Improving the performance of shape similarity retrieval systems

Mats Carlin
University of Oslo and SINTEF Electronics & Cybernetics
Box 124 Blindern, N-0314 Oslo, NORWAY
Tel: +47 2206 7300, Fax: +47 2206 7350
Mats.Carlin@ecy.sintef.no
http://www.ifi.uio.no/~matsca
Preprint from NOBIM-2000, Trondheim, Norway, 6-7 June 2000.

Abstract

The main aim of this paper is to present methods and strategies improving the performance of feature-based shape similarity retrieval systems.

First we provide a framework for measuring the performance of shape similarity retrieval systems. The concepts of relevance and redundancy are used to select optimal feature subsets and to assess which features are most important to shape similarity retrieval. The most important features seem to be features with a high-level abstract interpretation.

The methods have been implemented and tested in a system for retrieval of Computer Aided Design (CAD) drawings of aluminium sections called the DieFinder, yielding significant improvement.

Keywords : shape, similarity, retrieval, performance, features, relevance, redundancy

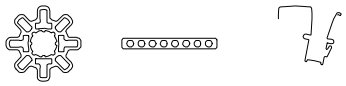
1 Introduction

When data mining large databases with images or drawings of objects, there is a need to search for objects with similar or resembling shapes [1]. Our interest in this field is inspired by a problem within the aluminium industry where a large database of Computer Aided Design (CAD) drawings of aluminium sections should be made searchable [2].

Exact shape similarity under a simple geometric transformation was solved over 2000 years ago by the ancient Greeks, but approximate shape similarity is still a difficult problem to solve [3], see fig. 1. In this paper we report the results of a dr.scient thesis recently submitted on the subject of improving the performance of shape similarity retrieval systems [4].

Feature descriptions can be computed automatically for large image databases provided that we have robust and accurate implementations of the feature extractors [5]. Retrieval is fast even for huge image databases, but we still lack an understanding of which features are most important for retrieval performance.

2 Measuring the performance of shape similarity retrieval methods



In the first paper we discuss how to evaluate the performance of different shape similarity retrieval methods [6].

2.1 Motivation

To enable systematic improvement of some computational method one must be able to measure the performance of that computational method. According to recent cover papers the issue of comparing different shape similarity retrieval methods and systems has mainly been neglected in the research community due to the subjective character of such comparisons [3, 7]. In shape similarity retrieval the common measure is relevance feedback [8, 9]. Relevance feedback is based on performing shape similarity retrieval and counting the number of hits and misses of each new search. The results of relevance feedback can be used to adjust the retrieval algorithm. Our main idea was to use several sources of information to measure the performance of the shape similarity retrieval methods, not only our perceived relevance. We also wanted to rate the similarity of different objects on a continuous scale, not only as hits and misses.

2.2 Methodology

We have developed a framework for measuring the performance of shape similarity retrieval methods based on three different knowledge sources; Human perception of similarity, the application-specific information and the mathematical representation of shape. The measures are based on measuring the difference between a number of reference objects and all the other objects in the database. Seven different measures are presented and four of them are tested on a database of Computer Aided Design drawings of aluminium extruded sections.

We apply the different performance measures to different specific groups of features, including skeleton-based features, three different kinds of moments, elliptic Fourier descriptors, fuzzy/symmetry features and a mixed set of features. To each specific group of features, a common group of five important general shape features were added providing basic information about each shape.

3 Selecting feature subsets and distance measures for shape similarity retrieval

In the second paper we discuss how to select a distance metric and identify the most important shape features for shape similarity retrieval [10].

3.1 Motivation

Redundant and irrelevant features are frequently encountered for many applications since it is easy to compute many features. In a feature-based shape similarity retrieval system, we wish to select a relatively small subset of features with optimal performance. We also want to select an optimal metric, since it influences the performance. These are difficult tasks since the combinatorial problem of selecting feature subsets is increasing exponentially with the number of available features. Recent cover papers have stated that there is lack of papers testing feature subset selection methods on large real-world problems [11].

