

# Adaptable Similarity Search in 3-D Spatial Database Systems

Dissertation im Fach Informatik  
an der Fakultät für Mathematik und Informatik  
der Ludwig-Maximilians-Universität München

von

Thomas Seidl

Oktober 1997

## **Content of this Survey:**

- Abstract (pp 2-3)
- Table of Contents (pp 4-10)
- Conclusion (Summary, pp 11-15)

Berichterstatter:

Prof. Dr. Hans-Peter Kriegel, Ludwig-Maximilians-Universität München

Prof. Dr. Oliver Günther, Humboldt-Universität zu Berlin

# Abstract

Similarity search in database systems is becoming an increasingly important task in modern application domains such as multimedia, CAD, molecular biology, medical imaging and many others. By investigating novel adaptable similarity models and new algorithms for efficient similarity query processing a fundamental requirement of similarity search in very large database systems is targeted. Particular attention is paid to the basic observation that similarity highly depends upon the requirements of specific applications and on the changing needs of individual users.

After some preliminary work concerning the subject of similarity search in spatial database systems and a survey of the 3-D protein database system which is covered in part I, several adaptable similarity models are presented in part II. Firstly, the fundamental notion of quadratic form distance functions,  $d_A^2(x, y) = (x - y) \cdot A \cdot (x - y)^T$ , is introduced. By providing appropriate feature transforms such as the computation of histograms and by specifying and modifying similarity matrices  $A$ , the basic model is shown to support particular adaptability for a wide variety of applications. The potentials of this multi-purpose approach are demonstrated through several examples including color-orientated similarity of image, shape-orientated similarity of images, histogram-based similarity of 3-D shapes and approximation-based shape similarity of 3-D surface segments.

In addition to the availability of adaptable similarity models, the aspect of efficient query processing becomes ever more important due to the increasing sizes of large and vary large databases. Part III is dedicated to the topic of efficient support for adaptable similarity query processing and begins with the presentation of a novel algorithm for optimal multi-step nearest-neighbor search. The multi-purpose quadratic form similarity distance functions represent ellipsoid queries which are a new query type for spatial database systems. New algorithms are introduced for efficiently processing ellipsoid

queries on multidimensional index structures and particular care is taken for the adaptability of the similarity model by the user. Specifically, this algorithm supports modifications of the similarity matrix (i.e: the query ellipsoid) at query time. After investigating several approximation techniques for ellipsoid queries a multi-step technique for efficient ellipsoid query processing in even high-dimensional spaces is presented that applies existing as well as new techniques for the reduction of dimensionality to query ellipsoids. The reduced similarity query is again an ellipsoid query and guarantees no false drops in the multi-step query processor. In particular, the reduced ellipsoid query represents the greatest of all lower-bounding filter distance functions thus ensuring the optimal filter selectivity. Extensive experiments on an image database containing 112,000 color images and 10,000 grayscale images as well as on the 3-D protein database system containing 5,000 molecules with 94,000 surface segments demonstrate the effectiveness and high efficiency of these new concepts in various dimensions from 5-D up to 4,096-D.

# Table of Contents

Acknowledgements .....	i
Abstract .....	iii
Survey of Chapters .....	v
Table of Contents .....	vii
List of Figures .....	xv
List of Tables .....	xxi
List of Definitions .....	xxiii

## PART I. PRELIMINARIES

<b>1</b>	<b>Introduction</b>	<b>3</b>
1.1	<i>Spatial Database Systems and Similarity Search</i> .....	3
1.1.1	Spatial Database Systems .....	3
1.1.2	Spatial Query Types .....	4
1.1.3	Similarity Search .....	5
1.1.4	Related Work .....	6
1.1.5	Adaptability of Similarity Search .....	7
1.2	<i>An Application from Molecular Biology</i> .....	9
1.2.1	Docking Search for Proteins .....	9
1.2.2	Object-oriented 3-D Protein Database System .....	11
1.3	<i>Outline of the Thesis</i> .....	12
1.3.1	Preliminaries (Part I) .....	12
1.3.2	Adaptable Similarity Models (Part II) .....	12
1.3.3	Efficient Processing of Similarity Queries (Part III) .....	13
<b>2</b>	<b>Similarity Query Types</b>	<b>15</b>
2.1	<i>Introduction</i> .....	15

