Systematic Methods for Multivariate Data Visualization and Numerical Assessment of Class Separability and Overlap in Automated Visual Industrial Quality Control *

Andreas König, Olaf Bulmahn, and Manfred Glesner
Institute for Microelectronic Systems
Darmstadt University of Technology
Karlstrasse 15, 64287 Darmstadt,Germany
koenig@microelectronic.e-technik.th-darmstadt.de

Abstract

The focus of this work is on systematic methods for the visualization and quality assessment with regard to classification of multivariate data sets. Our novel methods and criteria give in visual and numerical form rapid insight in the principal data distribution, the degree of compactness and overlap of class regions and class separability, as well as information to identify outliers in the data set and trace them back to data acquisition. Assessment by visualization and numerical criteria can be exploited for interactive or automatic optimization of feature generation and selection/extraction in pattern recognition problems. Further, we provide a novel criterion to assess the credibility and reliability of the visualization obtained from high dimensional data projection. Our methods will be demonstrated using data from visual industrial quality control and mechatronic applications.

1 Introduction

Systematic optimization of pattern recognition systems requires reliable criteria for performance evaluation. The objective of this optimization process is the dimensionality reduction and data compression under the constraint of retaining the significant information for discrimination in classification. Based on measurement data this can be achieved by systematic methods for feature selection or feature extraction [1], [5]. Feature extraction comprises systematic methods for strict mathematical treatment as well as heuristic approaches, where transformations for feature extraction are chosen from a multitude of available signal or image

*This work is under grant of the German Federal Ministry of Research and Technology (BMFT) grant number 01 IN 110 B/6 (SIOB). The responsibility for this paper is with the authors

BMVC 1994 doi:10.5244/C.8.19

transformations according to experience and apriori knowledge of the application to accomplish a compact and invariant representation for classification. This often intuitive approach will be denoted in the following as *feature generation* to achieve a distinction to systematic methods of feature selection/extraction. Examples for feature generation methods are coocurrence matrices for texture classification or gradient image and gradient histograms [8]. The parameters of the feature generation methods have to be tuned according to the pattern recognition problem. Tuning of these parameters is commonly carried out according to observations of the operator or system developer for a limited set of characteristic or pathological patterns or images. Generally, it is too time consuming and thus not feasible to assess parameter settings for all patterns or images of training and test sets individually and interactively. Our work provides methods for systematic parameter optimization of feature generation procedures according to complete sample sets based on interactive and non-interactive approaches. For this aim we employ mappings for dimensionality reduction and structure or topology preservation for visualization of high-dimensional data spaces. Furthermore, we develop non-parametric quality measures as criteria for overlap and separability of class regions in feature spaces of arbitrary dimensionality. These methods provide insight in the underlying distribution and data structure and thus allow the transparent and systematic development of a pattern recognition system. In the following we will present examples and applications which were examined as problem instances using our methods. Then we will proceed, introducing and demonstrating our methods. Concluding, we will indicate potential improvements.

2 Benchmarks and Applications

We have generated three two-dimensional Gaussian distributions for a two-class problem with a varying degree of overlap, denoted as $over_1$, $over_2$, and $over_3$ with 588 vectors and two classes each. Further, we generated a Gaussian distribution that is bimodal for class 2, denoted as *bimodal* with 780 vectors. The fifth data set with 1368 vectors, denoted as *banana*, consists of five Gaussian clusters per class, such that a banana shaped non-parametric distribution results for the class regions. These data sets are illustrated in fig. 1. Further we considered the well known *irisdata* [13], with three categories for the different iris species *setosa*, *virginica*, and *versicolor*, characterized by petal and sepal length and width. A train and a test set is available with 75 vectors each. The most important data sets come from visual industrial quality control problems. Within the national collaborative research project SIOB a generic system for visual object inspection in industrial manufacturing is developed. The research objective is the integration of image processing, knowledge processing, pattern recognition, and neural network technology to accomplish a flexible inspection system that is not specialized to a single task or object, but can be easily configured for variations of the inspection task [7]. Rapid system configuration requires a transparent and lucid user interface and a high degree of autonomy, achieved by self-monitoring and backtracking capabilities. Fig. 2 shows the components of the inspection system. The visualization interface and the arrow connections from the functional blocks to the control unit indicate the places were the methods presented in this work are required. Two relevant data sets were chosen from this research work for the demonstration of our methods. The first one was extracted from automatic cut-out objects (s. right part of Fig 2), which were examined for fissures and cracks of the casing. The

