3D Shape Histograms for Similarity Search and Classification in Spatial Databases

Mihael Ankerst, Gabi Kastenmüller, Hans-Peter Kriegel, Thomas Seidl

University of Munich, Institute for Computer Science
Oettingenstr. 67, 80538 Munich, Germany
http://www.dbs.informatik.uni-muenchen.de
{ankerst, kastenmu, kriegel, seidl}@dbs.informatik.uni-muenchen.de

Abstract. Classification is one of the basic tasks of data mining in modern database applications including molecular biology, astronomy, mechanical engineering, medical imaging or meteorology. The underlying models have to consider spatial properties such as shape or extension as well as thematic attributes. We introduce 3D shape histograms as an intuitive and powerful similarity model for 3D objects. Particular flexibility is provided by using quadratic form distance functions in order to account for errors of measurement, sampling, and numerical rounding that all may result in small displacements and rotations of shapes. For query processing, a general filter-refinement architecture is employed that efficiently supports similarity search based on quadratic forms. An experimental evaluation in the context of molecular biology demonstrates both, the high classification accuracy of more than 90% and the good performance of the approach.

Keywords. 3D Shape Similarity Search, Quadratic Form Distance Functions, Spatial Data Mining, Nearest Neighbor Classification

1 Introduction

Along with clustering, mining association rules, characterization and generalization, classification is one of the fundamental tasks in data mining [CHY 96]. Given a set of classes and a query object, the problem is to assign an appropriate class to the query object based on its attribute values. Many modern database applications including molecular biology, astronomy, mechanical engineering, medical imaging, meteorology and others are faced with this problem. When new objects are discovered through remote sensing, new tumors are detected from X-ray images, or new molecular 3D structures are determined by crystallography or NMR techniques, an important question is to which class the new object belongs. Further steps to deeper investigations may be guided by the class information: a prediction of primary and secondary effects of drugs could be tried, the multitude of mechanical parts could be reduced, etc.

As a basis for any classification technique, an appropriate model has to be provided. Classes represent collections of objects that have characteristic properties in common and thus are similar, whereas different classes contain objects that have more or less strong dissimilarities. In all of the mentioned applications, the geometric shape of the

R.H. Güting, D. Papadias, F. Lochovsky (Eds.): SSD'99, LNCS 1651, pp. 207-226, 1999.

objects is an important similarity criterion. Along with the geometry, also thematic attributes such as physical and chemical properties have an influence on the similarity of objects.

Data from real world applications inherently suffer from errors, beginning with errors of measurement, calibration, sampling errors, numerical rounding errors, displacements of reference frames, and small shifts as well as rotations of the entire object or even of local details of the shapes. Though no full invariance against rotations is generally required, if the objects are already provided in a standardized orientation, these errors have to be taken into account. In this paper, we introduce a flexible similarity model that considers these problems of local inaccuracies and may be adapted by the users to their specific requirements or individual preferences.

The paper is organized as follows: The remainder of this introduction surveys related work from molecular biology, data mining, and similarity search in spatial databases. In Section 2, we introduce the components of our similarity model: 3D shape histograms for object representation, and a flexible similarity distance function. Due to the large and rapidly increasing size of current databases, the performance of query processing is an important task and, therefore, we introduce an efficient multistep system architecture in Section 3. In Section 4, we present the experimental results concerning the effectiveness and efficiency of our technique in the context of molecular biology. Section 5 concludes the paper.

1.1 Classification in Molecular Databases

A major issue in biomolecular databases is to get a survey of the objects, and thus a basic task is classification: To which of the recognized classes in the database does a new molecule belong? In molecular biology, there are already classification schemata available. In many systems, classifying new objects when inserting them into the database requires supervision by experts that are very experienced and have a deep knowledge of the domain of molecular biology. What is desired is an efficient classification algorithm that may act as a fast filter for further investigation and that may be restricted e.g. to geometric aspects.

A sophisticated classification is available from the FSSP database (Families of Structurally Similar Proteins), generated by the Dali system [HS 94] [HS 98]. The similarity of two proteins is based on their secondary structure, that is substructures of the molecules such as alpha helices or beta sheets. The evaluation of a pair of proteins is very expensive, and query processing for a single molecule against the entire database currently takes an overnight run on a workstation.

Another classification schema is provided by CATH [OMJ+ 97], a hierarchical classification of protein domain structures, which clusters proteins at four major levels, class (C), architecture (A), topology (T) and homologous superfamily (H). The class label is derived from secondary structure content and cannot be assigned for all protein structures automatically. The architecture label, which describes the gross orientation of secondary structures, independent of connectivities, is currently assigned manually. The assignments of structures to topology families and homologous superfamilies are made by sequence and structure comparisons.

1.2 Nearest-Neighbor Classification

A lot of research has been performed in the area of classification algorithms; surveys are presented in [WK 91] [MST 94] [Mit 97]. All the methods require that a training set of objects is given for which both the attribute values and the correct classes are known a-priori. Based on this knowledge of previously classified objects, a classifier predicts the unknown class of a new object. The quality of a classifier is typically measured by the classification accuracy, i.e. by the percentage of objects for which the class label is correctly predicted.

Many methods of classification generate a description for the members of each class, for example by using bounding boxes, and assign a class to an object if the object matches the description of the class. Nearest neighbor classifiers, on the other hand, refrain from discovering a possibly complex description of the classes. As their name indicates, they retrieve the nearest neighbor p of a query object q and return the class label of p in order to predict the class label of q. Obviously, the definition of an appropriate distance function is crucial for the effectiveness of nearest neighbor classification. In a more general form, called k-nearest neighbor classification, k nearest neighbors of the query object q are used to determine the class of q. Thus, the effectiveness depends on the number k as well as on the weighting of the k neighbors. Both, appropriate similarity models as well as efficient algorithms for similarity search are required for successful nearest neighbor classification.

1.3 Geometry-Based Similarity Search

Considerable work on shape similarity search in spatial database systems has been performed in recent years. As a common technique, the spatial objects are transformed into high-dimensional feature vectors, and similarity is measured in terms of vicinity in the feature space. The points in the feature space are managed by a multi-dimensional index. Many of the approaches only deal with two-dimensional objects such as digital images or polygonal data and do not support 3D shapes.

Let us first survey previous 2D approaches from the literature. In [GM 93], a shape is represented by an ordered set of surface points, and fixed-sized subsets of this representation are extracted as shape features. This approach supports invariance with respect to translation, rotation and scaling, and is able to deal with partially occluded objects. The technique of [BKK 97] applies the Fourier transform in order to encode sections of polygonal outlines of 2D objects; even partial similarity is supported. Both methods exploit a linearization of polygon boundaries and, therefore, are hard to extend to 3D objects. In [Jag 91], shapes are approximated by rectangular coverings. The rectangles of a single object are sorted by size, and the largest ones are used for the similarity retrieval. The method of [KSF+ 96] is based on mathematical morphology and uses the max morphological distance and max granulometric distance of shapes. It has been applied to 2D tumor shapes in medical image databases. A 2D technique that is related to our 3D shape histograms is the Section Coding technique [Ber 97] [BK 97] [BKK 97a]. For each polygon, the circumscribing circle is decomposed into a given number of sectors, and for each sector, the area of the polygon inside of this sector divided by the total area of the polygon is determined. Similarity is defined in terms of the Euclidean dis-

tance of the resulting feature vectors. The similarity model in [AKS 98] handles 2D shapes in pixel images and provides a solution for the problem of small displacements.