2.2	<i>Formal Definitions</i> .....	16
2.2.1	Similarity Range Query .....	16
2.2.2	Nearest Neighbor Queries .....	17
2.2.3	k-Nearest Neighbor Queries .....	18
2.2.4	Incremental Similarity Ranking Queries .....	20
2.3	<i>Summary</i> .....	23
<b>3</b>	<b>Molecular Surfaces in Protein Database Systems</b>	<b>25</b>
3.1	<i>Introduction</i> .....	25
3.1.1	Surface Representation Techniques .....	26
3.1.2	Applications .....	26
3.2	<i>Related Work</i> .....	28
3.2.1	Quad-Edge Data Structure .....	28
3.2.2	Other Graph Representations .....	30
3.3	<i>Molecular Surfaces</i> .....	31
3.3.1	Structure of Molecular Surfaces .....	31
3.3.2	An Algebraic Specification for Molecular Surfaces .....	33
3.4	<i>Neighborhood Queries</i> .....	35
3.4.1	Specification .....	35
3.4.2	Applications .....	35
3.4.3	Neighborhood Query Processing .....	36
3.5	<i>The TriEdge Data Structure</i> .....	38
3.5.1	Motivation .....	38
3.5.2	Simplified Surface Graph .....	39
3.5.3	TriEdge Records .....	40
3.5.4	Experimental Evaluation .....	42
3.5.5	Concluding Remarks .....	44
3.6	<i>Integration into a Protein Database System</i> .....	45
3.6.1	Object-Oriented Database Schema .....	45
3.6.2	Derived Surface Points .....	46
3.6.3	Conclusion .....	47

## PART II. ADAPTABLE SIMILARITY MODELS

<b>4</b>	<b>Adaptability of Similarity Models</b>	<b>51</b>
4.1	<i>Similarity Models</i> .....	51
4.1.1	Feature-based Similarity .....	52
4.1.2	Geometry-based Similarity .....	53
4.1.3	Histograms as Feature Vectors .....	54

4.2	<i>Adaptable Similarity Distance Functions</i>	56
4.2.1	Introduction	56
4.2.2	Shortcomings of the Euclidean Distance	57
4.2.3	Quadratic Form Distance Functions	59
4.2.4	Non-Symmetric Similarity Matrices	60
4.3	<i>Color-orientated Similarity of Images</i>	61
4.3.1	Color Frequencies of Images	61
4.3.2	Color Layout Similarity of Images	62
4.3.3	Example Image Database	63
4.4	<i>Adaptability of the Similarity Models</i>	64
4.4.1	Aspects of Adaptability	64
4.4.2	Efficient Query Processing	65
4.4.3	Conclusions	66
<b>5</b>	<b>Shape-orientated Similarity of Images</b>	<b>67</b>
5.1	<i>Digital Images</i>	68
5.1.1	Color and Grayscale Images	68
5.1.2	A Sample Database of Grayscale Images	69
5.1.3	Normalization of Images	71
5.1.4	Representation of Images	72
5.1.5	Characteristics of the Image Data	74
5.2	<i>Basic Shape Similarity Model for Images</i>	76
5.2.1	Difference Images of Color and Grayscale Images	76
5.2.2	Total Power of Digital Images	77
5.2.3	A Similarity Distance Function for Images	78
5.2.4	Experimental Evaluation	81
5.3	<i>Adaptable Shape-oriented Similarity of Images</i>	82
5.3.1	Motivation	82
5.3.2	Neighborhood Influence Weights	85
5.3.3	Adaptable Image Distance Function	87
5.3.4	An Example for the Adaptable Similarity Distance	88
5.3.5	Displaced Shape Components	90
5.3.6	An Experimental Example	91
5.4	<i>A System for Shape-oriented Similarity Search</i>	93
5.4.1	A Graphical User Interface for Image Similarity Search	94
5.4.2	Processing Basic Image Similarity Queries	95
5.4.3	Processing Adaptable Image Similarity Queries	96
5.4.4	Conclusions	98
<b>6</b>	<b>Similarity Models for 3-D Objects</b>	<b>99</b>
6.1	<i>Shape Histogram Similarity of 3-D Objects</i>	100
6.1.1	Shape Histograms of 3-D Objects	100
6.1.2	Section Histograms in the 3-D	101
6.1.3	Similarity Distance Functions	102
6.1.4	Experimental Evaluation	103

