PERGAMON

Pattern Recognition 33 (2000) 617–634

PATTERN RECOGNITION
THE JOURNAL OF THE PATTERN RECOGNITION SOCIETY

www.elsevier.com/locate/patcog

# A theory of proximity based clustering: structure detection by optimization

Jan Puzicha[a,]*, Thomas Hofmann[b], Joachim M. Buhmann[a]

[a]*Institut für Informatik III, University of Bonn, Bonn, Germany*
[b]*Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA, USA*

Received 15 March 1999

## Abstract

In this paper, a systematic optimization approach for *clustering proximity or similarity data* is developed. Starting from fundamental invariance and robustness properties, a set of axioms is proposed and discussed to distinguish different cluster compactness and separation criteria. The approach covers the case of sparse proximity matrices, and is extended to nested partitionings for hierarchical data clustering. To solve the associated optimization problems, a rigorous mathematical framework for *deterministic annealing* and *mean-field approximation* is presented. Efficient optimization heuristics are derived in a canonical way, which also clarifies the relation to stochastic optimization by Gibbs sampling. Similarity-based clustering techniques have a broad range of possible applications in computer vision, pattern recognition, and data analysis. As a major practical application we present a novel approach to the problem of unsupervised texture segmentation, which relies on statistical tests as a measure of homogeneity. The quality of the algorithms is empirically evaluated on a large collection of Brodatz-like micro-texture Mondrians and on a set of real–word images. To demonstrate the broad usefulness of the theory of proximity based clustering the performances of different criteria and algorithms are compared on an information retrieval task for a document database. The superiority of optimization algorithms for clustering is supported by extensive experiments. © 2000 Pattern Recognition Society. Published by Elsevier Science Ltd. All rights reserved.

*Keywords:* Clustering; Proximity data; Similarity; Deterministic annealing; Texture segmentation; Document retrieval

## 1. Introduction

*Data clustering* is one of the core methods for numerous tasks in pattern recognition, exploratory data analysis, computer vision, machine learning, data mining, and in many other related fields. In a rather informal sense, the goal of clustering is to partition a given set of data into homogeneous groups. *Cluster homogeneity* is thus the central notion which needs to be formalized in order to give data clustering a precise meaning. In this paper, we focus on homogeneity measures which are defined in terms of pairwise similarities or dissimilarities between data entities or *objects*. The underlying data is usually called *similarity* or *proximity data* [1]. In proximity data, the elementary measurements are comparisons between two objects of a given data set. This data format differs from *vectorial data* where each measurement corresponds to a certain 'feature' evaluated at an external scale. Notice however, that pairwise dissimilarities can be canonically generated from vectorial data whenever a distance function or metric is available.

There exist numerous approaches to the data clustering problem, only to mention some of the most important: central clustering or vector quantization (e.g., the K-means algorithm [2]), linkage or agglomerative methods [3], mixture models [4], fuzzy clustering [5], and competitive learning [6,7]. These and other approaches offer a variety of clustering algorithms, and,

*Corresponding author.

*E-mail addresses:* {jan, jb}@cs.uni-bonn.de (J. Puzicha), hofmann@ai.mit.edu (T. Hofmann)

PII: S0031-3203(99)00076-X

more fundamentally, differ with respect to the framework in which data clustering is formulated. Typically, there are limitations to which kind of data a method can be applied. It is a commonly expressed opinion [1] that the application of partitional clustering methods which are based on explicit objective functions are only suitable for vectorial data. In addition, optimization methods are supposed to be inherently non-hierarchical. In this paper, a theory is outlined which advocates a formulation of proximity-based data clustering as a combinatorial optimization problem. The theory proposes solutions to the two fundamental problems: (i) the specification of suitable objective functions, and (ii) the derivation of efficient optimization algorithms.

To address the modeling problem, an *axiomatization* of clustering objective functions based on fundamental *invariance* and *robustness* properties is presented (Section 2). As will be rigorously shown, the proposed axioms imply restrictions on both the way cluster homogeneities are calculated from pairwise dissimilarities as well as the final combination of contributions from different clusters. These ideas are extended to cover two generalizations: *clustering with sparse or incomplete proximity data* and *hierarchical clustering*. These extensions are indispensable for the analysis of large data sets, since it is in general prohibitive (and also unnecessary due to redundancy) to exhaustively perform all $N^2$ pairwise comparisons for $N$ objects. Moreover, in large-scale applications group structure typically occurs at different resolution levels, which strongly favors hierarchical partitioning schemes.

Once a suitable objective function has been identified, in principle any known optimization technique could be applied to find optimal solutions. Yet, the $\mathscr{NP}$-hardness of most data partitioning problems renders the application of exact methods for large-scale problems intractable [8]. Therefore, heuristic optimization techniques are promising candidates to find at least approximate clustering solutions. In particular, stochastic optimization (Section 3) offers robustness and matches the peculiarities of data analysis problems [9]. Two closely related methods will be derived: a Monte Carlo algorithm known as the Gibbs sampler [10] and a deterministic variant known as *mean-field annealing* [11]. Both approaches rely on the introduction of a computational temperature, and offer a number of advantages: (i) they are general enough to cover all clustering objective functions, (ii) they yield *scalable* algorithms (in terms of the complexity-quality tradeoff), and (iii) the temperature defines a 'natural' resolution scale for hierarchical clustering problems [12]. The theory and the derived algorithms are tested and validated in two application areas: unsupervised segmentation of textured images [13,14] and information retrieval in document databases.

## 2. Axiomatization for clustering objective functions

Assume the data is given in the form of a proximity matrix $\mathbf{D} \in \mathbb{R}^{N^2}$ with entries $D_{ij}$ quantifying the dissimilarity between objects $\mathbf{o}_i$ and $\mathbf{o}_j$ from a given domain of $N$ objects. Furthermore, assume the number of clusters $K$ to be fixed. A Boolean representation of data partitionings is introduced in terms of assignment matrices $\mathbf{M} \in \mathscr{M}_{N,K}$, where

$$\mathscr{M}_{N,K} = \left\{ \mathbf{M} \in \{0, 1\}^{N \times K} \colon \sum_{\nu=1}^{K} M_{i\nu} = 1,\ 1 \leqslant i \leqslant N \right\}. \tag{1}$$

$M_{i\nu}$ is an indicator variable for an assignment of object $\mathbf{o}_i$ to cluster $\mathscr{C}_\nu$, hence $M_{i\nu} = 1$ if and only if object $\mathbf{o}_i$ belongs to cluster $\mathscr{C}_\nu$. The assignment constraints in the definition of $\mathscr{M}_{N,K}$ assure that the data assignment is complete and unique.

### 2.1. Elementary axioms

General principles of proximity-based clustering are expressed by the following axioms:

**Definition 1** (Clustering criterion). A cost function $\mathscr{H}_{N,K} \colon \mathscr{M}_{N,K} \times \mathbb{R}^{N^2} \to \mathbb{R}$ is a *clustering criterion* if the following set of axioms is fulfilled:

**Axiom 1** (Permutation invariance). $\mathscr{H}_{N,K}$ is invariant with respect to permutations of (a) object indices and (b) label indices. More precisely, for all $\mathbf{D} \in \mathbb{R}^{N^2}$, $\mathbf{M} \in \mathscr{M}_{N,K}$, permutations $\pi$ over $\{1, \dots, N\}$, and $\bar{\pi}$ over $\{1, \dots, K\}$:

$$\mathscr{H}_{N,K}(\mathbf{M};\, \mathbf{D}) = \mathscr{H}_{N,K}(\mathbf{M}^{\pi}_{\bar{\pi}};\, \mathbf{D}^{\pi}_{\pi}),$$

where $\mathbf{A}^{\pi}_{\bar{\pi}}$ is obtained from $\mathbf{A}$ by row permutation with $\pi$ and column permutation with $\bar{\pi}$.

**Axiom 2** (Monotonicity). For all $\mathbf{D} \in \mathbb{R}^{N^2}$, $\mathbf{M} \in \mathscr{M}_{N,K}$, $\triangle d \in \mathbb{R}^+$, and pairs of data indices $(i, j)$:

$$\sum_{\nu=1}^{K} M_{i\nu} M_{j\nu} = \left\{ \begin{smallmatrix} 1 \\ 0 \end{smallmatrix} \right\} \Rightarrow \mathscr{H}_{N,K}(\mathbf{M};\, \mathbf{D}) \left\{ \begin{smallmatrix} \leqslant \\ \geqslant \end{smallmatrix} \right\} \mathscr{H}_{N,K}(\mathbf{M};\, \mathbf{D}^{ij}), \tag{2}$$

where $\mathbf{D}^{ij}$ is obtained from $\mathbf{D}$ by the local modification $D_{ij} \to D_{ij} + \triangle d$.

$\mathscr{H}_{N,K}$ is a *strict* clustering criterion if for each $(i, j)$ at least one of the inequalities in (2) is strict.

Axiom 1 prevents that the quality of a data partitioning depends on additional information hidden in the data labels or cluster labels. Axiom 2 states that increasing the dissimilarity between objects in the same cluster can never be advantageous. The same is true for decreasing the dissimilarity between objects belonging to different clusters.

To further limit the functional dependency on proximities, an additivity axiom is introduced. Additivity reduces the noise sensitivity of a clustering criterion by averaging, as opposed to other approaches which completely discard the dissimilarity values and only keep their order relation (e.g., single linkage).

**Definition 2** (Additivity). A clustering criterion $\mathscr{H}_{N,K}$ is *additive* if it has the following functional form:[1]

$$\mathscr{H}_{N,K}(\mathbf{M}; \mathbf{D}) = \sum_{i=1}^{N} \sum_{j=1, j\neq i}^{N} \psi_{N,K}(i, j, D_{ij}, \mathbf{M}).$$

$\psi_{N,K}: \{1, \dots, N\}^2 \times \mathbb{R} \times \mathscr{M}_{N,K} \to \mathbb{R}$ is called the *contribution* function. Furthermore, for $M_{i\alpha} = 1$ and $M_{j\nu} = 1$, $\psi_{N,K}(i, j, D_{ij}, \mathbf{M})$ does not depend on assignments to clusters $\mu \neq \alpha, \nu$.

The contribution function $\psi_{N,K}$ of an additive clustering criterion is in fact further restricted by Axioms 1 and 2.

**Proposition 1.** *Every additive clustering criterion can be rewritten as a combination of* (**D**-*monotone*) *intra- and inter-cluster contribution functions*, $\psi^{(1)}$ *and* $\psi^{(2)}$ *respectively*,

$$\mathscr{H}_{N,K}(M; \mathbf{D}) = \sum_{\nu=1}^{K} \sum_{\substack{i,j=1,\\ i\neq j}}^{N} M_{i\nu} \left[ M_{j\nu}\psi^{(1)}_{N,K}(D_{ij}, n_\nu) - \sum_{\substack{\mu=1\cdot\\ \mu\neq\nu}}^{K} M_{j\mu}\psi^{(2)}_{N,K}(D_{ij}, n_\nu, n_\mu)\right].$$

*Here* $n_\nu = \sum_{i=1}^{N} M_{i\nu}$ *denotes the size of cluster* $\nu$.

Each additive clustering criterion can thus be linearly decomposed into one part measuring intra-cluster compactness (to be minimized) and a second part measuring inter-cluster separation (to be maximized). A proof of Proposition 1 is given in the Appendix.