The QBIC (Querying By Image Content) system [FBF+ 94] [HSE+ 95] contains a component for 2D shape retrieval where shapes are given as sets of points. The method is based on algebraic moment invariants and is also applicable to 3D objects [TC 91]. As an important advantage, the invariance of the feature vectors with respect to rigid transformations (translations and rotations) is inherently given. However, the adjustability of the method to specific applications is restricted. From the available moment invariants, appropriate ones have to be selected, and their weighting factors may be modified. Whereas the moment invariants are abstract quantities, the shape histograms presented in this paper are more intuitive and may be graphically visualized, thus providing an impression of the suitability for specific applications.

The Geometric Hashing paradigm for model-based 3D object recognition was introduced by [LW 88]. The objects are represented by sets of points; from these points, non-collinear triplets are selected to represent different orientations of a single object. For each of these orientations, every point of an object is stored in a hash table that maps 3D points to objects and their orientations. The query processing heuristic requires a certain threshold provided by the user. This threshold has a substantial impact on the effectiveness of the technique and, thus, an appropriate choice is crucial. If the threshold is too high, no answer is reported; if the threshold is too low, however, there is no guarantee and, moreover, no feedback whether the best matching object with respect to the underlying similarity model is returned. In contrast to that, the *k*-nearest neighbor algorithm used in our approach ensures that the *k* most similar objects are returned. There are no objects in the database which are more similar than the retrieved ones.

The approximation-based similarity model presented in [KSS 97] and [KS 98] addresses the retrieval of similar 3D surface segments. These surface segments occur in the context of molecular docking prediction where they represent potential docking sites. Since the segments are designed to model local portions of 3D surfaces but not to model the entire contour of a 3D solid, this technique is not applicable for searching 3D solids having similar global shapes.

1.4 Invariance Properties of Similarity Models

All the mentioned similarity models incorporate invariance against translation of the objects, some of them also include invariance against scaling which is not necessarily desired in the context of molecular or CAD databases. With respect to invariance against rotations, two approaches can be observed. Some of the similarity models inherently support rotational invariance, e.g. by means of the Fourier transform [BKK 97] or the algebraic moment invariants [TC 91]. Most of the techniques, however, include a preprocessing step that rotates the objects to a normalized orientation, e.g. by a Principal Axis Transform. If rotations should be considered nevertheless, the objects may be rotated artificially by certain angles as suggested in [Ber 97]. For some applications, eventually, rotational invariance may be not required, e.g. if mechanical parts in a CAD database are already stored in a standardized orientation.

An important kind of invariance has not yet be considered in previous work, the robustness of similarity models against errors of measurement, calibration, sampling

errors, errors of classification of object components, numerical rounding errors, and small displacements such as shifts or slight rotations of geometric details. In our model, these problems are addressed and may be controlled by the user by specifying and adapting a similarity matrix for histogram bins.

2 A 3D Shape Similarity Model

In this section, we introduce our 3D shape similarity model by defining the two major ingredients: First, the shape histograms as an intuitive and discrete representation of complex spatial objects. Second, an adaptable similarity distance function for the shape histograms that may take small shifts and rotations into account by using quadratic forms.

2.1 Shape Histograms

The definition of an appropriate distance function is crucial for the effectiveness of any nearest neighbor classifier. A common approach for similarity models is based on the paradigm of feature vectors. A *feature transform* maps a complex object onto a feature vector in a multidimensional space. The similarity of two objects is then defined as the vicinity of their feature vectors in the feature space.

We follow this approach by introducing 3D shape histograms as intuitive feature vectors. In general, histograms are based on a partitioning of the space in which the objects reside, i.e. a complete and disjoint decomposition into cells which correspond to the bins of the histograms. The space may be geometric (2D, 3D), thematic (e.g. physical or chemical properties), or temporal (modeling the behavior of objects).

We suggest three techniques for decomposing the space: A shell model, a sector model, and a spiderweb model as the combination of the former two (cf. Figure 1). In a preprocessing step, a 3D solid is moved to the origin. Thus the models are aligned to the center of mass of the solid.

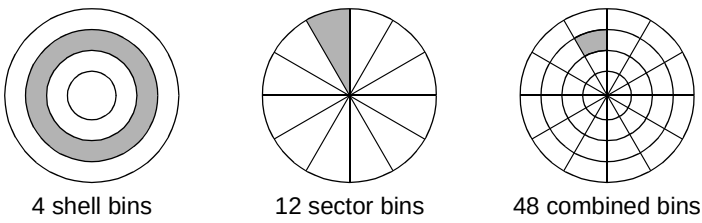

Figure 1. Shells and sectors as basic space decompositions for shape histograms. In each of the 2D examples, a single bin is marked

Shell Model. The 3D is decomposed into concentric shells around the center point. This representation is particularly independent from a rotation of the objects, i.e. any rotation of an object around the center point of the model results in the same histogram. The radii of the shells are determined from the extensions of the objects in the database. The outermost shell is left unbound in order to cover objects that exceed the size of the largest known object.

Sector Model. The 3D is decomposed into sectors that emerge from the center point of the model. This approach is closely related to the 2D section coding method [BKK 97a]. However, the definition and computation of 3D sector histograms is more sophisticated, and we define the sectors as follows: Distribute the desired number of points uniformly on the surface of a sphere. For this purpose, we use the vertices of regular polyhedrons and their recursive refinements. Once the points are distributed, the Voronoi diagram of the points immediately defines an appropriate decomposition of the space. Since the points are regularly distributed on the sphere, the Voronoi cells meet at the center point of the model. For the computation of sector-based shape histograms, we need not to materialize the complex Voronoi diagram but simply apply a nearest neighbor search in 3D since typical number of sectors are not very large.

Combined Model. The combined model represents more detailed information than pure shell models and pure sector models. A simple combination of two fine-grained 3D decompositions results in a high dimensionality. However, since the resolution of the space decomposition is a parameter in any case, the number of dimensions may easily be adapted to the particular application.

In Figure 2, we illustrate various shape histograms for the example protein, 1SER-B, which is depicted on the left of the figure. In the middle, the various space decompositions are indicated schematically, and on the right, the corresponding shape histograms are depicted. The top histogram is purely based on shell bins, and the bottom histogram is defined by 122 sector bins. The histograms in the middle follow the combined model, they are defined by 20 shell bins and 6 sector bins, and by 6 shell bins and 20 sector bins, respectively. In this example, all the different histograms have approximately the same dimension of around 120. Note that the histograms are not built from volume elements but from uniformly distributed surface points taken from the molecular surfaces.

