also highlighted in [7, 6].

We argue that although all similarity metrics use the same features as input, geometric measures provide a more parametric method of matching, and is sensitive to the biases of the underlying representation of similarity. On the other hand, the probabilistic measure is derived more directly from what is objectively similar/different (through the classes S & D) and effectively contain more information (in μ_S, Σ_S, μ_D, Σ_D) than what the representation for similarity (including feature weights) provide in geometric measures. The density estimates P(S|d) and P(D|d) explicitly capture the type of pattern variations which are critical in the formulation of a good measure of similarity. Furthermore, geometric metrics are more sensitive to how features are processed: normalization [1], scaling and weighting. These reasons probably explain why the probabilistic measure is more robust as compared to geometric ones, especially in high-dimensional, noisy domains (incorrect data and irrelevant features).

A number of related ideas can be investigated to further enhance the accuracy of retrieval of electron density patterns for TEXTAL^{TM}: aggregation of various rankings and other methods to combine different measures of similarity; runtime creation of a distance measure [2]. The probabilistic measure can be improved by a more careful choice of the similar and different classes such that they contain more relevant, domain-specific information to enable discriminate between what is really similar and different.

7.0 References

[1] S. Aksoy and R.M. Haralick, "Probabilistic vs. Geometric Similarity Measures for Image Retrieval", *Proceedings of Computer Vision and Pattern Recognition (CPRV)*, pp. 112-128, 2001.

[2] A. Berman and L. Shapiro, "Efficient Image Retrieval with Multiple Distance Measures", *Proc. of SPIE/IS&T Conf. on Storage and Retrieval for Image and Video Databases V*, vol. 3022, San Jose, CA, 1997.

[3] P. Domingos, "Context-Sensitive Feature Selection for Lazy Learners", *Artificial Intelligence Review*, 11, pp. 227-253, 1997.

[4] T.R. Ioerger and J.C. Sacchettini, "The TEXTAL system: Artificial Intelligence Techniques for Automated Protein Model-Building", In R.M. Sweet and C.W. Carter, eds., *Methods in Enzymology*, 374: 244-270, 2003.

[5] T.R. Holton, J.A. Christopher, T.R. Ioerger, and J.C. Sacchettini, "Determining Protein Structure from Electron Density Map Using Pattern Recognition", *Acta. Cryst.*, D46, pp. 722-734, 2000.

[6] P. Kontkanen, P. Myllymaki, T. Silander, and H. Tirri, "A Bayesian approach for retrieving relevant cases", In P. Smith, editor, *Artificial Intelligence Applications* (Proceedings of the EXPERSYS-97 Conference), Sunderland, UK, pp. 67-72, 1997.

[7] B. Moghaddam, C. Nastar, and A. Pentland, "A Bayesian Similarity Measure for Direct Image Matching", M.I.T. Media Laboratory Perceptual Computing Section Technical Report No. 393, 1996.

[8] E.B. Rogowitz, T. Frese, J. Smith, A.C. Bouman, and E. Kalin, "Perceptual Image Similarity Experiments", *Proceedings of the IS&T/SPIE Conference on Human Vision and Electronic Imageing III*, 1998.

[9] Y. Rui, T.S. Huang, and F. Chang, "Image Retrieval: Past, Present, and Future", *Journal of Visual Communication and Image Representation*, 1999.