### 2.2. Invariance axioms

While Axioms 1 and 2 ensure elementary requirements for a clustering criterion, and additivity narrows the focus to a particular simple class of objective functions, the following invariance and robustness properties have to be considered as the core of the proposed axiomatization. Assuming $N$ and $K$ to be fixed, the explicit dependency in our notation is dropped whenever possible.

**Definition 3** (Invariance). An objective functions $\mathscr{H}$ is invariant with respect to linear data transformations, if the following set of axioms is fulfilled:

**Axiom 3** (Scale invariance). For all $\mathbf{D} \in \mathbb{R}^{N^2}$, $\mathbf{M} \in \mathscr{M}$, $c \in \mathbb{R}^+$:

$$\mathscr{H}(\mathbf{M}; c\mathbf{D}) = c\mathscr{H}(\mathbf{M}; \mathbf{D}).$$

**Axiom 4** (Shift invariance). For all $\mathbf{D} \in \mathbb{R}^{N^2}$, $\mathbf{M} \in \mathscr{M}$, $\triangle d \in \mathbb{R}$:

$$\mathscr{H}(\mathbf{M}; \mathbf{D} + \triangle d) = \mathscr{H}(\mathbf{M}; \mathbf{D}) + N\triangle d.$$

Axiom 3 ensures that rescaling of the data solely rescales the cost function. A scale-invariant criterion has the advantage not to introduce an implicit bias towards a particular data scale. Axiom 4 is crucial for data that is only meaningful on an interval scale and not on an absolute or ratio scale. Invariant clustering criteria are thus non-committal with respect to scale and origin of the data, a property which is especially useful in applications where these quantities are not a priori known.[2] For additive clustering criteria scale invariance restricts $\psi^{(1)}$ and $\psi^{(2)}$ in Proposition 1 to a linear data dependency. The first argument of $\psi^{(1,2)}$ is therefore dropped with the understanding $\psi^{(1,2)}(D_{ij}, \cdot) = D_{ij}\psi^{(1,2)}(\cdot)$. The number of additive clustering criteria is further reduced by the shift invariance axiom. For cluster compactness measures the following result is obtained.

**Proposition 2.** *For every invariant additive clustering criterion with* $\psi^{(2)} = 0$ (*intra-cluster compactness measure*), *the contribution function* $\psi^{(1)}$ *can be written as*

$$\psi^{(1)}(n_\nu) = \lambda\frac{1}{n_\nu - 1} + (1-\lambda)\frac{1}{n_\nu(n_\nu - 1)}, \quad \lambda \in \mathbb{R}.$$

The number of admissible weighting functions $\psi^{(2)}$ is less significantly reduced by the invariance property. Therefore, a special class of contribution functions is considered which possess a natural decomposition.

**Definition 4** (Decomposability). An inter-cluster contribution function $\psi^{(2)}$ is *decomposable*, if there exists a function $f(n_\nu)$ such that either

$$f(n_\nu) = \sum_{\mu\neq\nu} n_\mu\psi^{(2)}(n_\nu, n_\mu) \text{ or } f(n_\nu) = \sum_{\mu\neq\nu} n_\mu\psi^{(2)}(n_\mu, n_\nu).$$

**Proposition 3.** *For every invariant additive clustering criterion with* $\psi^{(1)} = 2C$ *and decomposable* $\psi^{(2)}$, *the contribution*

[1] Self-dissimilarities $D_{ii}$ are excluded for simplicity. All objective functions can however be easily modified, if diagonal contributions should be included.

[2] Both invariance axioms can in fact be weakened with respect to additional multiplicative or additive constants, but this does not result in qualitatively different criteria.

*function can be written as an affine combination*

$$\psi^{(2)}(n_\nu, n_\mu) = \sum_{r=1}^{7} \lambda_r \psi_r(n_\nu, n_\mu) - C,$$

*with an additive offset*[3] $C = 1/(N-1)$ *and the following elementary functions and its* $(\nu, \mu)$*-symmetric counterparts* $(\psi_5, \psi_6, \psi_7)$:

$$\psi_1 = \frac{1}{N - n_\nu}, \quad \psi_2 = \frac{1}{(K-1)n_\nu},$$

$$\psi_3 = \frac{N}{K(K-1)(N-n_\nu)n_\nu}, \quad \psi_4 = \frac{N}{K(K-1)n_\nu n_\mu}.$$

Proofs of both propositions are given in the appendix. For simplicity, we focus on the special case of symmetric proximity matrices in the sequel, for which one may set $\lambda_5 = \lambda_6 = \lambda_7 = 0$ without loss of generality. In summary, we have derived 2 elementary intra-cluster compactness criteria

$$\mathscr{H}^{\mathrm{pc1}}(\mathbf{M}; \mathbf{D}) = \sum_{\nu=1}^{K} n_\nu D_\nu, \quad \mathscr{H}^{\mathrm{pc2}}(\mathbf{M}; \mathbf{D}) = \sum_{\nu=1}^{K} \frac{N}{K} D_\nu, \tag{3}$$

where

$$D_\nu = \frac{\sum_{i=1}^{N}\sum_{j=1, j\neq i}^{N} M_{i\nu} M_{j\nu} D_{ij}}{\sum_{i=1}^{N}\sum_{j=1, j\neq i}^{N} M_{i\nu} M_{j\nu}} \tag{4}$$

is the average dissimilarity in cluster $\mathscr{C}_\nu$. Moreover, by restriction to decomposable contribution functions and symmetric data we have derived 4 elementary inter-cluster separation criteria:

$$\mathscr{H}^{\mathrm{ps1a}}(\mathbf{M}; \mathbf{D}) = -\sum_{\nu=1}^{K} n_\nu \frac{\sum_{\mu=1,\mu\neq\nu}^{K} D_{\nu\mu}}{K-1},$$
$$\mathscr{H}^{\mathrm{ps1b}}(\mathbf{M}; \mathbf{D}) = -\sum_{\nu=1}^{K} n_\nu \frac{\sum_{\mu=1,\mu\neq\nu}^{K} n_\mu D_{\nu\mu}}{N-n_\nu}, \tag{5}$$

$$\mathscr{H}^{\mathrm{ps2a}}(\mathbf{M}; \mathbf{D}) = -\sum_{\nu=1}^{K} \frac{N}{K} \frac{\sum_{\mu=1,\mu\neq\nu}^{K} D_{\nu\mu}}{K-1},$$
$$\mathscr{H}^{ps2b}(\mathbf{M}; \mathbf{D}) = -\sum_{\nu=1}^{K} \frac{N}{K} \frac{\sum_{\mu=1,\mu\neq\nu}^{K} n_\mu D_{\nu\mu}}{N-n_\nu}, \tag{6}$$

where

$$D_{\nu\mu} = \frac{\sum_{i=1}^{N}\sum_{j=1}^{N} M_{i\nu} M_{j\mu} D_{ij}}{\sum_{i=1}^{N}\sum_{j=1}^{N} M_{i\nu} M_{j\mu}} \tag{7}$$

denotes the average inter-cluster dissimilarity. The most fundamental distinction between the different criteria concerns the weighting of the average cluster compactness ($D_\nu$) or separation ($\sum_{\mu\neq\nu} n_\mu D_{\nu\mu}$) with either the cluster size ($\mathscr{H}^{\mathrm{pc1}}$, $\mathscr{H}^{\mathrm{ps1a}}$, $\mathscr{H}^{\mathrm{ps1b}}$) or a constant ($\mathscr{H}^{\mathrm{pc2}}$, $\mathscr{H}^{\mathrm{ps2a}}$, $\mathscr{H}^{\mathrm{ps2b}}$). The second distinction concerns the way the average cluster separation is computed for separation measures: either these averages are performed by pooling all dissimilarities together ($\mathscr{H}^{\mathrm{ps1b}}$, $\mathscr{H}^{\mathrm{ps2b}}$), or by a two-stage procedure which first calculates averages for every pair of clusters ($D_{\nu\mu}$) and combines those with a constant weight ($\mathscr{H}^{\mathrm{ps1a}}$, $\mathscr{H}^{\mathrm{ps2a}}$). These differences are crucial for the robustness properties, formalized in the following axioms.

### 2.3. Robustness axioms

The following set of axioms is concerned with the sensitivity of the quality criterion with respect to perturbations of the proximities. Since inaccurate (e.g., quantized) or noisy measurements are common, robustness is a key issue. Two different notions of robustness are distinguished:

**Axiom 5** (Weak robustness). A family of objective functions $\mathscr{H} = (\mathscr{H}_{N,K})_{N\in\mathbb{N}}$ is robust in the weak sense if for all $\triangle d \in \mathbb{R}$, $\varepsilon \in \mathbb{R}^+$ there exists $N_0 \in \mathbb{N}$ such that for all $N \geqslant N_0$, $M \in \mathscr{M}_{N,K}$, $\mathbf{D} \in \mathbb{R}^{N^2}$, and pairs of data indices $(i, j)$:

$$\frac{1}{N}|\mathscr{H}_{N,K}(\mathbf{M}|\mathbf{D}) - \mathscr{H}_{N,K}(\mathbf{M}|\mathbf{D}^{ij})| < \varepsilon, \tag{8}$$

where $\mathbf{D}^{ij}$ is defined as in Axiom 2.

**Axiom 6** (Strong robustness). $\mathscr{H}$ is robust in the strong sense if condition (8) holds for all $\mathbf{D}^i \in \mathbb{R}^{N^2}$ defined by $\mathbf{D}^i = \mathbf{D} + \triangle d(\delta_{ik} + \delta_{il} - \delta_{ik}\delta_{il})_{k,l}$.

More intuitively, robustness assures that single measurements (weak robustness) or measurements belonging to a single object (strong robustness) do not have a macroscopic influence on the costs of a configuration. The robustness properties of the invariant criteria are summarized without proof in the following table:

| | $\mathscr{H}^{\mathrm{pc1}}$ | $\mathscr{H}^{\mathrm{pc2}}$ | $\mathscr{H}^{\mathrm{ps1a}}$ | $\mathscr{H}^{\mathrm{ps1b}}$ | $\mathscr{H}^{\mathrm{ps2a}}$ | $\mathscr{H}^{\mathrm{ps2b}}$ |
|---|---|---|---|---|---|---|
| Weak robustness | Yes | No | Yes | Yes | No | Yes |
| Strong robustness | Yes | No | No | No | No | No |

We emphasize the most remarkable facts: (i) $\mathscr{H}^{\mathrm{pc1}}$ is the only criterion which fulfills the strong robustness axiom, (ii) no invariant inter-cluster separation criterion is robust in the strong sense, (iii) $\mathscr{H}^{\mathrm{ps2b}}$ is the only criterion

[3] The constant $C$ is a technical requirement to obtain the correct sign for the additive offset in Axiom 4, it will be dropped in the sequel.

with constant cluster weights being robust in the weak sense. All cluster separation criteria lack strong robustness. The reason for this are configurations with only one large (macroscopic) cluster and $(K-1)$ small (microscopic) clusters where the number of inter-cluster dissimilarities scales only with $\mathcal{O}(N)$ compared to the total number scaling with $\mathcal{O}(N^2)$. Strong robustness can be obtained by restricting $\mathcal{M}$ to the following sets of admissible solutions: (1) at least two macroscopic clusters for $\mathcal{H}^{\mathrm{ps1b}}$ and (2) $K$ macroscopic clusters for $\mathcal{H}^{\mathrm{ps1a}}$ and $\mathcal{H}^{\mathrm{ps2b}}$. These considerations yield the following ranking of invariant clustering criteria with respect to their asymptotic robustness properties:

$$\mathcal{H}^{\mathrm{pc1}} \succ \mathcal{H}^{\mathrm{ps1b}} \succ \{\mathcal{H}^{\mathrm{ps1a}}, \mathcal{H}^{\mathrm{ps2b}}\} \succ \{\mathcal{H}^{\mathrm{ps2a}}, \mathcal{H}^{\mathrm{pc2}}\}. \tag{9}$$

By the axiomatization, the criterion $\mathcal{H}^{\mathrm{pc1}}$ thus is clearly distinguished from all other additive criteria. Among the cluster separation measures $\mathcal{H}^{\mathrm{ps1b}}$ has been identified as the most promising candidate due to its robustness properties.