6.2	<i>Volume-orientated Similarity of 3-D Solids</i>	105
6.2.1	Difference Volume of 3-D Solids	106
6.2.2	Volume Similarity Distance Function	108
6.2.3	Analysis of the Difference Volume Distance Function	109
6.2.4	Volume-orientated Similarity and Voxel Representation	111
6.2.5	Adaptable Volume-oriented Similarity of 3-D Solids	113
6.2.6	Efficient Query Processing	114
6.3	<i>Potentials of Histogram-based Molecular Similarity</i>	114
6.3.1	Combined Shape and Structure Histograms	115
6.3.2	Fuzzy Objects and Molecular Flexibility	115
6.3.3	Conclusions	117
<b>7</b>	<b>Approximation-based Similarity of 3-D Segments</b>	<b>119</b>
7.1	<i>Introduction</i>	119
7.1.1	Problem and Applications	120
7.1.2	3-D Surface Segments	121
7.1.3	Basic Idea of our Approach	122
7.2	<i>Approximation of 3-D Segments</i>	123
7.2.1	Approximation Models	124
7.2.2	(Unique) Approximation of a 3-D Segment	125
7.2.3	Approximation by Singular Value Decomposition	128
7.2.4	Normalization in the 3-D	129
7.3	<i>Shape Similarity of 3-D Segments</i>	129
7.3.1	Approximation-based Similarity Distance Function	129
7.3.2	Sample Application	132
7.3.3	Multi-Step Similarity Query Processing	133
7.3.4	A Lower Bound for Shape Similarity	135
7.3.5	Query Ellipsoids in the Filter Step	137
7.3.6	Experimental Evaluation	138
7.3.7	Conclusions	140

## PART III. EFFICIENT PROCESSING OF SIMILARITY QUERIES

<b>8</b>	<b>Multi-Step Similarity Query Processing</b>	<b>143</b>
8.1	<i>Introduction</i>	143
8.2	<i>Basic Spatial Query Processing</i>	144
8.2.1	Multidimensional Access Methods	144
8.2.2	A Framework for Query Processing on Indexes	146
8.2.3	Particular Range Query Classes	149

8.3	<i>Multi-Step Query Processing</i>	149
8.3.1	Requirements	149
8.3.2	Correctness of Multi-Step Query Processing	152
8.3.3	Related Work	153
8.3.4	Complex Region Query Processing	154
8.3.5	Similarity Range Query Processing	155
8.3.6	Lower-Bounding Distance Functions	156
8.3.7	Conclusion	157
<b>9</b>	<b>Optimal k-Nearest Neighbor Search</b>	<b>159</b>
9.1	<i>k-Nearest Neighbor Search using Index Structures</i>	159
9.1.1	Introduction	160
9.1.2	Cell-based Nearest Neighbor Search	161
9.1.3	k-Nearest Neighbor Search on Indexes	163
9.1.4	Incremental Similarity Ranking	165
9.2	<i>Multi-Step k-Nearest Neighbor Search</i>	167
9.2.1	Introduction	167
9.2.2	State-of-the-Art Algorithm	169
9.2.3	Efficiency of k-Nearest Neighbor Search	171
9.3	<i>Optimal k-Nearest Neighbor Search</i>	174
9.3.1	Optimal Multi-Step Algorithm	174
9.3.2	Analysis of the Optimal Algorithm	176
9.3.3	Performance Evaluation	180
9.3.4	Conclusions	183
<b>10</b>	<b>Ellipsoid Query Processing in Low-Dimensional Spaces</b>	<b>185</b>
10.1	<i>Ellipsoids as Query Objects</i>	185
10.1.1	Quadratic Form Distance Functions	186
10.1.2	Properties of Similarity Matrices	187
10.1.3	Geometry of Ellipsoids	189
10.2	<i>Exact Ellipsoid Query Processing</i>	192
10.2.1	Adaptability at Query Time	192
10.2.2	Ellipsoid Query Processing on Multidimensional Index Structures	194
10.2.3	Basic Distance Algorithm for Ellipsoids and Boxes	197
10.3	<i>Evaluation of the Exact Algorithm</i>	200
10.3.1	Runtime Complexity of the Algorithm	201
10.3.2	Performance of Ellipsoid Queries on Indexes	202
10.3.3	Performance of Ellipsoid Queries as Filter Step	204
10.3.4	Conclusions	206
<b>11</b>	<b>Ellipsoid Query Processing using Approximations</b>	<b>207</b>
11.1	<i>Approximation Techniques</i>	207
11.1.1	Previous Work	208
11.1.2	Requirements, Advantages, and Problems	209
11.1.3	Approximations and Lower-Bounding Property	209
11.1.4	Conservative Approximation Techniques	210