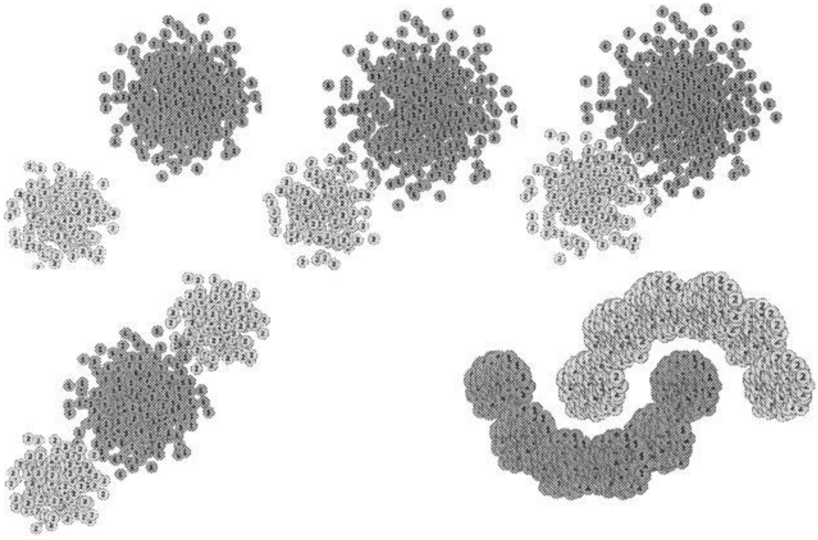

Figure 1: Benchmark data sets *over*$_1$, *over*$_2$, *over*$_3$, *bimodal*, and *banana*

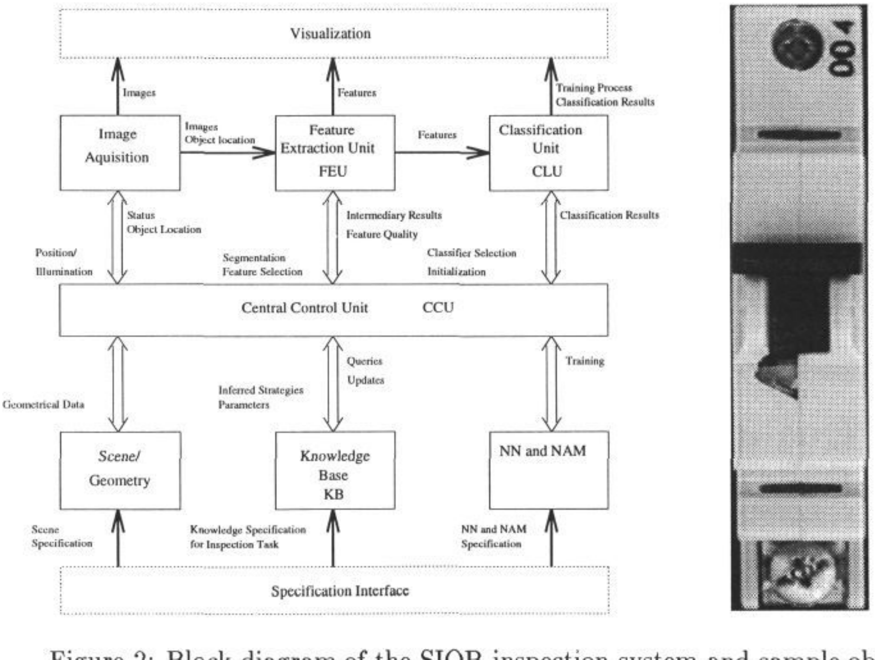

Figure 2: Block diagram of the SIOB inspection system and sample object

resulting data sets comprise 240 38-dimensional vectors with 3 categories and will be denoted as cut_1 to cut_5. The second was obtained from stuffed PCB-inspection, where pins of a large bus connector were checked for correct penetration.

Coocurrence matrices [3] were computed of individual pins and from these seven characteristic moments were determined. This 7-dimensional data set with 185 vectors and 3 categories will be denoted as *pins*. To demonstrate the general validity of our methods, we will show their application to a mechatronic problem. From operating point estimation for stall monitoring of compressors [12] 29-dimensional data sets with 420 vectors per set and four categories for the different operating point regions,denoted as $comp_5$ to $comp_7$, were also regarded in this work.