Interestingly, there is an intrinsic connection to the $K$-means objective function for central clustering. Assume that the data were generated from a vector space representation $\mathbf{v}_i \in \mathbb{R}^d$ by $D_{ij} = (\mathbf{v}_i - \mathbf{v}_j)^2$, then $\mathcal{H}^{\mathrm{pc1}}(\mathbf{M}, \mathbf{D}) \cong \mathcal{H}^{\mathrm{km}}(\mathbf{M}, \mathbf{D})$, with

$$\mathcal{H}^{\mathrm{km}}(\mathbf{M}, \mathbf{D}) = \sum_{\nu=1}^{K} \sum_{i=1}^{N} M_{i\nu}(\mathbf{v}_i - \mathbf{y}_\nu)^2, \tag{10}$$

and the usual centroid definition

$$\mathbf{y}_\nu = \frac{\sum_{j=1}^{N} M_{j\nu}\mathbf{v}_j}{\sum_{j=1}^{N} M_{j\nu}}.^{4}$$

Moreover, there exists an intimate relation to Ward's agglomerative clustering method [15]. If the distance between a pair of clusters is defined by the cost increase after merging both clusters, any objective function $\mathcal{H}$ is heuristically minimized by greedy merging starting from singleton clusters. In case of $\mathcal{H}^{\mathrm{pc1}}$ this procedure exactly yields Ward's method. It was often conjectured that Ward's method depends on the definition of centroids and the usage of a squared error measure, however, as demonstrated, Ward's method can be understood as a greedy algorithm to minimize $\mathcal{H}^{\mathrm{pc1}}$, which does not involve centroids or any other geometrical concepts.

It is worth mentioning that the graph partitioning objective function defined by

$$\mathcal{H}^{\mathrm{gp}}(\mathbf{M}; \mathbf{D}) = \sum_{\nu=1}^{K} \sum_{i=1}^{N} \sum_{j=1,\, j\neq i}^{N} M_{i\nu} M_{j\nu} D_{ij}, \tag{11}$$

is scale invariant and robust in the strong sense, but *not* shift invariant. $\mathcal{H}^{\mathrm{gp}}$ has been utilized for data analysis in the operations research context [16] and also for texture segmentation [17]. The missing shift invariance is obvious: in the limit of $\Delta d \to \infty$ the optimal solution is an equipartitioning, in the limit of $\Delta d \to -\infty$ the configuration inevitably collapses into one cluster. A 'good' balance between positive and negative contribution is necessary to avoid this type of degeneration [17] (see Section 4). This is a consequence of the ratio scale interpretation of dissimilarities which requires the specification of a scale origin.

The recently proposed *normalized cut* approach [18] provides interesting normalized cluster criteria which are not additive. The normalized cut cost function has been introduced only for two clusters. But as it is equal to the minimization of the normalized association

$$\mathcal{H}^{\mathrm{nc}}(\mathbf{M}; \mathbf{D}) = \sum_{\nu=1}^{K} \frac{\sum_{i=1}^{N}\sum_{j=1,\, j\neq i}^{N} \mathbf{M}_{i\nu}\mathbf{M}_{j\nu} D_{ij}}{(\sum_{i=1}^{N}\sum_{j=1,\, j\neq i}^{N} \mathbf{M}_{i\nu} + \mathbf{M}_{j\nu} - \mathbf{M}_{i\nu}\mathbf{M}_{j\nu}) D_{ij}} \tag{12}$$

it is naturally extended to multiple clusters. It should be noted that by using similarities and maximizing (12) a qualitatively different criterion is obtained, which is well-defined even for highly sparse dissimilarity data. Both criteria are scale invariant and robust in the strong sense, but they are not shift invariant. The normalized cut is well-defined only for positive proximities and is thus only defined on a ratio scale.

### 2.4. Sparse proximity data

As discussed in the introduction, it is important for large-scale applications of proximity-based clustering to develop methods which apply to arbitrary sparse proximity matrices. In order to distinguish between known and unknown dissimilarities an irreflexive graph $(V, E)$ with $V = \{1, \ldots, N\}$, $E \subset V \times V$ is introduced. For notational convenience denote by $\mathcal{N}_i \subset V$ the set of graph neighbors of node or object $i$, i.e., $\mathcal{N}_i = \{j \in V\colon (i, j) \in E\}$. The above axiom system can be extended to cover this case [19], but we restrain from a formal treatment which involves a lot of technical details. The essential result is that there are two ways of averaging which result in invariant criteria for sparse proximity matrices. For intra-cluster compactness objective functions the average intra-cluster dissimilarity is calculated either by one-step averages $D_\nu^{(I)}$ or by the cascaded averaging $D_\nu^{(II)}$, where

$$D_\nu^{(I)} = \frac{\sum_{(i,j)\in E} M_{i\nu} M_{j\nu} D_{ij}}{\sum_{(i,j)\in E} M_{i\nu} M_{j\nu}},$$

$$D_\nu^{(II)} = \frac{\sum_{i=1}^{N} M_{i\nu} \dfrac{\sum_{j\in\mathcal{N}_i} M_{j\nu} D_{ij}}{\sum_{j\in\mathcal{N}_i} M_{j\nu}}}{\sum_{i=1}^{N} M_{i\nu}}. \tag{13}$$

[4] The almost equal relation '$\cong$' refers to the additional diagonal contributions $D_{ii}$ which is negligible for large $N$. Alternatively, the definition of additivity could be extended to cover the reflexive case to get a true identity.

In a similar way, average inter-cluster dissimilarities $D_{\nu\mu}$ are generalized. Since the different possibilities of weighting clusters and averaging are independent, they can be freely combined. The two sparse data variants obtained from $\mathscr{H}^{\mathrm{pc1}}$ are given by

$$\mathscr{H}_I^{\mathrm{pc1}} = \sum_{\nu=1}^{K}\left(\sum_{i=1}^{N} M_{i\nu}\right)\frac{\sum_{(i,j)\in E} M_{i\nu}M_{j\nu}D_{ij}}{\sum_{(i,j)\in E} M_{i\nu}M_{j\nu}},$$
$$\mathscr{H}_{II}^{\mathrm{pc1}} = \sum_{\nu=1}^{K}\sum_{i=1}^{N} M_{i\nu}\frac{\sum_{j\in\mathscr{N}_i} M_{j\nu}D_{ij}}{\sum_{j\in\mathscr{N}_i} M_{j\nu}}. \quad (14)$$

### 2.5. Hierarchical clustering

The second extension of data clustering objective functions concerns the problem of hierarchical clustering [20]. This is in particular important, as it has often been claimed, that partitional methods are inherently non-hierarchical [1]. By a hierarchical clustering solution data partitionings $\mathbf{M}^K$ are specified on all levels of $K = 2,\dots, K_{\max}$. In order for the different data partitionings to be consistent, it is required that between consecutive solutions always one cluster is split into two sub-clusters. For simplicity a consistent numbering of clusters is enforced.

**Definition 5.** A sequence of data partitionings $\mathbf{M}^K \in \mathscr{M}_{N,K}$ with $K = 2,\dots, K_{\max}$ is *hierarchical*, if cluster indices $\alpha(K)$ exist such that

$$M_{i\nu}^K = \begin{cases} M_{i\nu}^{K+1} & \text{if } \nu \neq \alpha, \\ M_{i\nu}^{K+1} + M_{i(K+1)}^{K+1} & \text{if } \nu = \alpha(K). \end{cases}$$

Definition 5 implies that a hierarchical clustering solution is completely described by the finest data partitioning $\mathbf{M}^{K_{\max}}$ and the sequence of splits $\alpha(K)$, which implicitly encode the topology of a complete binary tree. Following the same underlying principle as in Definition 2, hierarchical clustering criteria are defined by additive composition of the single level solutions.

**Definition 6** (Hierarchical clustering). A clustering criterion $\mathscr{H}_N$ is *hierarchical* if it has the following functional form:

$$\mathscr{H}_N(\mathbf{M}^{K_{\max}}; \alpha(K), \mathbf{D}) = \sum_{K=2}^{K_{\max}} w_K \mathscr{H}_{N,K}(\mathbf{M}^K; \mathbf{D}),$$

where $\mathscr{H}_{N,K}$ are clustering criteria and $w_K \in \mathbb{R}_0^+$ with $\sum_{K=2}^{K_{\max}} w_K = 1$. $\mathscr{H}_N$ is *additive*, if all $\mathscr{H}_{N,K}$ are additive. $\mathscr{H}_N$ is *invariant* if all $\mathscr{H}_{N,K}$ are invariant.

The least biased choice is a constant weighting $w_K = 1/(K_{\max} - 1)$ of all data partitionings. But if prior knowledge about the data 'granularity' is available it can be incorporated by an appropriate non-constant weighting. In the limiting case of $w_{K+1}/w_K \to 0$ an objective function is obtained which corresponds to greedy splitting, while the inverse limit $w_K/w_{K+1} \to 0$ corresponds to greedy cluster merging. In this limit Ward's method is guaranteed to find the minimum, if $\mathscr{H}_{N,K} = \mathscr{H}_{N,K}^{\mathrm{pc1}}$ is utilized, and hence the definition naturally includes this agglomerative technique as a special case.

The presented optimization approach has the advantage to guarantee a consistent and strictly nested hierarchy of clusters. Of course, not all $K$-partitionings have to be considered as defining 'natural' solutions. Therefore the following validation criterion is proposed in order to eliminate intermediate levels of the hierarchy. To all $K$-partition costs $\mathscr{H}_{N,K}(\mathbf{M}(k))$ *complexity costs* $\mathscr{H}^{\mathrm{cmx}} = \lambda N \log(K)$, $\lambda \in \mathbb{R}^+$ are added. Only those $K$-partitions are kept which possess a range of $\lambda$ where they have the lowest total costs. The motivation of complexity costs proportional to $\log K$ stems from the expected cost decay for a random instance in the $N \to \infty$ limit. Notice, that this index does not necessarily identify a single 'true' partition, but only eliminates implausible partitions, which are sub-optimal for *all* choices of $\lambda$. In particular, it is not a model selection criterion in the Bayesian sense.