3.2 Methodology

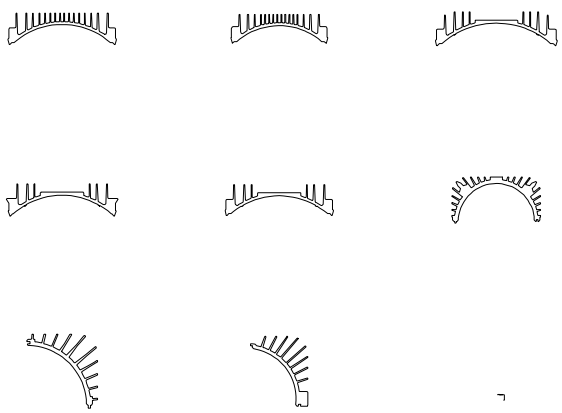
We have introduced the concepts of relevance and redundancy with respect to shape similarity retrieval and have applied these to select optimal relevant and non-redundant feature subsets. Our definition of relevance differs from recent papers on relevance [12, 13].

Definition 1 (Relevance to retrieval) *Let F be the complete set of features $\{f_k\}_{k=1}^{m}$ and let $g(\{f\})$ be the performance of shape similarity retrieval with respect to a feature subset $\{f\}$. A feature f_i is to if the retrieval performance $g(F) > g(F\backslash\{f_i\})$. A feature f_i is to similarity retrieval if it is possible to remove a subset of the features from F such that f_i becomes strongly relevant.*

Then we define redundancy with respect to subsets of features based on covers. Our covers are inspired by the use of blankets to select optimal feature subsets for classification tasks [14], but are defined for continuous variables.

Definition 2 (Redundancy) *Let F be the complete set of features $\{f_k\}_{k=1}^{m}$. A feature f_i is linearly if the feature can be described as a linear function $f_i = h(\{f_j\}|\{f_j\} \subseteq F\backslash\{f_i\})$ of a subset of the other features. A minimal subset implying redundancy of a feature is called a .*

The degree of partial redundancy can be measured by the multiple correlation coefficient [15]. The two definitions are used to select feature subsets and distance measures for shape similarity retrieval. We use forward selection, backward elimination, hybrid methods and stochastic methods such as genetic algorithms and simulated annealing.



4 Which shape features are most important for shape similarity retrieval?

In the third paper we assess which individual shape features are best suited for shape similarity retrieval [16].

4.1 Motivation

It is easy to compute a large number of features from well-defined shapes, the literature on feature extractors is enormous, see e.g. [5, 17]. It is however an unsolved issue in shape similarity retrieval which features are really important for retrieval performance. There is an obvious need to assess the individual features in a shape similarity retrieval.

4.2 Methodology

First, we provide a hierarchy of requirements to each individual feature. Necessary requirements are invariance, global scope, indexability and computability. To assess the individual features we have focused on the objective requirements of relevance, redundancy, independence and statistical orthogonality. We present a much longer list of wanted requirements, but most of them are not evaluated in the paper.

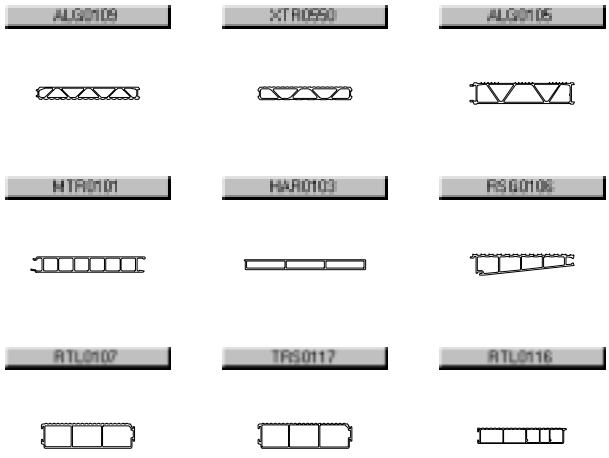


Figure 4: Query result for a building section (ALG0109)

4.3 Results

110 different shape features have been assessed using redundancy and relevance as the objective measures. We are able to eliminate many features using covers to measure the redundancy of each feature. Features with high-level abstract interpretations are most relevant to the task of shape similarity retrieval.

The 12 most important features for our application are rectangularity (area and perimeter fraction), diameter, circularity (area fraction), area, perimeter, U-shape, length of fins, thickness (absolute and ratio), area of holes and number of fins. We would like to stress that other features may be important for other applications.