3 Visualization of Multivariate Data

The visualization of high-dimensional data by projection in two or three dimensions can give considerable insight in the principal data distribution, substructure in data, shape of class regions, and an estimation of the complexity of the class borders for classification problems. Straight forward visualization methods in two or three dimensions include arbitrary selection of components, first principal components [1], first components for parametric or non-parametric scatter matrices [1], and by plotting data points using the values of the discriminance functions as coordinates. The dimensionality of the resulting plot depends on the number of categories [1]. Due to the nature of these methods the structure of the data set is not necessarily preserved, resulting in distorted and unreliable displays. One approach to obtain a topology preserving mapping is the Kohonen self-organizing feature map (SOFM) [6] which is a very useful tool for exploratory data analysis [11]. But due to the quantization carried out by SOFM along with the topology preserving mapping the SOFM is not suited for tracing the position of individual objects from the sample data set. The algorithm of Sammon [10] provides the interesting property of structure preserving mapping of the complete sample set. Sammon's *Non-Linear Mapping* (NLM) proved to be well suited for our applications and we found it an extremely useful tool for feature space evaluation. In the left column of Fig. 3 NLM-mappings of our data sets are displayed. Due to its gradient procedure the NLM-mapping is extremely slow and cannot deal with identical vectors in the data set. The quality of the resulting mapping, which implies its reliability of drawing conclusions from the NLM with regard to the data arrangement in high-dimensional space, is assessed by Sammon using the MSE of distance preservation. We enhance the quality assessment by introducing a topology quality measure for the mapping in the next section. Further, we introduce a specialized mapping algorithm for two-dimensional data displays of multivariate data that is superior to the NLM in approximately two orders of speed.

4 Quality Assessment of Visualization Reliability

For assessment of mapping quality in terms of topology preservation we designed a non-parametric credit assignment scheme based on the nearest neighbor principle. The neighborhood in the original space X and the mapped space Y is examined, concerning the order of the occurrence of nearest neighbors in both spaces of every