## 3. Optimization by annealing

### 3.1. Simulated annealing and Gibbs sampling

*Simulated annealing* [21] is a popular stochastic optimization heuristic which has successfully been applied to large-scale problems, e.g., in computer vision and in operations research. Simulated annealing performs a random search which can be modeled by an inhomogeneous discrete-time Markov chain $(\mathbf{M}^{(t)})_{t\in\mathbb{N}}$ converging towards its equilibrium distribution, the *Gibbs distribution*

$$\mathbf{P}_{\mathscr{H}}(\mathbf{M}) = \frac{1}{\mathscr{Z}_T}\exp(-\mathscr{H}(\mathbf{M})/T),$$
$$\mathscr{Z}_T = \sum_{\mathbf{M}\in\mathscr{M}} \exp(-\mathscr{H}(\mathbf{M})/T). \quad (15)$$

Formally, denote by $\mathscr{P} = \{\mathbf{P} : \mathscr{M} \to [0,1]: \sum_{\mathbf{M}\in\mathscr{M}} \mathbf{P}(\mathbf{M}) = 1\}$ the space of probability distributions on $\mathscr{M}$, by $\mathscr{S}(\mathbf{P})$ the entropy of $\mathbf{P}$ and by

$$\mathscr{F}_T(\mathbf{P}) = \langle \mathscr{H} \rangle_{\mathbf{P}} - T\mathscr{S}(\mathbf{P}) = \sum_{\mathbf{M}\in\mathscr{M}} \mathbf{P}(\mathbf{M})\mathscr{H}(\mathbf{M}) + T\sum_{\mathbf{M}\in\mathscr{M}} \mathbf{P}(\mathbf{M})\log\mathbf{P}(\mathbf{M}) \quad (16)$$

the *generalized free energy*, which plays the role of an objective function over $\mathscr{P}$. For arbitrary assignment problems $\mathscr{H}$ the Gibbs distribution $\mathbf{P}_{\mathscr{H}}$ minimizes the generalized free energy, i.e., $\mathbf{P}_{\mathscr{H}} = \arg\min_{\mathbf{P}\in\mathscr{P}} \mathscr{F}_T(\mathbf{P})$. For the data clustering problem we focus on *local* algorithms with state transitions restricted to pairs of configurations which differ in the assignment of at most one object or *site*. Denote by $s_i(\mathbf{M}, \boldsymbol{e}_\alpha)$ the matrix obtained by substituting the $i$th row of $\mathbf{M}$ by the unity vector $\boldsymbol{e}_\alpha$. For convenience introduce a *site visitation schedule* as a map $v: \mathbb{N} \to \{1, \dots, N\}$ such that, in the limit, every site is visited infinitely often. A sampling scheme known as the *Gibbs sampler* [10] is advantageous, if it is possible to efficiently sample from the conditional distribution of $\mathbf{P}_{\mathscr{H}}$ at site $v(t)$, given the assignments at all other sites $\{j \neq v(t)\}$. For a given schedule $v$ the Gibbs sampler is defined by the non-zero transition probabilities

$$S_t(s_{v(t)}(\mathbf{M}, \boldsymbol{e}_\alpha), \mathbf{M}) = \frac{\exp[-\mathscr{H}(s_{v(t)}(\mathbf{M}, \boldsymbol{e}_\alpha))/T(t)]}{\sum_{v=1}^{K} \exp[-\mathscr{H}(s_{v(t)}(\mathbf{M}, \boldsymbol{e}_v))/T(t)]}. \quad (17)$$

The basic idea of annealing is to gradually lower the temperature $T(t)$, on which the transition probabilities depend. For the zero temperature limit a local optimization algorithm known as *Iterative Conditional Mode* [22] (ICM) is obtained.

### 3.2. Deterministic and mean-field annealing

An approach known as *deterministic annealing* (DA) combines the advantages of a temperature controlled continuation method with a fast, purely deterministic computational scheme. The key idea of DA is to calculate the relevant expectation values of system parameters, e.g., the variables of the optimization problem with analytical techniques. In DA a combinatorial optimization problem with objective function $\mathscr{H}$ over $\mathscr{M}$ is relaxed to a family of stochastic optimization problems with objective functions $\mathscr{F}_T$ over a subspace $\mathscr{Q} \subseteq \mathscr{P}$. The subspace $\mathscr{Q}$ discussed in the context of assignment problems is the space of all factorial distributions given by

$$\mathscr{Q} = \left\{ \mathbf{Q} \in \mathscr{P} : \mathbf{Q}(\mathbf{M}) = \prod_{i=1}^{N} \sum_{v=1}^{K} M_{iv} q_{iv},\ \forall \mathbf{M} \in \mathscr{M} \right\}. \quad (18)$$

With this specific choice of $\mathscr{Q}$, DA is more specifically called *mean-field annealing* (MFA) whenever $\mathbf{P}_{\mathscr{H}} \notin \mathscr{Q}$. $\mathscr{Q}$ is distinguished from other subspaces of $\mathscr{P}$ in many respects: (i) the dimensionality of $\mathscr{Q}$ increases only linearly with $N$, (ii) an efficient alternation algorithm exists for a very general class of objective functions, (iii) in the limit $T \to 0$ locally optimal solutions to the combinatorial optimization problem can be recovered. $\mathscr{F}_T$ is an entropy-smoothed version of the original optimization problem which becomes convex over $\mathscr{Q}$ for sufficiently large $T$. DA tracks solutions from high to low temperatures, where gradually more and more details of the original objective function appear. The most important properties of factorial distributions are

1. All correlations w.r.t. $\mathbf{Q}$ vanish for assignment variables at different sites.
2. The parameters $q_{iv}$ can be identified with the $\mathbf{Q}$-averages $\langle M_{iv} \rangle$.

Calculating stationary conditions by differentiation of the free energy (16) with respect to the parameters $q_{iv}$ a system of coupled transcendental, so-called *mean-field equations*, is obtained, which is solved by a convergent asynchronous update scheme.

**Theorem 1.** *Let $\mathscr{H}$ be an arbitrary clustering criterion and $v$ a site visitation schedule. Then, for arbitrary initial conditions, the following asynchronous update scheme converges to a local minimum of the generalized free energy $\mathscr{F}_T$ (16) over $\mathscr{Q}$:*

$$q_{iv}^{\text{new}} = \frac{\exp[-\frac{1}{T} h_{iv}]}{\sum_{\mu=1}^{K} \exp[-\frac{1}{T} h_{i\mu}]}$$

$$\textit{where } h_{iv} = \left. \frac{\partial \langle \mathscr{H} \rangle}{\partial q_{iv}} \right| q_{iv}^{\text{old}} = \langle \mathscr{H} \rangle_{s_i(\mathbf{Q}^{\text{old}}, \boldsymbol{e}_v)} \quad \textit{and } i = v(t). \quad (19)$$

Notice, that the variables $h_{iv}$, called *mean-fields* by analogy to the naming convention in statistical physics, are only auxiliary parameters to compactify the notation. The update scheme is essentially a non-linear Gauß–Seidel relaxation to iteratively solve the coupled transcendental equations. Combining the convergent update scheme with an annealing schedule yields a GNC-like [23] algorithm, because $\mathscr{F}_T$ is convex over $\mathscr{Q}$ for $T$ sufficiently large. For a derivation and more mathematical details please see Refs. [9,14].

There is a tight relationship between the quantities $g_{iv} = \mathscr{H}(s_i(\mathbf{M}, \boldsymbol{e}_v))$ involved in implementing the Gibbs sampler in Eq. (17) and mean-field equations. Rewriting Eq. (19) we arrive at

$$h_{iv} = \sum_{\mathbf{M} \in \mathscr{M}} \frac{M_{iv}}{q_{iv}} \mathscr{H}(\mathbf{M})\ \mathbf{Q}(\mathbf{M})$$

$$= \sum_{\mathbf{M} \in \mathscr{M}} \mathscr{H}(s_i(\mathbf{M}, \boldsymbol{e}_v)) \mathbf{Q}(\mathbf{M}) = \langle g_{iv} \rangle_{\mathbf{Q}}. \quad (20)$$

Thus the mean-field $h_{iv}$ is a $\mathbf{Q}$-averaged version of the local costs $g_{iv}$.

### 3.3. Gibbs sampling for clustering of proximity data

To efficiently implement the Gibbs sampler one has to optimize the evaluation of $\mathscr{H}$ for a sequence of locally modified assignment matrices. It is an important observation that the quantities $g_{iv}$ only have to be computed up to an additive shift, which may depend on the site index $i$, but not on $v$. The choice of $g_{iv}(\mathbf{M}) = \mathscr{H}(s_i(\mathbf{M}, \boldsymbol{e}_v)) - \mathscr{H}(s_i(\mathbf{M}, 0))$ leads to compact analytical expressions, because the contributions of the reduced system without site $i$ are subtracted. For the functional representation of additive clustering criteria in Proposition 1 the general formula for the Gibbs fields is given by

$$\begin{aligned} g_{i\nu} &= 2\sum_{j\neq i} M_{j\nu}\psi^{(1)}(D_{ij}, n_\nu^{+i}) \\ &\quad - 2\sum_{j\neq i}\sum_{\mu\neq\nu} M_{j\mu}\psi^{(2)}(D_{ij}, n_\nu^{+i}, n_\mu) \\ &\quad + \sum_{j\neq i,\ k\neq i,j} M_{j\nu}M_{k\nu}[\psi^{(1)}(D_{jk}, n_\nu^{+i}) - \psi^{(1)}(D_{jk}, n_\nu^{-i})] \\ &\quad - \sum_{j\neq i,\, k\neq i,\, j} M_{j\nu}\sum_{\mu\neq\nu} M_{k\mu}[\psi^{(2)}(D_{jk}, n_\nu^{+i}, n_\mu) \\ &\quad - \psi^{(2)}(D_{jk}, n_\nu^{-i}, n_\mu)]. \end{aligned} \tag{21}$$

Here, $n_\nu^{+i} = n_\nu + 1 - M_{i\nu}$ is the cluster size after adding object $\mathbf{o}_i$ to cluster $\mathscr{C}_\nu$ and $n_\nu^{-i} = n_\nu - M_{i\nu}$ the cluster size after eliminating the object from that cluster. In order to derive the Gibbs fields for specific cost functions the occuring differences between the contribution functions are calculated. Exemplary, we display the Gibbs field equation for $\mathscr{H}^{\mathrm{pc}1}$ explicitly,

$$\begin{aligned} g_{i\nu}^{pc1} &= \sum_{j\neq i}\left[\frac{1}{n_\nu^{+i}} + \frac{1}{n_\nu^{-i}}\right] M_{j\nu}D_{ij} \\ &\quad - \frac{1}{n_\nu^{+i}n_\nu^{-i}}\sum_{j\neq i}\sum_{k\neq i,j} M_{j\nu}M_{k\nu}D_{jk}. \end{aligned} \tag{22}$$

In order to obtain an efficient implementation of the Gibbs sampler, we propose to utilize book keeping quantities, e.g., for intra-cluster compactness criteria the cluster sizes $n_\nu$, $a_{i\nu} = \sum_{j\neq i} M_{j\nu}D_{ij}$, and $A_\nu = \sum_{i=1}^{N} M_{i\nu}a_{i\nu}$. The computation of all Gibbs fields based on the book-keeping quantities then requires $\mathcal{O}(NK)$ arithmetical operations. After locally changing an assignment of a single object, the update of all book-keeping quantities requires $\mathcal{O}(N)$ operations, for a complete sweep thus $\mathcal{O}(N^2)$.