5 Conclusion

The major contributions in the dr.scient thesis are:

- A framework for **measuring the performance** of shape similarity retrieval methods based on human perception of similarity, application-specific information and the mathematical representation has been established [6].
- We have defined the two concepts of **relevance** and **redundancy** with respect to shape similarity retrieval. This approach is new within shape similarity retrieval. We have used these two concepts to select **optimal** feature subsets containing those features which are most important to shape similarity retrieval. We have showed empirically that we can reduce a large feature set with many redundant and a few irrelevant features by 80 % and at the same time increase retrieval performance significantly [10].
- We have assessed the **importance** of individual shape features with respect to shape similarity retrieval by the relevance and redundancy of the individual features [16].

In addition some minor contributions have been achieved that are not reported here.

- A robust method for computing geometric moments from polygons have been described and we have proved that some commonly used methods are sensitive to the slope of the individual line segments in a polygon [18].
- We have presented a new method for measuring complexity of non-fractal objects based on the divider-step method which is commonly used to estimate the fractal dimension of fractal objects [19].

The methods have been implemented and tested in a system for retrieval of Computer Aided Design (CAD) drawings of aluminium sections called the DieFinder, yielding significant improvement [4].

6 Acknowledgements

A special thanks to my supervisor Tom Kavli and to my colleagues at SINTEF.

References

[1] D. Forsyth, J. Malik & R. Wilensky, Searching for digital pictures, *Scientific American*, 72-77, June 1997.

[2] M. Carlin, T. Kavli, Similarity analysis of extruded aluminium sections, *Proceedings of NOBIM-98, Norwegian Image Processing and Pattern Recognition conference* (June 1998), Oslo, Norway, 28–38.

[3] A. Gupta & R. Jain, Visual Information Retrieval, *Communications of the ACM* **40(5):**71-79, 1997.

[4] M. Carlin, Improving the performance of shape similarity retrieval systems, dr.scient thesis submitted to the Department for Informatics, University of Oslo, 2000.

[5] Sven Loncaric, A survey of shape analysis techniques, *Pattern Recognition* **31(8)**:983-1001, 1998.

[6] M. Carlin, Measuring the performance of shape similarity retrieval methods, submitted to *Computer Vision & Image Understanding*, special issue on empirical evaluation of computer vision algorithms.

[7] V. N. Gudivada & V. V. Raghavan, Content-Based Image Retrieval Systems, *IEEE Computer*, September 1995, 18-22.

[8] R. Price, T.-S. Chua & S. Al-Hawamdeh, Applying relevance feedback to a photo archival system, *Journal of Information Science* **18**:203-215, 1992.

[9] J. Peng, B. Bhanu & S. Qing, Probabilistic feature relevance learning for content-based image retrieval, *Computer Vision and Image Understanding* **75(1-2)**:150-164, 1999.

[10] M. Carlin, Selecting feature subsets and distance measures for shape similarity retrieval, submitted to *Pattern Recognition*.

[11] H. Liu & H. Motoda, Feature transformation and subset selection, *IEEE Intelligent Systems* 26-28, March-April 1998.

[12] A. L. Blum & P. Langley, Selection of relevant features and examples in machine learning, *Artificial Intelligence* **97**:245-271, 1997.

[13] R. Kohavi & G. H. John, Wrappers for feature subset selection, *Artificial Intelligence* **97**:273-324, 1997.

[14] D. Koller & M. Sahami, Toward optimal feature selection, *Proceedings of the 13th International Conference on Machine Learning (ML)*, Bari, Italy, pp.284-292, July 1996.

[15] K.V. Mardia, J.T. Kent & J.M. Bibby, *Multivariate analysis*, Academic Press, 1979.

[16] M. Carlin, Which shape features are most important for shape similarity retrieval?, submitted to *Pattern Recognition*.

[17] Ø. D. Trier, A. K. Jain & T. Taxt, Feature extraction methods for character recognition - a survey, *Pattern Recognition* **29(4)**:641-662, 1996.

[18] M. Carlin, Computing geometric moments for objects with an exact polygon representation, *Proceedings of VI-98, Vision Interface* (June 1998), Vancouver, Canada, 319–324.

[19] M. Carlin, Measuring the complexity of non-fractal shapes by a fractal method, accepted for publication in *Pattern Recognition Letters*.