<i>11.2 Minimum Bounding Box Approximation</i>	212
11.2.1 Introduction	212
11.2.2 MBB Approximation for Range Queries	213
11.2.3 Greatest Lower-Bounding Box Distance Function	214
11.2.4 Box Query Processing	216
<i>11.3 Minimum Rotated Bounding Box Approximation</i>	217
11.3.1 Geometry of an MRBB	217
11.3.2 Approximation Quality	219
11.3.3 Rotated Box Query Processing	221
11.3.4 Advantages and Problems	223
<i>11.4 Minimum Bounding Sphere Approximation</i>	225
11.4.1 Geometry of the Minimum Bounding Sphere	225
11.4.2 Sphere Approximation Quality	227
11.4.3 Greatest Lower-Bounding Sphere Distance Function	228
11.4.4 Sphere Query Processing	229
<i>11.5 Combined Conservative Approximations</i>	229
11.5.1 Combination of Approximations	230
11.5.2 Greatest Lower-Bounding Combined Distance Function	232
11.5.3 Combined Query Processing	233
<i>11.6 Evaluation of the Approximation Techniques</i>	235
11.6.1 Approximation Quality	236
11.6.2 Performance of Single-Step Queries on Indexes	236
11.6.3 Performance of Multi-Step Query Processing	239
11.6.4 Conclusions	240
<b>12 Ellipsoid Query Processing in High-Dimensional Spaces</b>	<b>243</b>
<i>12.1 High-Dimensional Spaces</i>	244
12.1.1 Problems with High Dimensions	244
12.1.2 Reduction of Dimensionality	245
12.1.3 Principle of Linear Reduction Techniques	248
<i>12.2 Linear Techniques for Reducing the Dimensionality</i>	249
12.2.1 Karhunen-Loève Transform (KLT)	250
12.2.2 Discrete Fourier Transform (DFT)	252
12.2.3 Histogram Coarsening Technique (HCT)	254
12.2.4 Matri-Dependent Reduction Techniques	257
<i>12.3 Reduction of General Ellipsoid Queries</i>	258
12.3.1 Complementing Reduction Matrices	258
12.3.2 Distance-Preserving Transformation	259
12.3.3 Optimal Lower Bounding of Similarity Distances	260
12.3.4 Overall Similarity Matrix Reduction	261
<i>12.4 Minimum Bounding Box of a Reduced Ellipsoid</i>	263
12.4.1 Direct Computation of the Reduced Approximation	263
12.4.2 Greatest Lower-Bounding Reduced Box Distance Function	265

<i>12.5 Experimental Evaluation</i> .....	268
12.5.1 Performance of Query Processing .....	269
12.5.2 Comparison of Reduction Techniques .....	271
12.5.3 Performance in Ultra-High Dimensions .....	272
12.5.4 Conclusions .....	275
<b>13 Conclusions</b> .....	<b>277</b>
<i>13.1 Summary</i> .....	277
13.1.1 Preliminaries (Part I) .....	277
13.1.2 Adaptable Similarity Models (Part II) .....	278
13.1.3 Efficient Similarity Query Processing (Part III) .....	278
<i>13.2 System Architecture</i> .....	282
<i>13.3 Potentials and Future Work</i> .....	284
References .....	285
Index .....	293
Curriculum Vitae .....	297

# Chapter 13

## Conclusions

We conclude this work by a summary of the concepts as well as of the theoretical and practical results. After an illustration of the system architecture which we implemented for our experiments, we provide an outlook to future work and discuss the potentials of our approaches.