The generalization to the case of sparse proximity matrices is straightforward. However, the averages $D_\nu^{(II)}$ yield more complex equations, since it involves object-specific normalization functions, while $D_\nu^{(I)}$ has a cluster-specific normalization. In the case of hierarchical clustering, the additive combination of contributions from different data partitionings allows us to reduce the calculation of Gibbs fields to the case of 'flat' clustering, $g_{i\nu} = \sum_{K=2}^{K_{\max}} g_{i\alpha(\nu,K)}^{K}$, where $g_{i\alpha}^{K}$ denotes the Gibbs field corresponding to cluster $\mathscr{C}_\alpha$ in the $K$-cluster partitioning and $\alpha(\nu, K)$ denotes the index of the (super-)cluster the leaf cluster $\mathscr{C}_\nu$ is associated with in the coarse level partitioning with $K$ clusters.

### 3.4. Mean-field annealing for clustering of proximity data

According to Theorem 1 and exploiting Eq. (20) the problem of calculating the mean-fields $h_{i\nu}$ is reduced to the problem of Q-averaging the quantities $g_{i\nu}$. The main technical difficulty in calculating the mean-field equations are the averages of the normalization constants. Although every Boolean function has a polynomial normal form, which would in principle eliminate the denominator, some approximations have to be introduced to avoid exponential order in the number of conjunctions. This is done by independently averaging the numerator and the normalization in the denominator. Using the correlation properties of factorial distributions this leads to $h_{i\nu}(\mathbf{M}) = g_{i\nu}(\langle\mathbf{M}\rangle)$. Thus approximate mean-field equations are obtained by simply replacing the Boolean variables in the Gibbs field equations by its probabilistic counterparts.

MFA offers the possibility to track the phase transitions (cluster splits) in order to obtain a meaningful tree topology. This strategy has been first pursued by Rose et al. for vector quantization [12,24] and can be implemented by tentatively splitting clusters at each temperature level and fusing degenerated clusters after convergence of the mean-field equations until the maximal number of clusters $K_{\max}$ is reached.

## 4. Applications and experimental results

Performance comparisons for different optimization algorithms are straightforward, because the obtained solution quality is assessed by the utilized objective function. In contrast, the empirical verification is much more difficult for the modeling problem. A decisive evaluation of the quality of a clustering solution is only possible if information about ground truth is available. This is the case in both application discussed in the sequel: texture segmentation and document retrieval. These problems are interesting on their own, but the obtained results promise that the developed methods are of relevance in many other application domains as well.

### 4.1. Unsupervised segmentation of textured images

#### 4.1.1. Proximity-based texture segmentation

The *unsupervised segmentation* of textured images is widely recognized as a challenging computer vision problem. The main conceptual difficulty is the definition of an appropriate homogeneity measure in mathematical terms. Many explicit texture models have been considered in the last three decades. Textures are often represented by feature vectors, e.g., by the means and variances of a filter bank output [25], wavelet coefficients [26] or as parameters of an explicit Markov random field model [27]. Feature-based approaches, however, often suffer from the inadequacy of the metric utilized in parameter space to appropriately represent visual dissimilarities between different textures, a problem which is severe for unsupervised segmentation. It is an important observation due to Geman et al. [17], that the segmentation problem can be defined in terms of pairwise dissimilarities between textures without the need for explicit texture features.

Our approach to unsupervised texture segmentation is based on a *Gabor wavelet* scale-space image representation with frequency-tuned filters [13,14,25,28]. The similarity between pairs of texture patches is then measured by a statistical test applied to the empirical feature distribution functions of locally sampled Gabor coefficients. We have intensively investigated several tests with respect to their texture discrimination ability [29]. As a result throughout this work a $\chi^2$-statistic will be used. Denote by $\boldsymbol{I}(\boldsymbol{x})$ the vector of Gabor coefficients at a position $\boldsymbol{x}$. $\boldsymbol{I}(\boldsymbol{x})$ contains information about the spatial relationship between pixels in the neighborhood of $\boldsymbol{x}$, but may not capture the complete characteristics of the texture. Therefore, the (*weighted*) *empirical distribution* of Gabor coefficients in a window around $\boldsymbol{x}_i$ is considered,

$$f_i^{(r)}(t) = \sum_{y:\ t_{k-1} \leqslant I_r(y) < t_k} W_r(\|\boldsymbol{x}_i - \boldsymbol{y}\|) \bigg/ \sum_{y} W_r(\|\boldsymbol{x}_i - \boldsymbol{y}\|),$$

$$\text{for } t \in [t_{k-1}; t_k]. \tag{23}$$

$W_r$ denotes a non-negative, monotonely decreasing window function centered at the origin. $t_0 = 0 < t_1 < \cdots < t_L$ is a suitable binning. Here we consider only the computationally simplest choice of squared windows with a constant weight inside and zero weight outside. The window size is chosen proportional to the standard deviation $\sigma_r$ of the Gabor filter [25]. The dissimilarity between textures at two positions $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ is evaluated independently for each Gabor filter according to

$$D_{ij}^{(r)} = \chi^2 = \sum_{k=1}^{L} \frac{(f_i^{(r)}(t_k) - \hat{f}(t_k))^2}{\hat{f}(t_k)}$$

$$\text{where } \hat{f}(t_k) = [f_i^{(r)}(t_k) + f_j^{(r)}(t_k)]/2. \tag{24}$$

The dissimilarities $D_{ij}^{(r)}$ are finally combined by the $L_1$-norm, $D_{ij} = \sum_r D_{ij}^{(r)}$. The $L_1$-norm is less sensitive to differences in single channels than the $L_\infty$-norm proposed in Ref. [17] and empirically showed the best performance within the $L_p$-norm family [29].

### 4.1.2. Sparse pairwise clustering

We selected a representative set of 86 micro-patterns from the Brodatz texture album [30] to empirically test the segmentation algorithms on a wide range of textures.[5] A database of random mixtures (512 × 512 pixels each) containing 100 entities of $K = 5$ textures each (as depicted in Fig. 1) was constructed from this collection. All segmentations are based on a filter bank of twelve Gabor filters with four orientations and three scales. For each image a subset of 64 × 64 sites was considered. For each site we used a square window of size 8 × 8 for the smallest scale. A sparse neighborhood including the 4 nearest neighbors and (on average) 80 randomly selected neighbors was chosen.

As an independent reference algorithm using a frequency-based feature extraction we selected the approach of Jain and Farrokhnia [25], which we refer to as Gabor Feature Clustering (GFC). The vector of Gabor coefficients at a position $x_i$ is non-linearly transformed by using the absolute value of the hyperbolic tangents of the real part. Then a Gaussian smoothing filter is applied and the resulting feature vectors are rescaled to zero mean and unit variance. The texture segmentation problem is formulated as a clustering problem using the $K$-means clustering criterion with an Euclidean norm (10). We have chosen a deterministic annealing algorithm for clustering of vectorial data due to Rose et al. [31], which was empirically found to yield slightly better results than the clustering technique proposed in Ref. [25]. In order to obtain comparable results we used the same 12 Gabor filters and extracted feature vectors on the 64 × 64 regular sub-lattice of sites.

As an example for an agglomerative clustering method we selected *Ward's method* [1], which experimentally achieved substantially better results than single and complete linkage. For all methods small and narrow regions were removed in a simple postprocessing step to avoid the typical speckle-like noise inherent to all clustering methods under consideration [17].

Table 1 summarizes the obtained mean and median values for all cost functions under consideration, evaluated on the database of mixture images with $K = 5$ textures each. In addition, we report the percentage of outliers with more than 20% segmentation error, which we define as structural segmentation errors, since typically complete textures are missed. For $\mathscr{H}_I^{\mathrm{pc}1}$ ($\mathscr{H}_{II}^{\mathrm{pc}1}$) a median segmentation error rate as low as 3.7% (3.6%) was obtained. Both cost functions yield very similar results as expected and exhibit only few outliers. We recommend the use of $\mathscr{H}_I^{\mathrm{pc}1}$, because it can be implemented more efficiently. For $\mathscr{H}_I^{\mathrm{pc}2}$ both mean and median error are larger.[6] We conclude, that in most cases the invariant

[5] We a priori excluded the textures d25-d26, d30-d31, d39-d48, d58-d59, d61-d62, d88-d89, d91, d96-d97, d99, d107-d108 by visual inspection due to missing micro-pattern properties, i.e., all textures are excluded, where the texture property is lost when considering small image parts.

[6] The missing robustness properties render $\mathscr{H}_{II}^{\mathrm{pc}1}$ inapplicable. As a compensation one may add *prior costs* penalizing the generation of extremely small clusters, e.g., by adding $\lambda_s(\mathbf{D})\sum_\nu n_\nu^{-1}$. Yet, such prior costs violate the invariance axioms, as $\lambda_s(\mathbf{D}) \sim \sum_{i \neq j} D_{ij}/(N-1)$ in order to fulfill scale invariance, but on the other hand $\lambda_s(\mathbf{D}) = \lambda_s(\mathbf{D} + \Delta d)$ by the shift invariance requirement. The robustness deficiency is partially compensated by choosing an appropriate prior, but at the cost of empirically fixing an additional, data-dependent algorithmic parameter, which has to be considered as a major deficiency in the context of unsupervised texture segmentation.

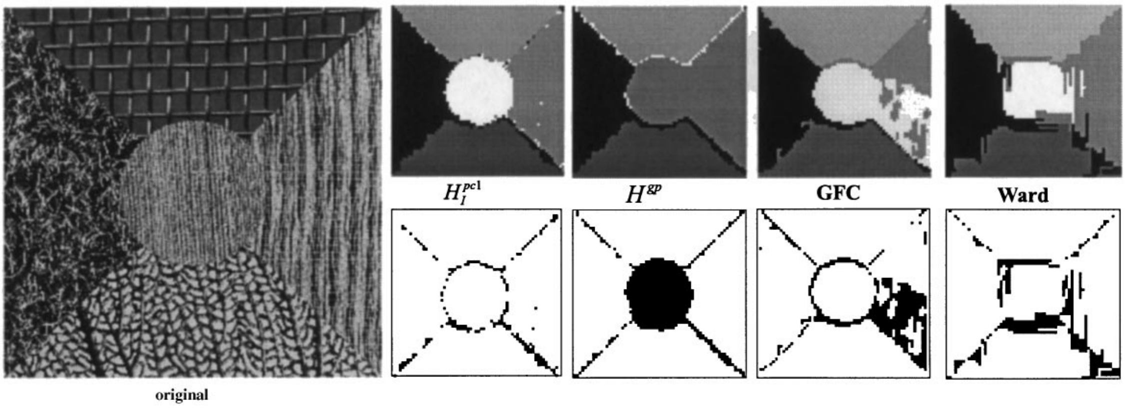

Fig. 1. Typical segmentation results using 5 clusters for different cost functions before postprocessing. Misclassified sites are depicted in black.

Table 1
Mean and median error compared to ground truth for segmenting 100 randomly generated images with $K = 5$ textures each using MFA

| | $\mathscr{H}_I^{\mathrm{pc1}}$ (%) | $\mathscr{H}_{II}^{\mathrm{pc1}}$ (%) | $\mathscr{H}_I^{\mathrm{pc2}}$ (%) | $\mathscr{H}^{\mathrm{gp}}$ (%) | $\mathscr{H}^{\mathrm{nc}}$(%) | Ward (%) | GFC (%) |
|---|---|---|---|---|---|---|---|
| Median | 3.7 | 3.6 | 5.0 | 4.0 | 4.0 | 7.7 | 6.7 |
| Mean | 5.8 | 6.0 | 7.7 | 7.7 | 6.6 | 11.5 | 10.6 |
| 20%-quantile | 6 | 5 | 9 | 11 | 9 | 18 | 18 |

*Note*: The columns correspond to different cost functions $\mathscr{H}$. For $\mathscr{H}_I^{\mathrm{pc2}}$ a prior with $\lambda_s = (150/N)\mathbf{E}[D_{ij}]$ was used, while the data were shifted by $\triangle d = 0.1 - \mathbf{E}[D_{ij}]$ for $\mathscr{H}^{\mathrm{gp}}$.