2.2 Shortcomings of the Euclidean Distance

In order to quantify the dissimilarity of objects, an appropriate distance function of feature vectors has to be provided. An obvious solution is to employ the classic Euclidean distance function which is well-defined for feature spaces of arbitrary dimension. In a squared representation, the Euclidean distance of two N-dimensional vectors p and q is defined as:

$$d^2_{\text{euclid}}(p, q) = \sum_{i=1}^{N} (p_i - q_i)^2 = (p - q) \cdot (p - q)^{\mathrm{T}}.$$

However, the Euclidean distance exhibits severe limitations with respect to similarity measurement. In particular, the individual components of the feature vectors which correspond to the dimensions of the feature space are assumed to be independent from each other, and no relationships of the components such as substitutability and compensability may be regarded. The following example demonstrates these shortcomings in more detail.

Let us consider the three objects *a*, *b*, and *c* from Figure 3. From a visual inspection, we assess the objects *a* and *b* to be more similar than *a* and *c* or *b* and *c* since the two characteristic peaks are located more close together in the objects *a* and *b* than in object *c*. However, the peaks of *a* and *b* do not overlap the same sectors and, therefore, are mapped to distinct histogram bins. The Euclidean distance neglects any relationship of

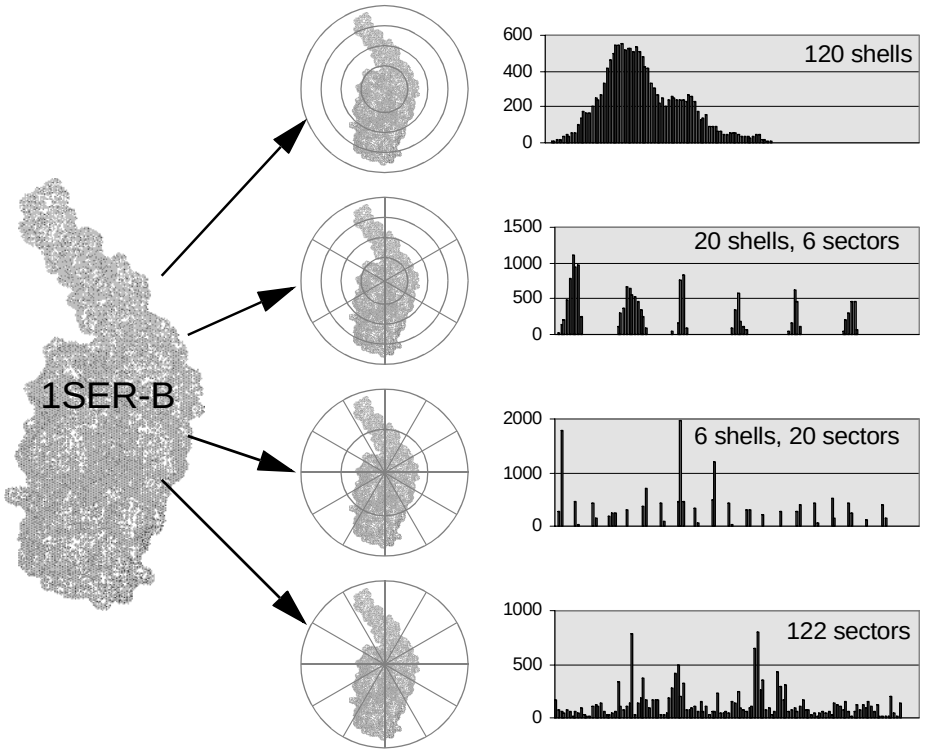

Figure 2. Several 3-D shape histograms of the example protein 1SER-B. From top to bottom, the number of shells decreases and the number of sectors increases

the vector components and does not reflect the close similarity of *a* and *b* in comparison to *c*. Thus, the three objects count for being equally similar, because their feature vectors have the same distance in pairs.

2.3 Quadratic Form Distance Functions

An approach to overcome these limitations has been investigated for color histograms in the QBIC project (Query by Image Content) at IBM Almaden [FBF+ 94] [HSE+ 95]. The authors suggest to use quadratic form distance functions which have also been successfully applied to several multimedia database applications [Sei 97] [SK 97] [KSS 97] [AKS 98] [KS 98]. A quadratic form distance function is defined in terms of a similarity matrix *A* as follows where the components a_{ij} of the matrix *A* represent the similarity of the components *i* and *j* in the underlying vector space.

$$d_A^2(x, y) = (x - y) \cdot A \cdot (x - y)^T = \sum_{i=1}^{N} \sum_{j=1}^{N} a_{ij}(x_i - y_i)(x_j - y_j).$$

From this definition, it becomes clear that the standard Euclidean distance is a special case of the quadratic form distance which is achieved by using the identity matrix *Id*

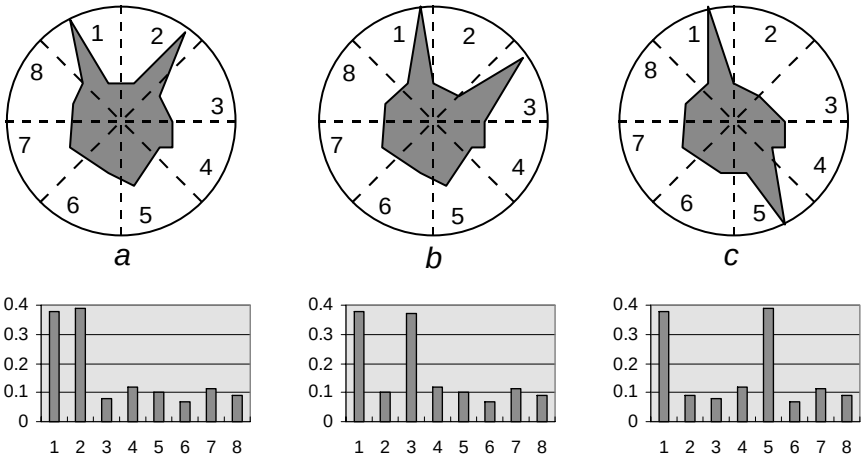

Figure 3. Shortcomings of the Euclidean distance. The Euclidean distance of the shape histograms does not reflect the similarity that is due to the proximity of neighboring sectors

as similarity matrix. Analogously, we obtain a weighted Euclidean distance function that has the weights $(w_1, w_2, \ldots, w_n)$ by using the diagonal matrix $\text{diag}(w_1, w_2, \ldots, w_n)$ as similarity matrix. In both cases, the non-diagonal components are set to zero which exactly corresponds to the fact that no cross-similarities of the dimensions are assumed.

The Euclidean distance of two vectors p and q is totally determined, there is no parameter which may be tuned. The weighted Euclidean distance is a little more flexible because it controls the effect of each vector component onto the overall distance by specifying individual weights for the dimensions. A new level of flexibility is supported by the general quadratic form distance function. On top of specifying the effect of individual dimensions onto the overall distance, cross-dependencies of the dimensions may be handled.