### 13.1 Summary

In this thesis, we present our research on adaptable similarity search in large spatial database systems. Particular attention is paid to the adaptability of similarity models, and we distinguish two levels: The *application level* aims at the adaptability of a particular similarity models to the characteristics of individual applications. For instance, the structure and resolution of feature vectors such as histograms may be tuned to fit the specific domain. The *user level* addresses the adaptability of the similarity distance function to the specific requirements of individual users. The personal needs may vary from query to query, and a modern similarity search system should support the interactive modification of the similarity distance function at query time.

#### 13.1.1 Preliminaries (Part I)

The preliminaries in part I illustrate the topic and the background of this work. Typical similarity query types are formally specified, and an application from molecular

biology is explained in more detail. We introduce our 3-D protein database system including our new TriEdge data structure that aims at the efficient storage of molecular surfaces and at the support of topological neighborhood queries.


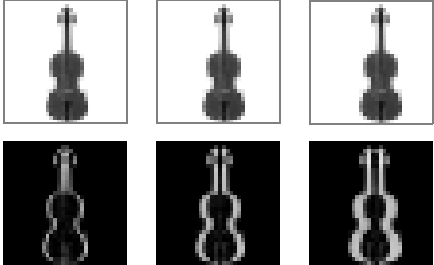
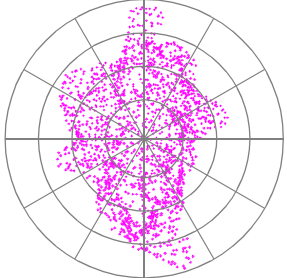
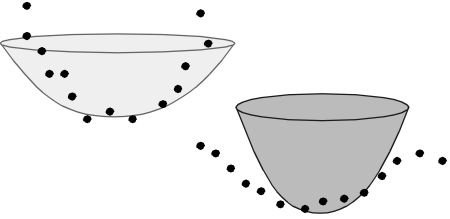
### 13.1.2 Adaptable Similarity Models (Part II)

Part II of the thesis provides a variety of adaptable similarity models, and figure 1 illustrates the approaches. At the beginning, we characterize several approaches to define similarity models in general and emphasize the aspect of adaptability. The first application is the *color-oriented similarity model* for image databases where the powerful quadratic form distance functions were originally introduced for similarity search. The approach employs color histograms that represent the frequency of certain colors in an image, and the components of the similarity matrix correspond to the cross-similarity of individual hues.

Based on this fundamental model, we present adaptable similarity models for several other applications: The *shape-oriented similarity of images* aims at the level of individual pixels. By our novel approach to take a user-specified neighborhood influence area into account, the similarity distance function can be adapted to regard small displacements of shapes within the images. The *shape-oriented similarity of 3-D objects* is demonstrated by examples from our 3-D protein database system. We use shape histograms which provide a discrete and tight representation of 3-D solids such as molecules or mechanical parts. Again, the user may specify her or his personal requirements by modifying the similarity matrix. By the *approximation-based shape similarity of 3-D surface segments*, we present a similarity model that measures the similarity of 3-D segments by using geometric approximations. Multi-parametric surface functions are employed in order to represent complex 3-D segments in a way that is quite convenient for similarity comparisons, and the similarity distance is defined in terms of the mutual approximation error. Sample query results from similarity search on large databases demonstrate the usefulness and the potentials of the presented similarity models.

### 13.1.3 Efficient Similarity Query Processing (Part III)

For the presented similarity models, part III provides positive and constructive answers to the question whether and how efficient query processing on large databases is supported. Figure 2 illustrates the proposed new methods. We start with an introduction into multi-step similarity query processing based on multidimensional index structures