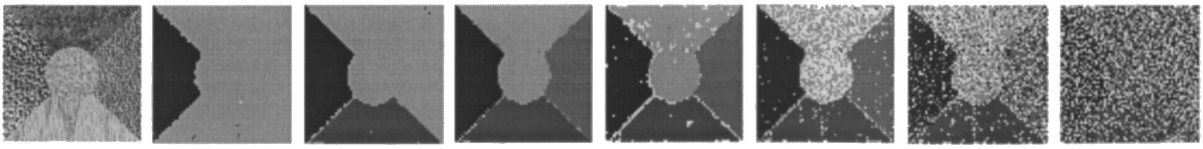

Fig. 2. Segmentations obtained by $\mathscr{H}^{\mathrm{gp}}$ for several data shifts: original image and segmentations with a mean dissimilarity of $-0.05$, 0, 0.05, 0.1, 0.15, 0.2 and 0.25 are depicted. Segments collapse for negative shifts. For large positive shifts the sampling noise induced by the random neighborhood system dominates the data contributions.

cost functions based on a pairwise data clustering formalization capture the true structure of the image. Furthermore, the robustness property of $\mathscr{H}^{\mathrm{pc1}}$ has proven to be advantageous. The feature-based GFC as well as Ward's method are clearly outperformed.

The unnormalized cost function $\mathscr{H}^{\mathrm{gp}}$ severely suffers from the missing shift-invariance property as shown in Fig. 2. Depending on the shift the unnormalized cost function often completely misses several texture classes. There may not even exist a parameter value to find all five textures. Even worse, the optimal value depends on the data at hand and varies for different images. With $\mathscr{H}^{\mathrm{gp}}$ a median error rate of 4.0% with substantially more outliers was achieved. The data were shifted to a mean dissimilarity of $\mathbf{E}[D_{ij}] = 0.1$, a value which was obtained after extensive experimentation. For the normalized cut $\mathscr{H}^{\mathrm{nc}}$ a median error rate of 4.0% and 9% outliers were achieved, which is better than the unnormalized graph partitioning cost function, but worse than the invariant normalized criteria $\mathscr{H}^{\mathrm{pc1}}$. The dissimilarity data has been scaled to a maximal value 1 and has than been transformed by $D_{ij}^{\mathrm{new}} = \exp(-D_{ij}/c)$ as suggested by Shi et al. [18] with a parameter $c = 0.25$ determined by extensive benchmarking.

We thus conclude that shift and scale invariance are important properties to avoid parameter fine tuning of

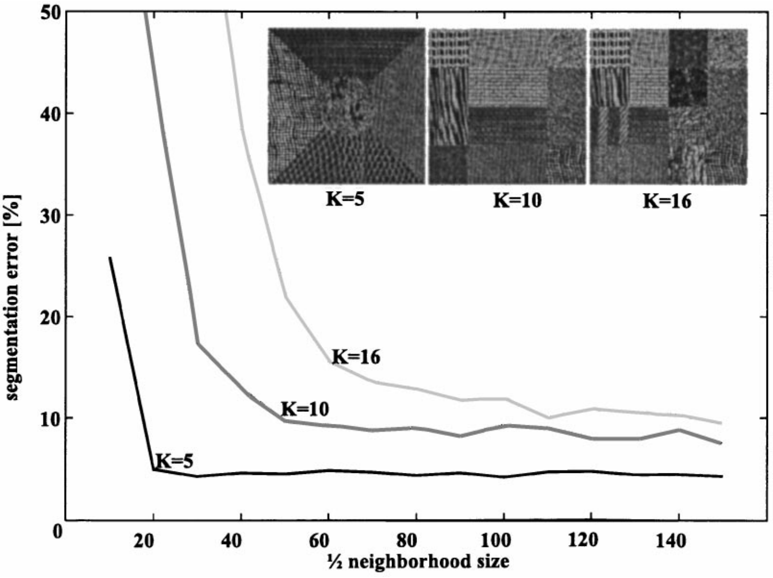

Fig. 3. Segmentation error for different neighborhood sizes for $K = 5$, $K = 10$ and $K = 16$ before postprocessing.

sensitive parameters and that the increased computational complexity for additional normalizations in $\mathscr{H}^{\mathrm{pc1}}$ is well-spent.

In Fig. 3 the effect of data sparseness is examined. The asymptotic segmentation error is already obtained for highly sparse data matrices. The neighborhood size needed grows moderately with the number $K$ of segments. Clustering of sparse data is therefore a successful concept for large-scale applications.

### 4.1.3. Gibbs sampling and MFA

Another important question concerns the quality of the MFA algorithm as opposed to stochastic procedures. The quality of the proposed clustering algorithm was evaluated by comparing the costs and the errors of the achieved segmentation with the local ICM algorithm and with the stochastic Gibbs sampling method. The error results are summarized in Table 2. For the graphical representation the distribution of the differences of costs were chosen. Exemplary the cost and error differences for $\mathscr{H}_I^{\mathrm{pc1}}$ using MFA versus ICM and MFA versus Gibbs sampling are depicted in Fig. 4. Compared with the ICM algorithm a substantial improvement both in terms of energy and segmentation quality has to be noted. As expected the ICM algorithm gets frequently stuck in inferior local minima. On the other hand the ICM algorithm runs notably faster then the other ones.

The comparison with the Gibbs sampler is more difficult, as the performance of the Gibbs sampler improves with slower cooling rates. We decided to use an approximately similar running time for both MFA and Gibbs sampler in our current implementation.[7] MFA and Gibbs sampling yield similar results. In all cases the differences are small. A rather conservative annealing schedule has been used. Empirically, little improvement has been observed for the Gibbs sampler with slower annealing, although it is well-established that for logarithmic annealing schedules the Gibbs sampling scheme converges to the global minimum in probability [33]. Because of this global optimization properties of Gibbs sampling, we conclude that MFA yields near-optimal solutions in most runs. Since the loss in segmentation quality for MFA by a faster annealing schedule is substantially lower than for Gibbs sampling, the MFA

Table 2
Mean and median error compared to ground truth for segmenting 100 randomly generated images with $K = 5$ textures each for ICM, MFA and the Gibbs–sampler before postprocessing

| | ICM (%) | MFA (%) | Gibbs sampling (%) |
|---|---|---|---|
| Median | 6.5 | 5.4 | 5.4 |
| Mean | 11.6 | 7.6 | 7.6 |
| 20%-quantile | 23 | 7 | 7 |

*Note*: The results have been obtained for the cost function $\mathscr{H}_I^{\mathrm{pc1}}$.

[7] About 300 s on a SUN Ultra-Sparc. For MFA this can be improved to 3–5 s using multiscale annealing techniques [32].

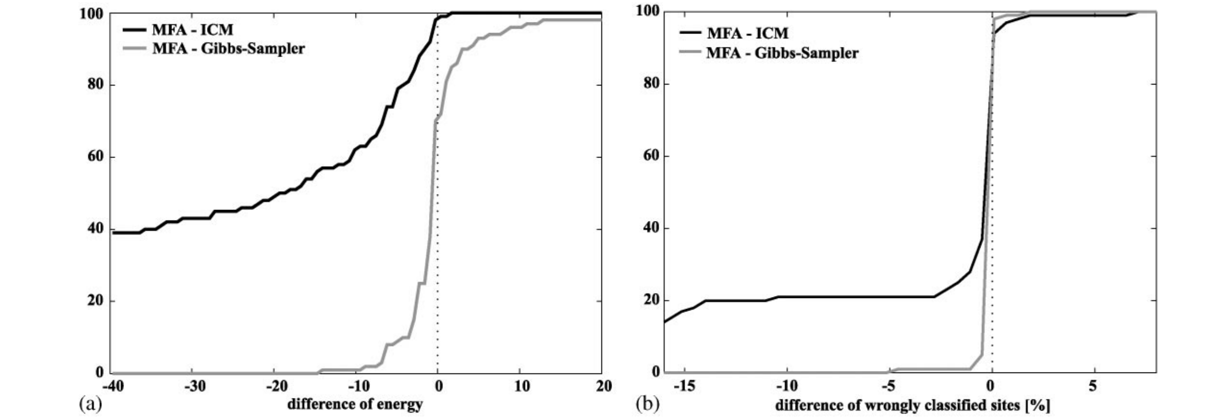

Fig. 4. The empirical density of (a) the cost difference and (b) the segmentation errors of MFA versus ICM and versus the Gibbs sampler evaluated over 100 images.

algorithm is a good choice within a certain window of the speed-quality trade-off.

### 4.1.4. Hierarchical clustering

The result of a hierarchical segmentation on a test image containing $K = 16$ Brodatz textures is depicted in Fig. 5. All textures have been correctly identified and the borders are localized precisely. Stable solutions according to our criterion have been detected for $K = 11$ and $K = 16$. The hierarchical structure detected is in accordance with the psychophysical expectation.

A segmentation example for an aerial image of San Francisco with the same set of parameters is shown in Fig. 6. Applying the proposed validation criterion the segmentations with $K = 3$, 4 and 9 are selected. $K = 6$ possesses significant local stability. The hierarchical organization is very intuitive: the first split separates land and ocean. At later stages homogeneously tilled areas are distinguished from vegetation. The results for Ward's method and for the complete linkage algorithm are less satisfying. In the segmentation obtained by Ward's method land and ocean are mixed, while for complete linkage several spurious segments occur. We conclude that by the optimization approach to hierarchical clustering semantically consistent segmentation hierarchies are obtained. These methods therefore offer an attractive alternative to the widely used family of agglomerative clustering algorithms.

## 4.2. Clustering for information retrieval

### 4.2.1. Proximity-based clustering of document databases

Information retrieval in large databases is one of the key topics in *data mining*. The problem is most severe in cases where the query cannot be formulated precisely, e.g., in natural language interfaces for documents or in image databases. Typically, one would like to obtain those entries which best match a given query according to some similarity measure. Yet, it is often difficult to reliably estimate similarities, because the query may not contain enough information, e.g., not all possibly relevant keywords might occur in a query for documents. Therefore, one often applies the *cluster hypothesis* [34]: if an entry is relevant to a query, similar entries may also be relevant to the query although they may not possess a high similarity to the query itself. Clustering thus provides a way of pre-structuring a database for the purpose of improved information retrieval [35].

Following state of the art techniques, we utilized a word stemmer and a stop word list to automatically generate index terms. A document is represented by a (sparse) binary vector $B$, where each entry corresponds to the occurrence of a certain index term. As a *measure of association* between two documents we utilized the cosine measure which normalizes the intersection with the geometrical mean of the number of index terms,

$$D_{ij} = 1 - \frac{B_i^t B_j}{\sqrt{|B_i||B_j|}}. \tag{25}$$

Other commonly applied measures are the Jaccard coefficient and the Dice coefficient [36].