By using a quadratic form distance function as an adaptable similarity function, the problems of the Euclidean distance may be overcome. The neighborhood of bins in general and of shells or sectors in particular may be represented as similarity weights in the similarity matrix A. The individual similarity weights depend on the distances of the corresponding bins. Let us denote by $d(i, j)$ the distance of the cells that correspond to the bins i and j. For shells, we define the bin distance to be the difference of the corresponding shell radii, and for sectors, we use the angle between the sector centers as bin distances. When provided with an appropriate bin distance function, we compute the corresponding similarity weights by an adapted formula from [HSE+ 95] as follows:

$$a_{ij} = e^{-\sigma \cdot d(i,j)}.$$

The parameter σ controls the global shape of the similarity matrix. The higher σ, the more similar is the resulting matrix to the identity matrix. In any case, a high value of σ yields the matrix to be diagonally dominant. We observed good results for σ between 1.0 and 10.

2.4 Invariance Properties of the Models

In general, the 3D objects are located anywhere in the 3D, and their orientation as well as their size can vary arbitrarily. For defining meaningful and applicable similarity models, we have to provide invariance for translations, scaling and rotation, depending on the application. We can ensure these invariances in three ways, by a preprocessed normalization step, by the similarity model itself or by both steps.

In a normalization step, we perform translation and rotation of all objects. After the translation which maps the center of mass of each object onto the origin, we perform a Principal Axes Transform on the object. The computation for a set of 3D points starts with the 3×3-covariance matrix where the entries are determined by an iteration over the coordinates (x, y, z) of all points:

$$\begin{bmatrix} \sum x^2 & \sum xy & \sum xz \\ \sum xy & \sum y^2 & \sum yz \\ \sum xz & \sum yz & \sum z^2 \end{bmatrix}.$$

The eigenvectors of this covariance matrix represent the principal axes of the original 3D point set, and the eigenvalues indicate the variance of the points in the respective direction. As a result of the Principal Axes Transform, all the covariances of the transformed coordinates vanish. Although this method in general leads to a unique orientation of the objects, this does not hold for the exceptional case of an object with at least two variances having the same value. In our experiments using the protein database, we almost never observed such cases and, therefore, assume a unique orientation of the objects.

The similarity models themselves have inherent invariance properties. Obviously, the sector model is invariant against scaling, whereas the shell model trivially has rotational invariance. Often, no full invariance is desired, instead just small displacement, shifts or rotations of geometric details occur in the data, for example caused by errors of measurement, sampling or numerical rounding errors. This variation of invariance precision which is highly application- and user-dependent is supported by the user-defined similarity matrix modeling the appropriate similarity weight for each pair of bins.

2.5 Extensibility of Histogram Models

What we have discussed so far is a very flexible and intuitive similarity model for 3D objects. However, the distance function of the similarity model is based just on the spatial attributes of the objects. Frequently, on top of the geometric information, a lot of thematic information is used to describe spatial objects. Particularly in protein databases, the chemical structure and physical properties are important. Examples include atom types, residue types, partial charge, hydrophobicity, electrostatic potential among others. A general approach to manage thematic information along with spatial properties is provided by combined histograms. Figure 4 demonstrates the basic principle. Assume we are given a spatial histogram structure as presented above, and additionally a thematic histogram structure to be given. A combined histogram structure is immediately obtained as the Cartesian product of the original structures.

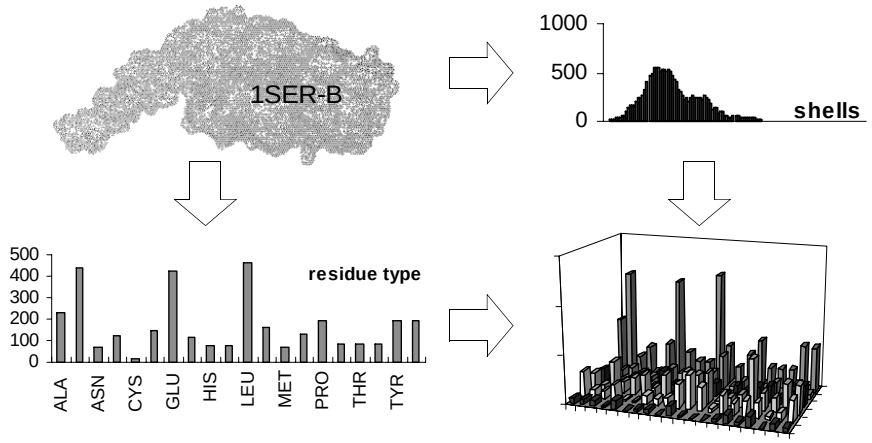

Figure 4. Example for a combined thematic and shape histogram for a molecule

Obviously, this product based approach leads to a tradeoff between a more powerful modeling versus a very high dimensionality. An investigation of the efficiency and effectiveness as well as the development of new techniques that meet the requirements of ultra high dimensional spaces is part of our future research plans.

3 Efficient Query Processing

Due to the enormous and still increasing size of modern databases that contain tens and hundreds of thousands of molecules, mechanical parts, or medical images, the task of efficient query processing becomes more and more important. In the case of quadratic form distance functions, the evaluation time of a single database object increases quadratically with the dimension. We measured 0.23 milliseconds in the average for 21D histograms, 6.2 milliseconds for 256D and 1,656 milliseconds in 4,096D space (cf. Figure 5). Thus, linearly scanning the overall database is prohibitive. In order to achieve a good performance, our system architecture follows the paradigm of multistep query processing: An index-based filter step produces a set of candidates, and a subsequent refinement step performs the expensive exact evaluation of the candidates [Sei 97] [AKS 98].

3.1 Optimal Multistep *k*-Nearest Neighbor Search

Whereas the refinement step in a multistep query processor has to ensure the correctness, i.e. no false hits may be reported as final answers, the filter step is primarily responsible for the completeness, i.e. no actual result may be missing from the final answers and, therefore, from the set of candidates. Figure 6 illustrates the architecture of our multistep similarity query processor that fulfills this property [SK 98]. Moreover, as an advantage over the related method of [KSF+ 96], our algorithm is proven to be optimal, i.e. it produces only the minimum number of candidates. Thus, expensive evaluations of unnecessary candidates are avoided, and we observed improvement factors of up to 120 for the number of candidates and 48 for the overall runtime.

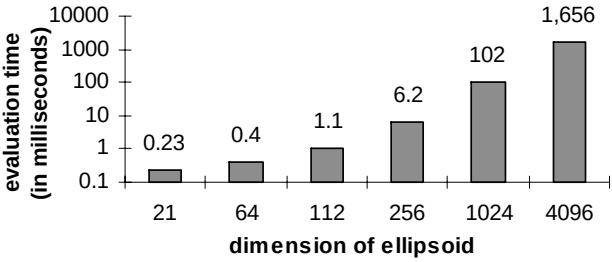

Figure 5. Average evaluation time for single ellipsoid distances

Based on a multidimensional index structure, the filter step performs an incremental ranking that reports the objects ordered by their increasing filter distance to the query object using an algorithm derived from [HS 95]. The number of accessed index pages is minimum as proven in [BBKK 97], and the termination is controlled by the refinement step in order to guarantee the minimum number of candidates [SK 98]. Only for the exact evaluation in the refinement step, the exact object representation is retrieved from the object server.