<b>Color-oriented Similarity of Images (chapter 4)</b>		
	<p><i>Method</i></p> <ul style="list-style-type: none"> <li>• color histograms</li> <li>• color similarity matrices</li> </ul> <p><i>Test database</i></p> <p>112,000 color images; 64-D; 112-D; 256-D</p>	
<b>Shape-oriented Similarity of Images (chapter 5)</b>		
	<p><i>Method</i></p> <ul style="list-style-type: none"> <li>• total power of difference images</li> <li>• neighborhood influence area</li> </ul> <p><i>Test database</i></p> <p>10,000 grayscale images; 256-D; 1,024-D; 4,096-D</p>	
<b>Shape-oriented Similarity of 3-D Objects (chapter 6)</b>		
	<p><i>Method</i></p> <ul style="list-style-type: none"> <li>• 3-D shape histograms</li> <li>• adaptable similarity matrices</li> </ul> <p><i>Test database</i></p> <p>5,000 proteins; 10 up to 1,000 dimensions</p>	
<b>Approximation-based Shape Similarity of 3-D Surface Segments (ch. 7)</b>		
	<p><i>Method</i></p> <ul style="list-style-type: none"> <li>• 3-D shape approximation</li> <li>• mutual approximation error</li> </ul> <p><i>Test database</i></p> <p>94,000 protein surface segments 7-D; 9-D; 11-D</p>	

**Figure 1:** Adaptable similarity models in part II.

and proceed with a quite general novel algorithm for *optimal multi-step k-nearest neighbor search*. Depending on the provided filter distance function, our procedure retrieves the minimum number of candidates from the index-based filter step. In our experiments, we obtained improvement factors of up to 99.7 for a large image database.

The following chapters of part III are dedicated to the ellipsoid query which has been identified as a fundamental query type for adaptable similarity search. By modifying the similarity matrix, the ellipsoid distance function is adapted to individual requirements. For processing *ellipsoid queries in low-dimensional spaces*, we propose a novel algorithm that is based on rectilinearly organized multidimensional index structures. Our new basic ellipsoid-and-box operation supports the modification of the query ellipsoid even at query specification time. Thus, the user may interactively adapt the similarity distance function to her or his preferences from query to query. Our experiments demonstrate a good performance in single-step as well as in multi-step environments. In order to support ellipsoid query processing in legacy systems that offer only window queries or spherical distance functions, for instance, we investigate methods for *approximate ellipsoid query processing*. Several conservative approximations for query ellipsoids are determined, and by a generalization of the approaches, we come up with spherical and box-shaped lower-bounding distance functions to support  $k$ -nearest neighbor search. However, the experiments demonstrate that in many cases, the exact ellipsoid query algorithm is superior to the approximate ones. Especially for low-dimensional range queries, the approximations effect a noticeable performance gain.

Part III is concluded by efficient techniques for processing *ellipsoid queries in high-dimensional spaces*. We investigate and classify existing techniques for the reduction of dimensionality and propose a new method, the Histogram Coarsening Technique. The main result of our analysis is that the reduction techniques are adaptable to our new query type, the ellipsoid query. We present a novel algorithm to reduce similarity matrices with respect to the reduction technique by which a given index has been created. The resulting filter distance function again represents an ellipsoid query, and it lower-bounds the original similarity distance function. In particular, we obtain the greatest of all lower-bounding filter distance functions for what reason the algorithm guarantees both, no false drops as well as the optimal filter selectivity. Several experiments on varying similarity matrix and for various techniques for reduction of dimensionality demonstrate the good performance of our approach even for high dimensions of 256, 1,024, or 4,096 which occur in our image database.

<b>Optimal k-Nearest Neighbor Search</b> (chapter 9)		
		<p><i>Contribution</i></p> <ul style="list-style-type: none"> <li>• optimal multi-step algorithm for <math>k</math>-nearest neighbor search</li> </ul>
<b>Ellipsoid Queries in Low-dimensional Spaces</b> (chapter 10)		
		<p><i>Contribution</i></p> <ul style="list-style-type: none"> <li>• ellipsoid query processing on multidimensional index structures</li> <li>• basic ellipsoid-and-box algorithm</li> </ul>
<b>Approximate Ellipsoid Query Processing</b> (chapter 11)		
		<p><i>Contribution</i></p> <ul style="list-style-type: none"> <li>• conservative approximations of ellipsoids</li> <li>• greatest lower-bounding filter distance functions</li> </ul>
<b>Ellipsoid Queries in High-dimensional Spaces</b> (chapter 12)		
		<p><i>Contribution</i></p> <ul style="list-style-type: none"> <li>• reduction of dimensionality for similarity matrices</li> <li>• greatest lower-bounding reduced ellipsoid distance</li> </ul>

**Figure 2:** Efficient similarity query processing in part III.