We have tested the different clustering criteria and algorithms on the Medline (*MED*) database consisting of $N = 1033$ abstracts of medical articles. The *MED* collection has 30 queries with known relevance assessment. Based on this ground truth data, we evaluate the two most important quantities for retrieval performance: the *precision* $P$ (% of returned documents which was relevant) and the *recall fraction* $R$ (% of relevant documents which was actually returned). There is obviously a trade-off between these quantities and we plot retrievals as

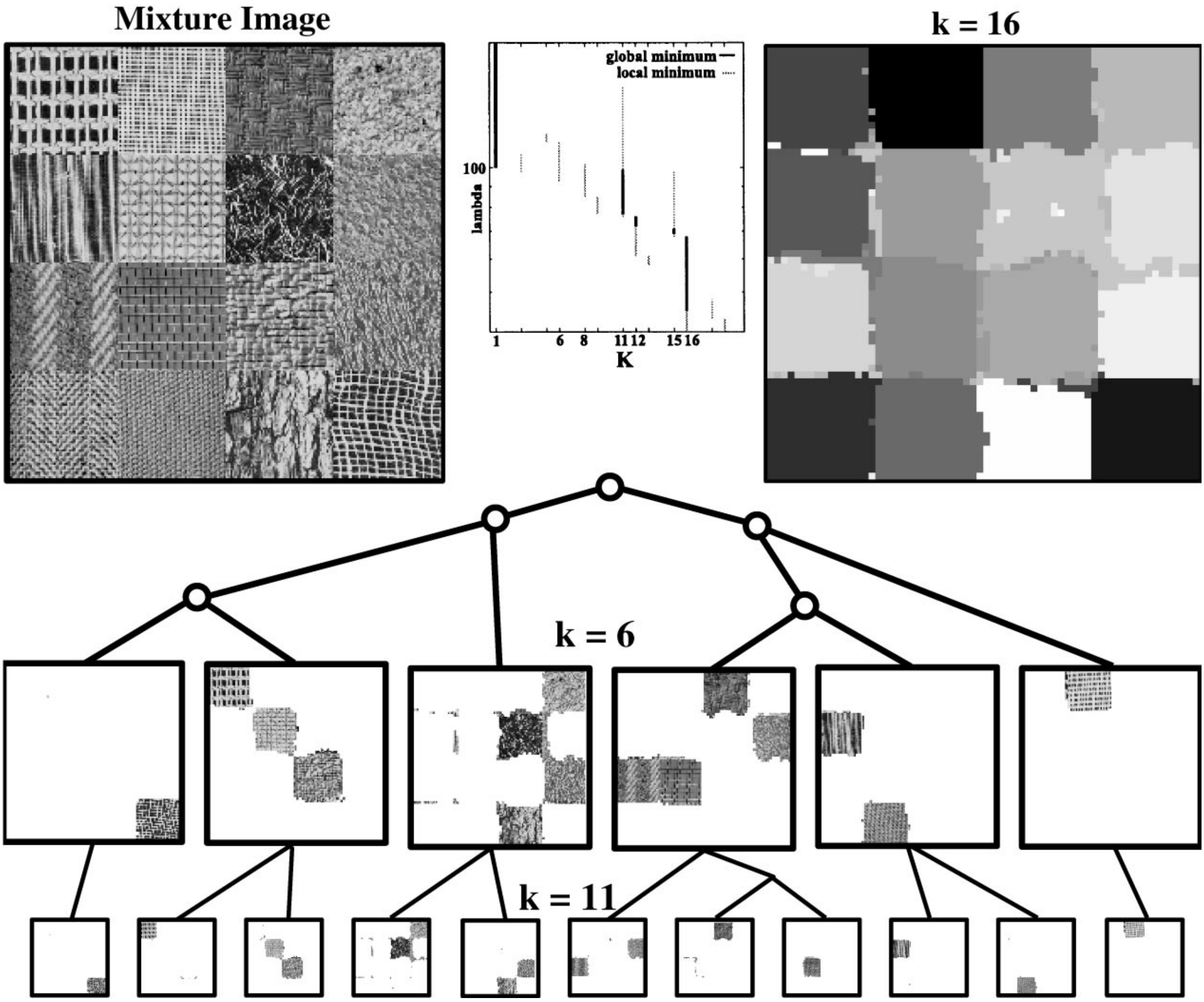

Fig. 5. Mixture image with 16 Brodatz micro-textures. For the segmentation $K = 24$ and $|\mathcal{N}|/2 = 150N$ evaluated dissimilarities were used.

points in the precision/recall plane. The measures are combined in terms of the so-called effectiveness [34], $E(\beta) = 1 - (1 + \beta^2)PR/(\beta^2 P + R)$, where $\beta$ weights precision versus recall. Since we are mainly interested in a comparison between different clustering solutions, we assume a simplified retrieval model, where the user interactively specifies the most promising cluster. Then, all documents in that cluster are returned. In a more realistic application this can be based on information on documents which are already known to be relevant, or on cluster summaries presented to the user (e.g., as shown in Fig. 8). Fig. 7 shows plots of $(P, R)$ pairs for solutions obtained by different clustering algorithms and different $K$ on the *MED* collection with ideal cluster search.

We summarize the most remarkable facts: (i) Among the linkage algorithms Ward's method shows consistently the best performance. (ii) Among the optimization methods, $\mathscr{H}^{\mathrm{pc1}}$ and $\mathscr{H}^{\mathrm{ps1b}}$ perform consistently better than the graph partitioning objective function $\mathscr{H}^{\mathrm{gp}}$, although the additive data shift has been empirically optimized. On coarser levels with small $K$ Ward's method performs better, while a global optimization of $\mathscr{H}^{\mathrm{pc1}}$ shows significant improvement for larger $K$. The reason for this behavior is the violation of the 'natural' data granularity for small $K$. In that regime the global maximization of cluster compactness leads to an unfavorable division of meaningful smaller clusters. If more documents should be returned at a lower reliability it might thus be advantageous to take a finer data partitioning and to additionally return documents of more than one cluster. Table 3 summarizes the effectiveness maximized over different $K$ for perfect retrieval ($E^*$) and best cluster match based on the query ($E$).

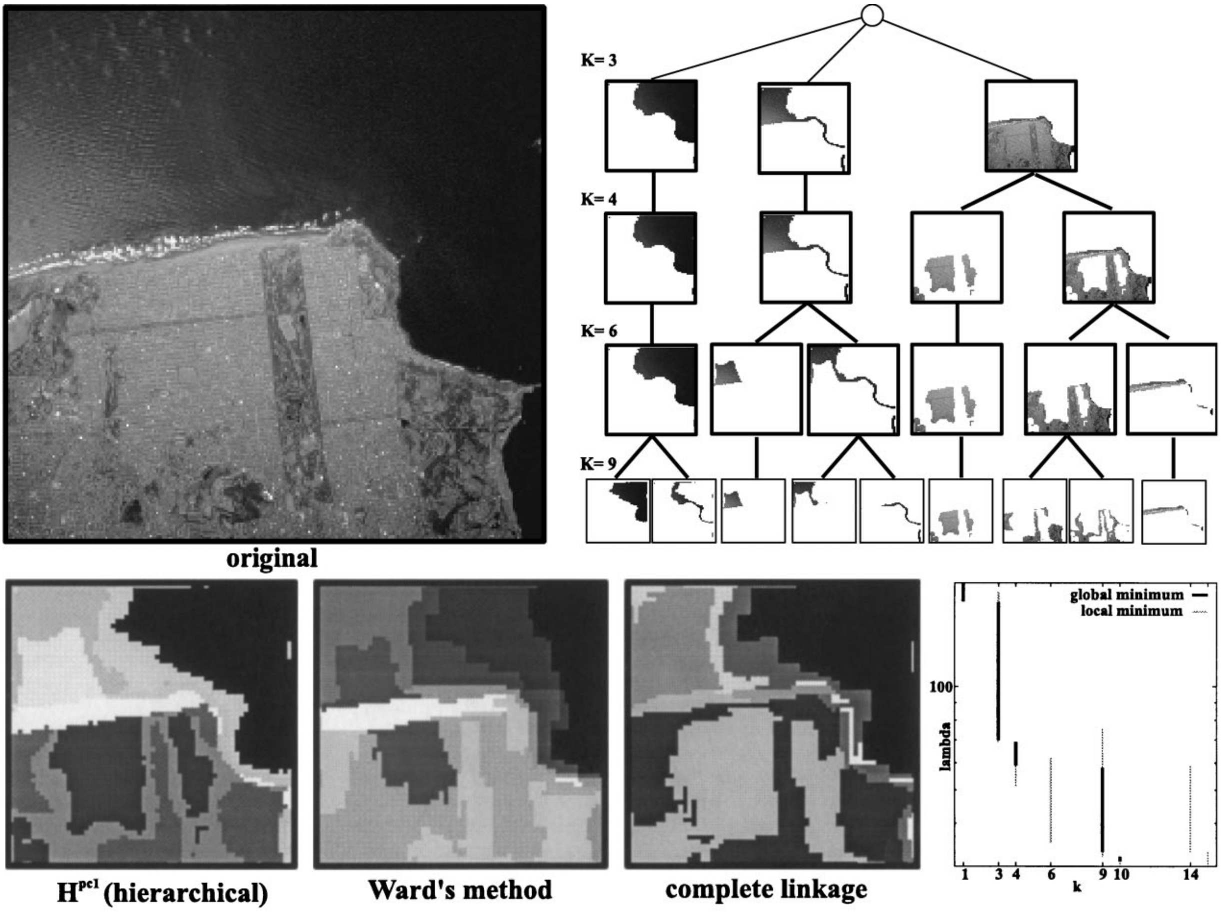

Fig. 6. Aerial image and hierarchical segmentation of a section of San Francisco.

As an illustrative example for hierarchical clustering of document databases we decided to cluster documents having the term *clustering* in their title. We collected 1568 abstracts from journal and conference papers. The top levels of a hierarchical solution with $K_{\max} = 60$ are visualized in Fig. 8. The clusters are characterized by terms having a high frequency of occurrence and being typical at the same time. More specifically, we utilized $t_v = p_v^2/\bar{p}$, where $p_v$ is the frequency inside a cluster $\mathscr{C}_v$ and $\bar{p}$ denotes the global frequency.

## 5. Conclusion

We have presented a rigorous optimization approach to similarity-based data clustering. The axiomatic approach attains a systematization of clustering criteria and yields theoretical insight which has proven to be highly relevant for practical applications. Our framework also provides a connection between graph partitioning optimization problems studied in operations research, and linkage algorithms like Ward's method known from cluster analysis. In particular, we have shown that partitional methods are not limited to vectorial data and a characterization of clusters by centroids, nor do they exclude the incomplete data case or nested cluster hierarchies. The second contribution of this paper concerns the derivation of efficient optimization heuristics based on annealing. These methods are essentially applicable to arbitrary clustering criteria, and are as universal as, for example, the agglomerative clustering scheme. Empirically, annealing methods have shown to yield significantly better solutions than local descend algorithms like ICM. Although they are not guaranteed to find the global minimum, the solutions found are often 'good enough', if the typical modeling uncertainty of unsupervised learning problems is taken into account (in the sense that the ground truth will most of the time not perfectly correspond to a global minimum of the objective function).