In order to guarantee no false dismissals caused by the filter step, the filter distance function d_f has to be a lower bound of the exact object distance function d_o that is evaluated in the refinement step. That is, for all database objects p and all query objects q, the following inequality has to hold:

$$d_f(p, q) \leq d_o(p, q) .$$

3.2 Reduction of Dimensionality for Quadratic Forms

A common approach to manage objects in high-dimensional spaces is to apply techniques to reduce the dimensionality. The objects in the reduced space are then typically managed by any multidimensional index structure [GG 98]. The typical use of common linear reduction techniques such as the Principal Components Analysis (PCA) or Karhunen-Loève Transform (KLT), the Discrete Fourier or Cosine Transform (DFT, DCT), the Similarity Matrix Decomposition [HSE+ 95] or the Feature Subselection [FBF+ 94] includes a clipping of the high-dimensional vectors such that the Euclidean distance in the reduced space is always a lower bound of the Euclidean distance in the high-dimensional space.

The question arises whether these approved techniques are applicable to general quadratic form distance functions. Fortunately, the answer is positive; an algorithm to reduce the similarity matrix from a high-dimensional space down to a low-dimensional space according to a given reduction technique was developed in the context of multimedia databases for color histograms [SK 97] and shapes in 2D images [AKS 98]. The method guarantees three important properties: First, the reduced distance function is a lower bound of the given high-dimensional distance function. Obviously, this criterion had to be a design goal in order to meet the requirements of multistep similarity query processing. Second, the reduced distance function again is a quadratic form and, there-

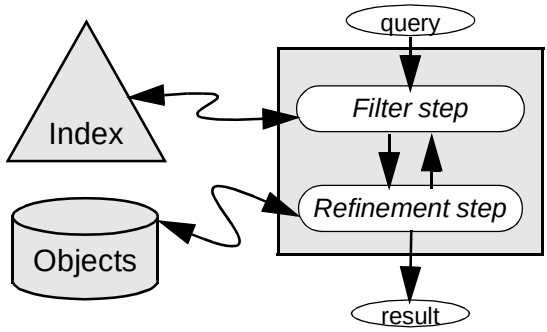

Figure 6. Multistep similarity query processing

fore, the complexity of the query model is not increased while decreasing the dimension of the space. Third, the reduced distance function is the greatest of all lower-bounding distance functions in the reduced space. As an important implication of this property, the selectivity in the filter step is optimal: In the reduced space, no lower-bounding distance function is able to produce a smaller set of candidates than the resulting quadratic form.

3.3 Ellipsoid Queries on Multidimensional Index Structures

The task remains to efficiently support *k*-nearest neighbor search and incremental ranking for quadratic form distance functions in low-dimensional spaces. Due to the geometric shape of the query range, a quadratic form-based similarity query is called an *ellipsoid query* [Sei 97]. An efficient algorithm for ellipsoid query processing on multidimensional index structures was developed in the context of approximation-based similarity search for 3-D surface segments [KSS 97] [KS 98]. The method is designed for index structures that use a hierarchical directory based on rectilinear bounding boxes such as the R-tree [Gut 84], R+-tree [SRF 87], R*-tree [BKSS 90], X-tree [BKK 96] [BBB+ 97], and Quadtrees among others; surveys are provided in [Sam 90] [GG 98]. The technique is based on measuring the minimum quadratic form distance of a query point to the hyperrectangles in the directory. Recently, an improvement by using conservative approximations has been suggested [ABKS 98].

An important property of the method is its flexibility with respect to the similarity matrix. The matrix does not have to be available at index creation time and, therefore, may be considered as a query parameter. Thus, the users may specify and adapt the similarity weights in the matrix even at query time according to their individual preferences or to the specific requirements of the application. In any case, the same precomputed index may be used. This property is the major advantage compared to previous solutions that were developed in the context of color histogram indexing in the QBIC project [FBF+ 94] [HSE+ 95] where the index depends on a specific similarity matrix that has to be given in advance.

The cost model of [BBKK 97] provides a theoretical analysis of the performance deterioration for multidimensional index structures with increasing dimensionality. An investigation in [WSB 98] results in the recommendation to use an accelerated sequen-

tial scan, and the VA-File was developed following this paradigm. However, the analyses are based on the L_2 (Euclidean distance), L_1, and L_∞ norms that may be evaluated in linear time depending on the dimension, and the results require careful reviewing and experimental evaluation when applied to quadratic form distance functions. Even if the index is substituted by a sequential scan, the filter-refinement architecture will still be necessary due to the high cost of exact quadratic form evaluations.

4 Experimental Evaluation

We implemented the algorithms in C++ and ran the experiments on our HP C160 workstations under HP-UX 10.20. For single queries, we also implemented a HTML/Java interface that supports query specification and visualization of the results. The atomic coordinates of the 3D protein structures are taken from the Brookhaven Protein Data Bank (PDB) [BKW+ 77]. For the computation of shape histograms, we use a representation of the molecules by surface points as it is required for several interesting problems such as the molecular docking prediction [SK 95]. The reduced feature vectors for the filter step are managed by an X-tree [BKK 96] of dimension 10.

The similarity matrices are computed by an adapted formula from [HSE+ 95] where the similarity weights a_{ij} of bin i and j are defined as $a_{ij} = e^{-\sigma \cdot d(i,j)}$. The distance $d(i,j)$ is equal to the difference of the corresponding shell radii in the shell model and is given by the angle between the sector axes in the sector model. In the combined model, the shell distance $d_{shell}(i,j)$ and the sector distance $d_{sector}(i,j)$ of the bins i and j are composed by using the Euclidean distance formula $d_{comb}(i,j) = \sqrt{d^2_{shell}(i,j) + d^2_{sector}(i,j)}$. We experimented with several values of the parameter σ but did not observe significant changes in the accuracy, so we set the parameter σ equal to 10 for the following evaluations [Kas 98].

4.1 Basic Similarity Search

In order to illustrate the applicability of the similarity model, we demonstrate the retrieval of the members of a known family. As a typical example, we chose the seven Seryl-tRNA Synthetase molecules from our database that are classified by CATH [OMJ+ 97] to the same family. The diagram in Figure 7 presents the result using shape histograms for 6 shells and 20 sectors. The seven members of the Seryl family rank on the top seven positions among the 5,000 molecules of the database. In particular, the similarity distance noticeable increases for 2PFK-A, the first non-Seryl protein in the ranking order.

4.2 Classification by Shape Similarity

For the classification experiments, we restricted our database to the proteins that are also contained in the FSSP database [HS 94] and took care that for every class, at least two molecules are available. From this preprocessing, we obtained 3,422 proteins assigned to 281 classes. The classes contain between 2 and 185 molecules. In order to measure the classification accuracy, we performed *leave-one-out* experiments for various histogram models. For each molecule in the database, the nearest neighbor classification was determined after removing that element from the database. Technically, we always used the same database and selected the second nearest neighbor since the query object itself is reported to be its own nearest neighbor.