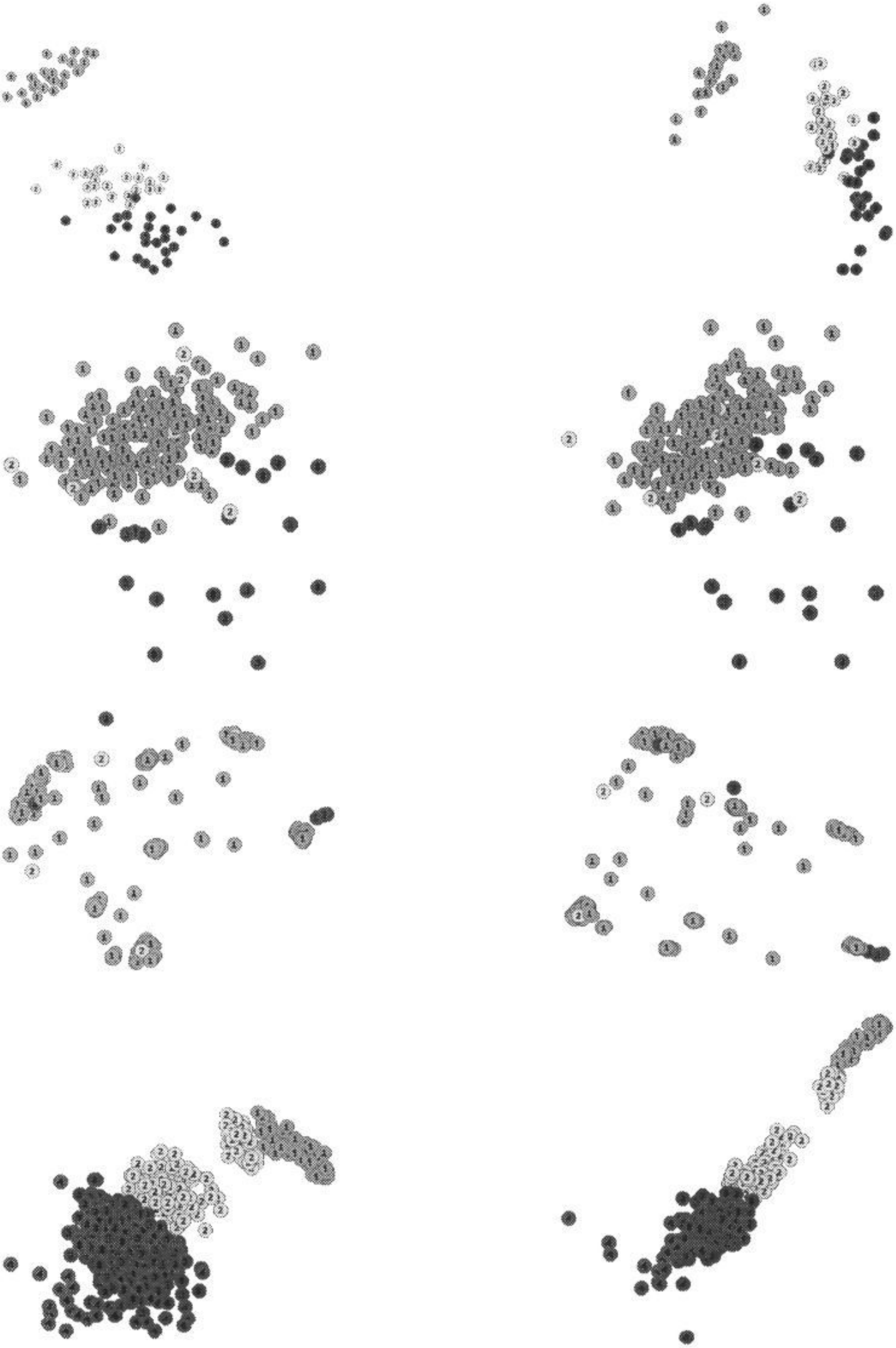

Figure 3: Two-dimensional projections of *irisdata* (top), *pins*, cut_4, and $comp_5$(bottom) by NLM (left column) and Visor (right column)

feature vector. Credits are assigned to a feature vector regarding the nearest neighbors NN_i ($i = 1..n$) according to the following criteria:

$$\begin{array}{lll ll}
3 \text{ credits}, & \text{if} & NN_i \text{ in X} = NN_i \text{ in Y} & \\
2 \text{ credits}, & \text{if} & NN_i \text{ in X} = NN_j \text{ in Y} & j = 1..n, \\
 & & & j \neq i \\
1 \text{ credit}, & \text{if} & NN_i \text{ in X} = NN_k \text{ in Y} & k = n..m \\
0 \text{ credits} & \text{else} & &
\end{array}$$

Here n denotes the number of nearest neighbors under consideration and m denotes the enlarged number of nearest neighbors for the third credit assignment criterion. The maximum quality for a feature vector with this scheme amounts to $3n$. Thus averaging over all feature vectors K of a data set, the quality measure of a mapping q_m is computed by:

$$q_m = \frac{1}{3n \times K} \sum_{i=0}^{K} credits_i \tag{1}$$

This quality measure provides a criterion together with the MSE to assess the quality and reliability of a mapping for a data set. The quality achievable by a mapping is of course dependent a the intrinsic dimensionality of the data [1]. Our criterion performed well and conformed with observations of mapping credibility. Individual quality measure instead of mean quality measure could be employed for the localization of mapping faults.

5 VISOR – a Fast Visualization Algorithm

Motivated by NLMs interesting properties and evident drawbacks, possessing a complexity of O(N^2*L*iterations) our research focused on the development of a fast alternative providing approximately O(N) complexity.
Thus, the Visor-algorithm was devised, based on a pivot-point strategy:

1. Find the pivot-vectors V_1, V_2, V_3 in the original L-dimensional space that provide a convex enclosure of the remaining data points by:
   (a) Compute centroid M of data set in original space
   (b) Determine pivot-vector V_1 such that $d(v_{max} = V_1, M) = max_{i=1}^{K}(d(v_i, M))$
   (c) Determine pivot-vector V_2 such that $d(v_{max} = V_2, V_1) = max_{i=1}^{K}(d(v_i, V_1))$ $\forall v_i \neq V_1$
   (d) Determine pivot-vector V_3 such that $(d(v_{max} = V_3, V_1) = max_{i=1}^{K}(d(v_i, V_1)) \wedge$ $d(v_{max} = V_3, V_1) = max_{i=1}^{K}(d(v_i, V_1)))$ $\forall (v_i \neq V_1 \wedge v_i \neq V_2)$
2. Placement of the corresponding pivot-points P_1, P_2, P_3 of the pivot-vectors V_1, V_2, V_3 in the two-dimensional mapping space
3. Placement of the remaining (N-3) feature vectors employing the pivot-points P_1, P_2, P_3 as follows:
   - Divide the lines $\overline{P_1 P_2}$ and $\overline{P_2 P_3}$ according to the distances to $\hat{P}$ in the original space, introducing division points D_1 and D_2
   - Compute the intersection of the two lines perpendicular to $\overline{P_1 P_2}$ and $\overline{P_2 P_3}$ through the division points D_1 and D_2. This intersection defines the 2D-mapping point of $\hat{P}$

data set	NLM		Visor	
	q_m	MSE	q_m	MSE
irisdata	0.6667	0.0098	0.6711	0.0093
pins	0.6310	0.0099	0.6145	0.0336
cut_1	0.6802	0.0322	0.6647	0.0184
$comp_5$	0.4527	0.0933	0.4207	0.1628

Table 1: Mapping reliability q_m for NLM and Visor. Minor topological deviations lead to the presented values of $q_m \in [0.7, 0.4]$, but the principal structure of the data still can reliably be observed. For values $q_m < 0.3$ the mapping is to be considered as unreliable

The Visor-algorithm preserves no distance value exactly but the same holds for the NLM, as all distances are approximated by the gradient procedure and due to the error criterion small distances are preserved with higher accuracy than large distances, resulting in undesired data clustering. Visor gives a rapid mapping of approximately the same quality as the NLM. Fig. 3 compares the Visor results in the right column with NLM results in the left column. In some cases of our work observed quality was less than in others beyond the quality obtained with the NLM. For the majority of cases Visor gives a satisfactory alternative for the two-dimensional visualization of multivariate data, providing $\approx$ O(N) complexity and thus a speed advantage of ≈ 100 with regard to NLM (s. Table 1). Recently, we found another method for 2D-visualization, based on the exact preservation of $2(M-3)$ distances and using a *Minimal Spanning Tree (MST)* approach [9]. We will compare this method with NLM and VISOR in future work. A further advantage is, that identical vectors in the data set are no obstacle for Visor which is a great benefit for the analysis and visualization of SOFMs. In case of a low mapping quality indicated by our measure q_m and the MSE and high reliability demands a complementing NLM run can be carried out for comparison with the fast preview of Visor.

6 Assessment of Class Regions Overlap

In addition to the visual information conveyed to the human operator of a pattern recognition system, which is in our case the visual quality inspection system (s. Fig 2), numerical measures are required for systematic and automatic parameter optimization in the feature generation process during system configuration. In this section we present a family of three non-parametric quality criteria measuring the overlap of class regions. The criteria of course can also serve for systematic feature selection/extraction computing overlap in the original and in the reduced space, indicating overlap increase/decrease by the selection/extraction process. We employ a nearest neighbor scheme largely motivated by the ideas of *non-parametric scatter matrices* and the *edited nearest neighbor* approach [1]. The n-nearest neighbors of every feature vector and their corresponding class affiliation are employed for overlap assessment. Overlap degree is determined for an individual feature vector

$\hat{P}$ by:

$$q_o = \frac{\sum_{i=1}^{n} q_{x_{NN_i}} + \sum_{i=1}^{n} k_i}{2\sum_{i=1}^{n} k_i} \; ; \; k_i = 1 - \frac{d_{NN_i}}{d_{NN_n}} \; ; \; q_{x_{NN_i}} = \begin{cases} k_i & : \; w_x = w_i \\ -k_i & : \; w_x \neq w_i \end{cases} \qquad (2)$$