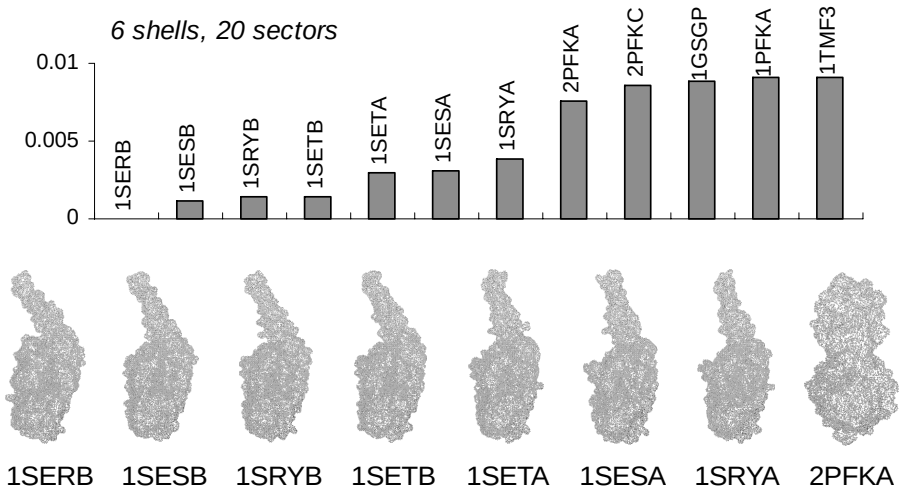

Figure 7. Similarity ranking for the Seryl-tRNA Synthetases 1SER-B. The diagram depicts the similarity distances of the 12 nearest neighbors to the query protein 1SER-B in ascending order. The illustration of the top eight molecules demonstrates the close similarity within the family, and the dissimilarity to the first non-Seryl protein in the ranking

Figure 8 demonstrates the results for histograms based on 12 shells, 20 sectors, and the combination of them. Obviously, the more fine-grained spiderweb model yields the best classification accuracy of 91.5 percent (top diagram), but even for the coarse sector histograms, a noticeable accuracy of 87.3 percent is achieved. These results compete with the accuracy of available protein classification systems such as CATH [OMJ 97] where also more than 90% of the class labels are predicted correctly. Whereas in CATH only four different class labels are used for the automatic classification, our experiments are based on a variety of 281 class labels.

The average overall runtime for a single query reflects the larger dimension of the combined model. It ranges from 0.05s for 12 shells over 0.2s for 20 sectors up to 1.42s for the combination. This runtime performance in the range of tens to thousands of milliseconds is a progress compared to established biomolecular systems for which query response times in the range of minutes and hours are reported [HS 98].

Figure 9 illustrates the effect of simply increasing the dimension of the model without combining orthogonal space partitionings. Again we observed the expected result that more information yields better accuracy. When increasing the histogram dimension by a factor of 10, the accuracy increases from 71.6 to 88.1 for the shell model, and from 87.3 to 91.6 for the sector model. For the task of classification, the increased granularity results in a better separation of the class members from the other objects. Obviously, the tradeoff for this gain is a larger space requirement and the increase of the runtime due to the high dimensionality. We plan to develop a cost model for obtaining the optimal number of bins in order to produce both accurate and fast results.

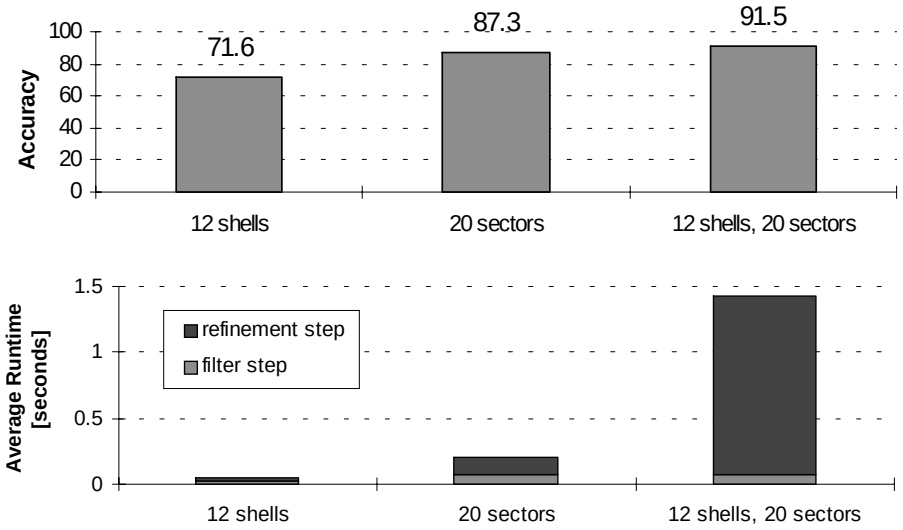

Figure 8. Classification accuracy (*top diagram*) and average runtime of query processing (*bottom diagram*) for histograms with 12 shells, 20 sectors, and their combination

In these experiments, we achieve the same accuracy for a fine-grained 122D sector model as we obtained from the 12 x 20 (240D) combined model. One may wonder why the combined model does not lead to the best accuracy. Although all proposed models yield good results in terms of accuracy and runtime, the sector model turns out to be most suitable for the tested data. One reason for this data-dependent result is that the decomposition of the 3D objects is computed for uniform sectors or equidistant shells. To reveal the properties of the space decompositions, we computed the standard deviation for each bin over all histograms. We present the observations by bar diagrams where the height of a bar represents the value of the standard deviation for the corresponding dimension.

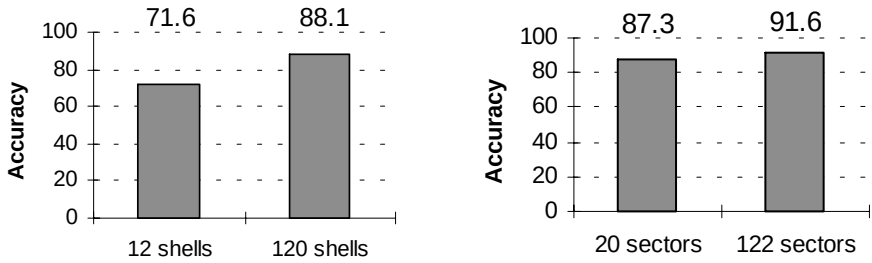

Figure 9. The accuracy increases with increasing granularity of the space partitioning for both, shell and sector histograms

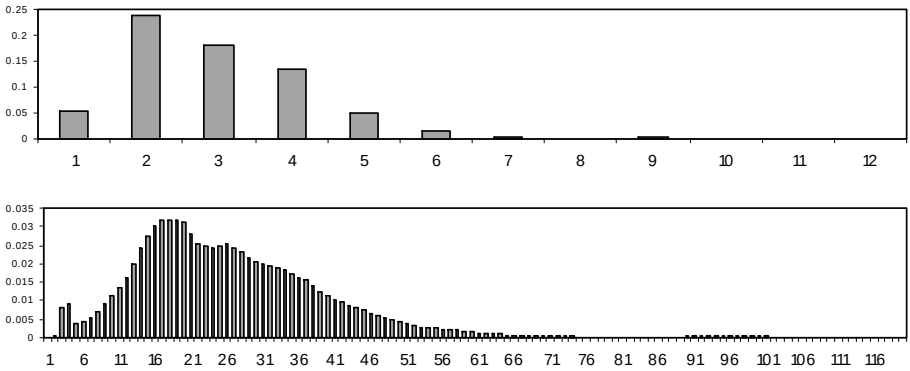

Figure 10. Standard deviations of the bins for the 12D (*top*) and 120D (*bottom*) shell models

Figure 10 demonstrates that for the shell model, the values of the standard deviations are distributed very unbalanced. For the 12D model, the highest standard deviation occurs in the shells 2 to 4 that contain the large majority of the surface points, and a low deviation is observed for the shells 7 to 12. For the 120D model, significant deviations occur only for the shells 3 to 60; there is only a low variance in the number of points for the other shells which are populated very sparse. Therefore, the corresponding histogram bins do not contribute to distinguish between different molecules but just increase the dimension and, as a consequence, the runtime becomes worse.

Figure 11 depicts the standard deviations for the two sector models, 20D and 122D. For every histogram bin, the standard deviation is high, and, therefore, all dimensions contribute to the distinction of different molecules. For the combined model, the standard deviations of the 240 bins are illustrated in Figure 12. These 240 bins result from the decomposition of the 3D space into 12 shells and 20 sectors. Therefore, the periodic pattern in the standard deviation reflects the previous observations for the shell and the sector models.

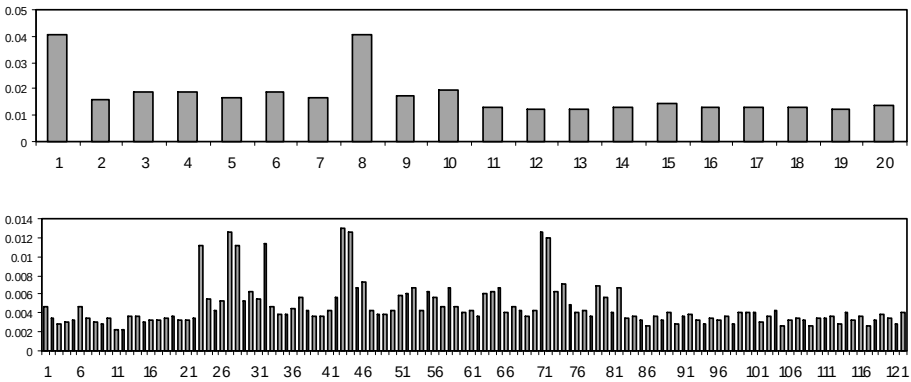

Figure 11. Standard deviations of the bins for the 20D (*top*) and 122D (*bottom*) sector models

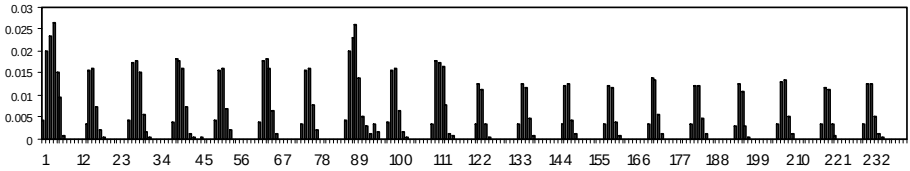

Figure 12. Standard deviations of the combined histogram bins. For each of the 20 sectors, the characteristic shape of the 12-shell histogram can be recognized

As a way to improve the properties of the shell model, we plan to use a more appropriate partitioning of the space. Instead of using equidistant radii to define the decomposition, the shell radii could be based on quantiles that are obtained from the distribution of the surface points in the space. As a consequence, this approach will also improve the effectiveness of the combined model.

5 Conclusions

In this paper, we presented a new intuitive and flexible model for shape similarity search of 3D solids. As a specific feature transform, 3D shapes are represented by using shape histograms for which several partitionings of the space are possible. This histogram model naturally is extensible to thematic attributes such as physical and chemical properties. In order to account for errors of measurement, sampling, numerical rounding etc., quadratic form distance functions are used that are able to take small displacements and rotations into account. For efficient query processing, a filter-refinement architecture is used that supports similarity query processing based on high-dimensional feature vectors and quadratic form distance functions. The experiments demonstrate both, the high classification accuracy of our similarity model, and the good performance of the underlying query processor.

The improvement of the space decomposition by using a quantile based method, the development of a cost model for determining the optimal number of bins, and the investigation of thematically extended histogram models are plans for our future work already mentioned so far. In addition, we will include a visualization of shape histograms as a Java applet in order to provide an explanation component for the classification system. This is an important issue since any notion of similarity is subjective in a high degree, and the users want to have as much feedback as possible concerning the behavior of the system depending on their queries and input parameters. Furthermore, the confidence of the users in an automatic classification increases with the reproducability of the decision by the user which can be enhanced by visualization methods. A more conceptual future work addresses the optimization of the space partitioning and the geometry of the cells which form the histogram bins. Both the number as well as the geometry of the cells affect the effectiveness and also the efficiency of similarity search and classification.

References

[ABKS 98] Ankerst M., Braunmüller B., Kriegel H.-P., Seidl T.: *Improving Adaptable Similarity Query Processing by Using Approximations*. Proc. 24th Int. Conf. on Very Large Databases (VLDB`98), New York, USA. Morgan Kaufmann (1998) 206-217

[AKS 98] Ankerst M., Kriegel H.-P., Seidl T.: *A Multi-Step Approach for Shape Similarity Search in Image Databases.* IEEE Transactions on Knowledge and Data Engineering, Vol. 10, No. 6 (1998) 996-1004

[BBB+ 97] Berchtold S., Böhm C., Braunmüller B., Keim D., Kriegel H.-P.: *Fast Parallel Similarity Search in Multimedia Databases*. Proc. ACM SIGMOD Int. Conf. on Management of Data, Tucson, AZ. ACM Press (1997) 1-12, Best Paper Award

[BBKK 97] Berchtold S., Böhm C., Keim D., Kriegel H.-P.: *A Cost Model for Nearest Neighbor Search in High-Dimensional Data Spaces*. Proc. 16th ACM SIGACT-SIGMOD-SIGART Symp. on Principles of Database Systems (PODS), Tucson, AZ (1997) 78-86

[Ber 97] Berchtold S.: *Geometry Based Search of Similar Mechanical Parts.* Ph.D. Thesis, Institute for Computer Science, University of Munich. Shaker Verlag, Aachen (1997) in German

[BKW+ 77] Bernstein F. C., Koetzle T. F., Williams G. J., Meyer E. F., Brice M. D., Rodgers J. R., Kennard O., Shimanovichi T., Tasumi M.: *The Protein Data Bank: a Computer-based Archival File for Macromolecular Structures.* Journal of Molecular Biology, Vol. 112 (1977) 535-542

[BKK 96] Berchtold S., Keim D., Kriegel H.-P.: *The X-tree: An Index Structure for High-Dimensional Data.* Proc. 22nd Int. Conf. on Very Large Data Bases (VLDB‘96), Mumbai, India. Morgan Kaufmann (1996) 28-39

[BK 97] Berchtold S., Kriegel H.-P.: *S3: Similarity Search in CAD Database Systems.* Proc. ACM SIGMOD Int. Conf. on Management of Data. ACM Press (1997) 564-567