Here k_i denotes the weighting factor for the position of the i-th nearest neighbor NN_i, d_{NN_i} denotes the distance between $\hat{P}$ and NN_i, d_{NN_n} denotes the distance between $\hat{P}$ and the most distant nearest neighbor NN_n, $q_{x_{NN_i}}$ denotes the overlap for $\hat{P}$ with regard to NN_i, w_x denotes the class affiliation of $\hat{P}$, and w_i denotes the class affiliation of NN_i. We define an influence of every NN_i decaying with the position in the nearest neighbor list, thus the weighting factors k_i are defined such that the influence decreases to zero for NN_n. The quality q_x is increased $(+k_i)$ for every NN_i with $w_x = w_i$ and decreased $(-k_i)$ for $w_x \neq w_i$. The range of $q_{x_{iNN}}$ is therefore between $\sum_{i=1}^{n} k_i$ and $-\sum_{i=1}^{n} k_i$. Normalization of the overlap measure q_o in the interval [0,1] is obtained by adding $\sum_{i=1}^{n} k_i$ to $\sum_{i=1}^{n} q_{x_{iNN}}$ and dividing the resulting sum by $2\sum_{i=1}^{n} k_i$.
Neglecting the weighting factor k_i for each NN_i reduces the overlap computation to:

$$q_o' = 1 - \frac{\sum_{i=1}^{n} d_{NN_{i,c}}}{\sum_{i=1}^{n} d_{NN_i}}. \qquad (3)$$

Here $\sum_{i=1}^{n} d_{NN_{i,c}}$ denotes the sum of all distances to nearest neighbors with $w_x \neq w_i$.
Additionally, neglecting the distances $d(NN_i)$ in the overlap computation leads to:

$$q_o'' = 1 - \frac{n_c}{n} \qquad (4)$$

where n_c denotes the number of nearest neighbors with $w_x \neq w_i$. During overlap degree computation a hypothesis for $\hat{P}$ being an isolated vector (outlier) or a member of a small isolated group is checked. The hypothesis for an outlier is considered true, if for all n nearest neighbors $w_x \neq w_i$ holds. The hypothesis for an isolated group of size v is considered true if for v nearest (closest) neighbors $w_x \neq w_i$ holds and for the remaining $n - v$ neighbors $w_x \neq w_i$ holds. This information complements the visual impression of outliers and isolated groups for a human observer, providing additional relevant criteria for automatic system optimization. In Table 2 overlap measures are computed for benchmark data sets and compared. For application data sets only the most exact measure q_o was computed.

7 Assessment of Separability

In addition to the numerical assessment of overlap a measure for the complexity of the class boundary is highly desirable. This complexity measure gives the separability of the pattern recognition problem. Though overlap and separability measure are related they have their distinct meaning. As can be seen from Table 2 for *banana*, no overlap does not imply easy (linear) separability. Thus we propose

data set	q_o	q_o'	q_o''		q_s
$over_1$	1.0	1.0	1.0		0.9966
$over_2$	0.9911	0.9931	0.9928		0.9820
$over_3$	0.9753	0.9794	0.9782		0.9575
banana	1.0	1.0	1.0		0.9927
bimodal	0.9909	0.9898	0.9900		0.9807
irisdata	0.9156	–	–		0.8666
pins	0.9527	–	–		0.8919
cut_4	0.8330	–	–		0.7901
$comp_5$	0.9788	–	–		0.9548

Table 2: Overlap degree and separability for benchmark data sets

two means for discerning the separability or the class boundary complexity. As a visual means we sketch a piecewise linear approximation of the actual class boundary in our projection plot (Voronoi diagram, s. Fig. 4), thus giving the user an idea of class regions and class boundaries. As a numerical means we propose to employ the piecewise linear boundary generated by a nearest neighbor approach as an approximate separability measure. The number of line segments of the piecewise linear boundary are proportional to the complexity of the class boundary and thus to the separability. Reducing nearest neighbor methods, e.g. *condensed/reduced nearest neighbors*[4][2], retain only vectors close to class boundaries for classification. These vectors approximate the class boundary by a Voronoi tessellation. Thus, the number of vectors retained after the reduction is itself a measure of separability. This can be expressed denoting M as the initial sample set size and K as the number of selected or retained vectors by:

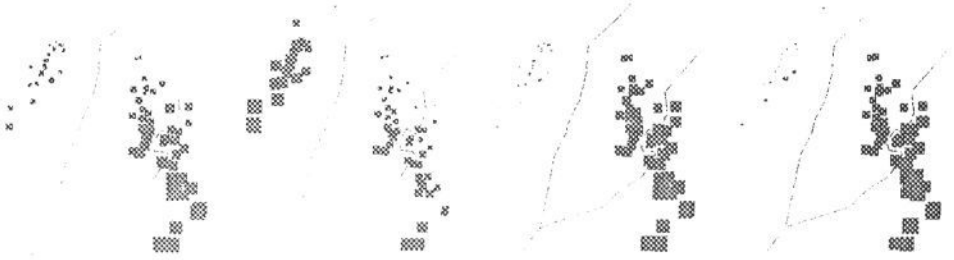

Figure 4: Sketch of class boundaries in the four components of *irisdata* (test set) by Voronoi plot. Observation of feature values within class regions can serve for visual selection of significant features

$$q_s = \frac{M - K}{M} \tag{5}$$

Obviously, q_s tends to one for well separable data sets and to zero for completely overlapping sets. The measure definitely is influenced by class overlap, but in contrast to the previous measures it gives no hint on outliers or isolated groups (s. Table 2).

8 Conclusion and Future Work

We have introduced and demonstrated methods for multivariate data visualization and non-parametric overlap assessment and separability measure. These tools offer support for interactive as well as for automatic optimization of pattern recognition systems. we have integrated our methods in the inspection system prototype for visual industrial quality control (s. Fig 2) and have achieved encouraging results for "in-the-loop" system optimization. In current work we examine the modification of NLM by introducing our quality measure q_m into the gradient procedure to obtain a topologically correct mapping. Further, we will investigate our criteria with regard to linearity and absolute interpretation independent of data sets.

References

[1] K. Fukunaga. *Introduction to Statistical Pattern Recognition.* ACADEMIC PRESS, INC. Harcourt Brace Jovanovich, Publishers Boston San Diego New York London Sydney Tokyo Toronto, 1990.

[2] G. W. Gates. The Reduced Nearest Neighbour Rule. In *IEEE Transactions on Information Theory, vol. IT-18*, pages 431 – 433, 1972.

[3] R. M. Haralick, K. Shanmugan, and I. Dinstein. Textural Features for Image Classification. In *IEEE Transactions on Systems, Man, and Cybernetics, vol. SMC-3(11)*, pages 610 –, 1973.

[4] P. E. Hart. The Condensed Nearest Neighbour Rule. In *IEEE Transactions on Information Theory, vol. IT-14*, pages 515 – 516, 1968.

[5] J. Kittler. *Feature Selection and Extraction.* ACADEMIC PRESS, INC. Tzai. Y. Young King Sun-Fu, Publishers Orlando San Diego New York Austin London Montreal Sydney Tokyo Toronto, 1986.

[6] T. Kohonen. *Self-Organization and Associative Memory.* Springer Verlag Berlin Heidelberg London Paris Tokyo Hong Kong, 1989.

[7] A. König, H. Genther, and M. Glesner et.al. A Generic Dynamic Inspection System for Visual Object Inspection and Industrial Quality Control. In *Proceedings of the International Joint Conference on Neural Networks IJCNN-'93, Nagoya, Japan*, volume II, pages 1243–1246. IEEE, 1993, ISBN 0-7803-1421-2.

[8] A. Korn. Toward a symbolic representation of intensity changes in images. In *IEEE Transactions on Pattern Analysis and Machine Intelligence PAMI-10*, pages 610–625, 1988.

[9] R. C. T. Lee, J. R. Slagle, and H. Blum. A Triangulation Method for the Sequential Mapping of Points from N-Space to Two-Space. In *IEEE Transactions on Computers C-26*, pages 288 – 292, 1977.

[10] J. W. Sammon. A Nonlinear Mapping for Data Structure Analysis. In *IEEE Transactions on Computers C-18*, pages 401–409, 1969.

[11] A. Ultsch and H. P. Siemon. Exploratory Data Analysis: Using Kohonen Networks on Transputers. In *Interner Bericht Nr. 329 Universität Dortmund, Dezember 1989*, 1989.

[12] H. Wang, D. K. Hennecke, A. König, P. Windirsch, and M. Glesner. An Approach to the Stall Monitoring in a Single Stage Axial Compressor. In *29th AIAA/SAE/ASME/ASEE Joint Propulsion Conference; Monterey, CA.*, 1993.

[13] C. T. Zahn. Graph-Theoretical Methods for Detecting and Describing Gestalt Clusters. In *IEEE Transactions on Computers, vol. C-20*, pages 68 – 86, 1971.