[BKK 97] Berchtold S., Keim D. A., Kriegel H.-P.: *Using Extended Feature Objects for Partial Similarity Retrieval.* VLDB Journal, Vol. 6, No. 4. Springer Verlag, Berlin Heidelberg New York (1997) 333-348

[BKK 97a] Berchtold S., Keim D.A., Kriegel H.-P.: *Section Coding: A Method for Similarity Search in CAD Databases.* Proc. German Conf. on Databases for Office Automation, Technology, and Science (BTW). Series Informatik Aktuell. Springer Verlag, Berlin Heidelberg New York (1997) 152-171; in German

[BKSS 90] Beckmann N., Kriegel H.-P., Schneider R., Seeger B.: *The R*-tree: An Efficient and Robust Access Method for Points and Rectangles.* Proc. ACM SIGMOD Int. Conf. on Management of Data, Atlantic City, NJ. ACM Press (1990) 322-331

[CHY 96] Chen M.-S., Han J. and Yu P. S.: *Data Mining: An Overview from a Database Perspective.* IEEE Transactions on Knowledge and Data Engineering, Vol. 8, No. 6 (1996) 866-883

[FBF+ 94] Faloutsos C., Barber R., Flickner M., Hafner J., Niblack W., Petkovic D., Equitz W.: *Efficient and Effective Querying by Image Content.* Journal of Intelligent Information Systems, Vol. 3 (1994) 231-262

[GG 98] Gaede V., Günther O.: *Multidimensional Access Methods.* ACM Computing Surveys, Vol. 30, No. 2 (1998) 170-231

[GM 93] Gary J. E., Mehrotra R.: *Similar Shape Retrieval Using a Structural Feature Index.* Information Systems, Vol. 18, No. 7 (1993) 525-537

[Gut 84] Guttman A.: *R-trees: A Dynamic Index Structure for Spatial Searching.* Proc. ACM SIGMOD Int. Conf. on Management of Data, Boston, MA. ACM Press (1984) 47-57

[HS 94] Holm L., Sander C.: *The FSSP Database of Structurally Aligned Protein Fold Families.* Nucleic Acids Research, Vol. 22 (1994) 3600-3609

[HS 95] Hjaltason G. R., Samet H.: *Ranking in Spatial Databases.* Proc. 4th Int. Symposium on Large Spatial Databases (SSD'95). Lecture Notes in Computer Science, Vol. 951. Springer Verlag, Berlin Heidelberg New York (1995) 83-95

[HS 98] Holm L., Sander C.: *Touring Protein Fold Space with Dali/FSSP.* Nucleic Acids Research, Vol. 26 (1998) 316-319

[HSE+ 95] Hafner J., Sawhney H. S., Equitz W., Flickner M., Niblack W.: *Efficient Color Histogram Indexing for Quadratic Form Distance Functions.* IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 17, No. 7. IEEE Press (1995) 729-736

[Jag 91] Jagadish H. V.: *A Retrieval Technique for Similar Shapes.* Proc. ACM SIGMOD Int. Conf. on Management of Data. ACM Press (1991) 208-217

[Kas 98] Kastenmüller G.: *Shape-oriented Similarity Search in 3D Protein Database Systems.* Diploma Thesis, Institute for Computer Science, University of Munich (1998) in German

[KS 98] Kriegel H.-P., Seidl T.: *Approximation-Based Similarity Search for 3-D Surface Segments*. GeoInformatica Journal, Vol. 2, No. 2. Kluwer Academic Publishers (1998) 113-147

[KSF+ 96] Korn F., Sidiropoulos N., Faloutsos C., Siegel E., Protopapas Z.: *Fast Nearest Neighbor Search in Medical Image Databases*. Proc. 22nd VLDB Conference, Mumbai, India. Morgan Kaufmann (1996) 215-226

[KSS 97] Kriegel H.-P., Schmidt T., Seidl T.: *3D Similarity Search by Shape Approximation.* Proc. Fifth Int. Symposium on Large Spatial Databases (SSD'97), Berlin, Germany. Lecture Notes in Computer Science, Vol. 1262. Springer Verlag, Berlin Heidelberg New York (1997) 11-28

[LW 88] Lamdan Y., Wolfson H.J.: *Geometric Hashing: A General and Efficient Model-Based Recognition Scheme*. Proc. IEEE Int. Conf. on Computer Vision, Tampa, Florida, 1988 238-249

[Mit 97] Mitchell T.M.: *Machine Learning.* McCraw-Hill, (1997)

[MST 94] Michie D., Spiegelhalter D.J., Taylor C.C.: *Machine Learning, Neural and Statistical Classification.* Ellis Horwood (1994)

[OMJ+ 97] Orengo C.A., Michie A.D., Jones S., Jones D.T. Swindells M.B., Thornton, J.M.: *CATH – A Hierarchic Classification of Protein Domain Structures.* Structure, Vol. 5, No. 8 (1997) 1093-1108

[Sam 90] Samet H.: *The Design and Analysis of Spatial Data Structures.* Addison Wesley (1990)

[Sei 97] Seidl T.: *Adaptable Similarity Search in 3-D Spatial Database Systems.* Ph.D. Thesis, Institute for Computer Science, University of Munich (1997). Herbert Utz Verlag, Munich, http://utzverlag.com, ISBN: 3-89675-327-4

[SK 95] Seidl T., Kriegel H.-P.: *A 3D Molecular Surface Representation Supporting Neighborhood Queries*. Proc. 4th Int. Symposium on Large Spatial Databases (SSD`95), Portland, Maine, USA. Lecture Notes in Computer Science, Vol. 951. Springer Verlag, Berlin Heidelberg New York (1995) 240-258

[SK 97] Seidl T., Kriegel H.-P.: *Efficient User-Adaptable Similarity Search in Large Multimedia Databases.* Proc. 23rd Int. Conf. on Very Large Databases (VLDB'97), Athens, Greece. Morgan Kaufmann (1997) 506-515

[SK 98] Seidl T., Kriegel H.-P.: *Optimal Multi-Step k-Nearest Neighbor Search.* Proc. ACM SIGMOD Int. Conf. on Management of Data, Seattle, Washington (1998) 154-165

[SRF 87] Sellis T., Roussopoulos N., Faloutsos C.: *The R+-Tree: A Dynamic Index for Multi-Dimensional Objects*. Proc. 13th Int. Conf. on Very Large Databases, Brighton, England (1987) 507-518

[TC 91] Taubin G., Cooper D. B.: *Recognition and Positioning of Rigid Objects Using Algebraic Moment Invariants.* in *Geometric Methods in Computer Vision*, Vol. 1570, SPIE (1991) 175-186

[WK 91] Weiss S.M., Kulikowski C.A.: *Computer Systems that Learn: Classification and Prediction Methods from Statistics, Neural Nets, Machine Learning, and Expert Systems*. Morgan Kaufmann, San Francisco (1991)

[WSB 98] Weber R., Schek H.-J., Blott S.: *A Quantitative Analysis and Performance Study for Similarity-Search Methods in High-Dimensional Spaces.* Proc. 24th Int. Conf. on Very Large Databases (VLDB`98), New York, USA. Morgan Kaufmann (1998) 194-205