To stress the generality of our optimization approach we have presented two large-scale applications from very

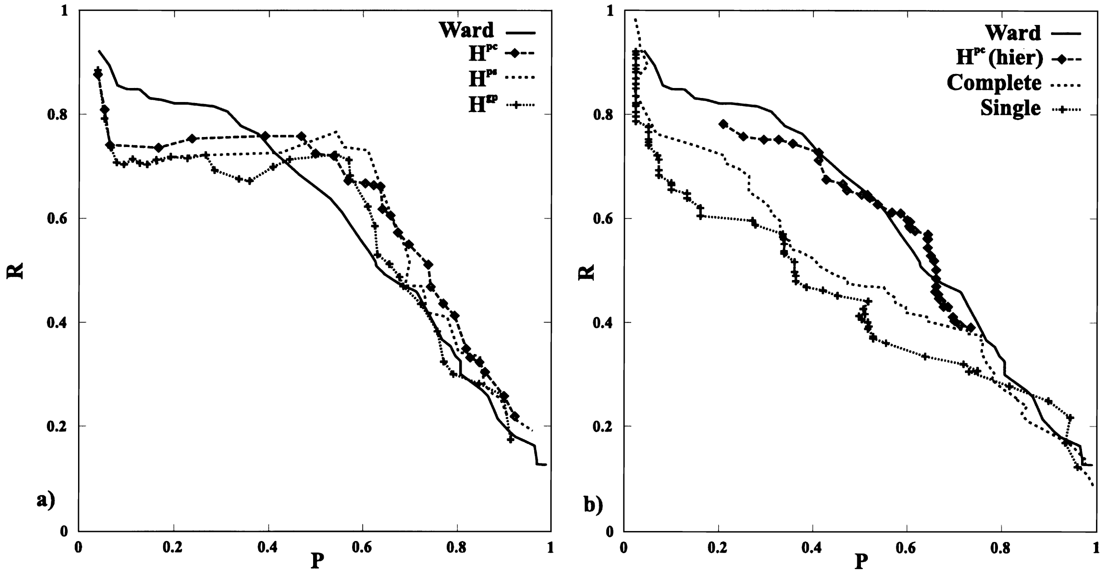

Fig. 7. Precision vs. recall results for different clustering algorithms on the *MED* collection. (a) Non-hierarchical optimizations method and Ward's method, (b) hierarchical methods.

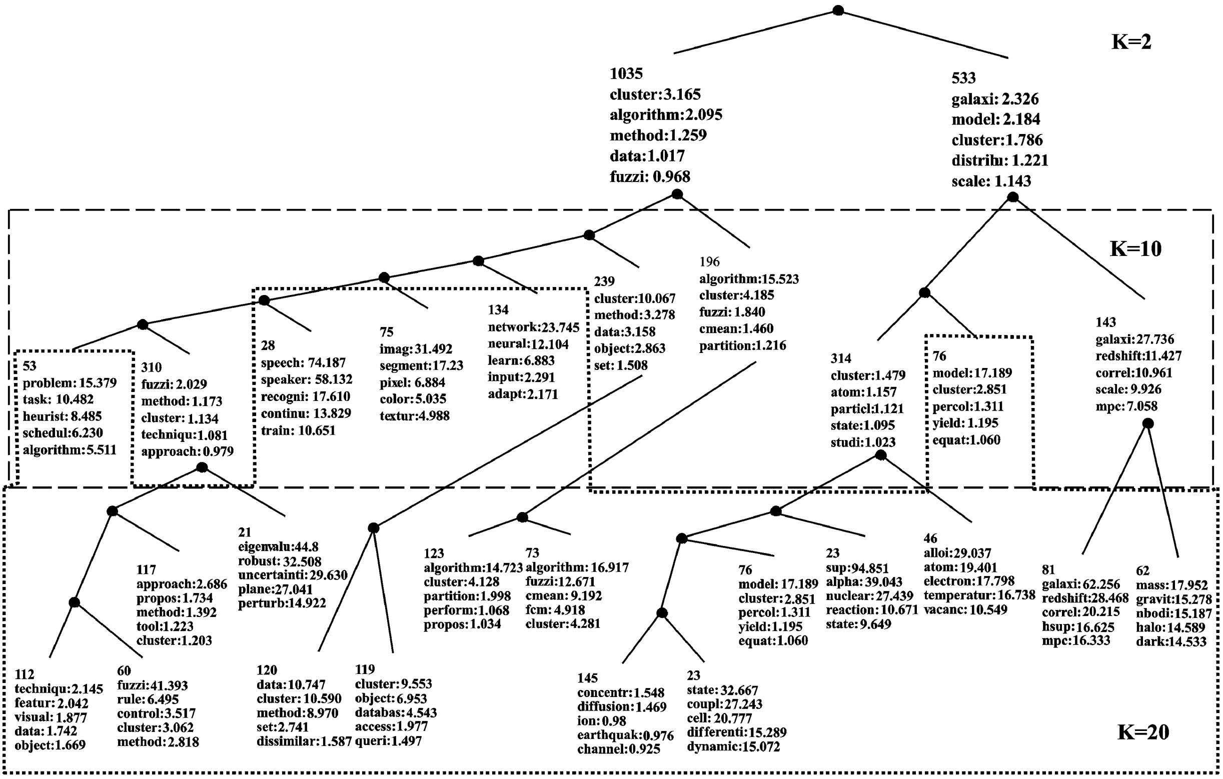

Fig. 8. Hierarchical clustering of 'clustering' documents. Numbers denote cluster sizes followed by the five most characteristic terms. The solutions $K = 2, 10, 20$ were selected according to the proposed pruning criterion.

Table 3
Effectiveness of document retrieval for different clustering models

| | $\mathscr{H}^{\text{pc1}}_{N,K}$ | $\mathscr{H}^{\text{pc1}}_{N}$ | $\mathscr{H}^{\text{ps1b}}_{N,K}$ | $\mathscr{H}^{\text{gp}}_{N,K}$ | Ward's | Complete | Single |
|---|---|---|---|---|---|---|---|
| $E^*(0.5)$ | **0.34** (70) | 0.38 (56) | 0.36 (60) | 0.38 (65) | 0.38 (84) | 0.41 (168) | 0.74 (426) |
| $E^*(1.0)$ | 0.37 (38) | 0.41 (56) | **0.35** (38) | 0.38 (40) | 0.45 (75) | 0.52 (140) | 0.76 (424) |
| $E^*(2.0)$ | 0.36 (36) | 0.42 (44) | **0.31** (32) | 0.31 (32) | 0.44 (33) | 0.54 (76) | 0.73 (426) |
| $E(0.5)$ | **0.61** (38) | 0.65 (66) | 0.66 (50) | 0.68 (45) | 0.69 (51) | 0.78 (78) | 0.94 (290) |
| $E(1.0)$ | **0.62** (38) | 0.69 (58) | 0.64 (32) | 0.67 (38) | 0.70 (39) | 0.82 (74) | 0.97 (268) |
| $E(2.0)$ | 0.61 (38) | 0.68 (16) | **0.59** (32) | 0.64 (38) | 0.68 (22) | 0.85 (70) | 0.97 (268) |

*Note*: Number in brackets denotes corresponding optimal value of $K$.

different application domains. The results on unsupervised texture segmentation show that similarity-based methods outperform other state of the art techniques. The data sparsening prevents an intractable scaling: even a large number of different textures can be reliable distinguished with reasonably small random graphs. In the context of document retrieval, where similarity-based clustering methods are commonly used, we have shown that optimization methods are a serious alternative to linkage algorithms and are able to identify meaningful document clusters. In contrast to agglomerative methods they have the further advantage not to require a complete re-computation if new documents are added to the database.

## Acknowledgements

This work was supported by the German Research Foundation DFG under grant # BU 914/3-1 and by the Federal Ministry of Education and Science BMBF under grant # 01 M 3021 A/4.

## Appendix

**Proof of Proposition 1.** For notational convenience denote by $\mathbf{M}_v$ the $v$th column of $\mathbf{M}$. We have to show that all possible dependencies of $\psi$ on its arguments take the form stated in the proposition. Therefore, we rewrite $\psi$ in a sequence of equalities, referring to the number of the axiom applied. For a given $\psi$ there exist functions $\hat{\psi}$, $\psi^{(1)}$ and $\psi^{(2)}$ such that

$$
\begin{aligned}
&\psi(i, j, D_{ij}, \mathbf{M})\\
&\quad =^{[1a]} \sum_{v,\mu=1}^{K} M_{iv}M_{j\mu}\hat{\psi}(D_{ij}, v, \mu, \mathbf{M})\\
&\quad =^{[2]} \sum_{v,\mu=1}^{K} M_{iv}M_{j\mu}\hat{\psi}(D_{ij}, v, \mu, \mathbf{M}_v, \mathbf{M}_\mu)\\
&\quad =^{[1b]} \sum_{v=1}^{K} M_{iv}M_{jv}\psi^{(1)}(D_{ij}, \mathbf{M}_v)\\
&\qquad - \sum_{\substack{v,\mu=1\\ v\neq\mu}}^{K} M_{iv}M_{j\mu}\psi^{(2)}(D_{ij}, \mathbf{M}_v, \mathbf{M}_\mu)\\
&\quad =^{[1a]} \sum_{v=1}^{K} M_{iv}M_{jv}\psi^{(1)}(D_{ij}, n_v)\\
&\qquad - \sum_{\substack{v,\mu=1\\ v\neq\mu}}^{K} M_{iv}M_{j\mu}\psi^{(2)}(D_{ij}, n_v, n_\mu).
\end{aligned}
$$

A reduced set of arguments for a function is used to indicate the corresponding invariance property of the function. For example, $\psi^{(1)}(D_{ij}, n_v)$ is defined by $\psi^{(1)}(Dij, n_v(\mathbf{M}_v)) = \psi^{(1)}(D_{ij}, \mathbf{M}_v)$ and, furthermore, indicate that for all $n_v(\mathbf{M}_v) = n_v(\hat{\mathbf{M}}_v)$: $\psi^{(1)}(D_{ij}, \mathbf{M}_v) = \psi^{(1)}(D_{ij}, \hat{\mathbf{M}}_v)$. The weighting functions $\psi^{(1)}$ and $\psi^{(2)}$ are non-decreasing in the first argument by Axiom 2. □

**Proof of Proposition 2.** From the shift invariance axiom we obtain the following condition:

$$
\begin{aligned}
&\sum_{v=1}^{K} \sum_{\substack{i,j=1\\ i\neq j}}^{N} M_{iv}M_{jv}\psi^{(1)}(n_v) = N\\
&\quad\Leftrightarrow \sum_{v=1}^{K} n_v(n_v - 1)\psi^{(1)}(n_v) = N.
\end{aligned}
$$

As will be proven in the subsequent lemma, $\sum_{v=1}^{K} f(n_v) = N$ requires $f$ to be an affine combination of $n_v$ and $N/K$. This implies $\psi^{(1)}(n_v)$ to be an affine combination of $\psi_1(n_v) = 1/(n_v - 1)$ and $\psi_2(n_v) = N/(Kn_v(n_v - 1))$. □

**Lemma A.1.** *Let* $f\colon \mathbb{R} \to \mathbb{R}$ *be a differentiable function, such that* $\sum_{v=1}^{K} f(n_v) = N$ *for all* $(n_1, n_2, \ldots, n_K) \in \mathbb{R}_+^K$ *with* $\sum_{v=1}^{K} n_v = N$. *Then* $f$ *can be written as an affine combination of* $n_v$ *and* $N/K$.

**Proof.** Calculating the directed derivative with $w \in \mathbb{R}^K$, $w_v = K - 1/K$ for an arbitrary but fixed $v$ and

$w_\mu = -1/K$, for all $\mu \neq \nu$, we obtain

$$\frac{\partial f(n_\nu)}{\partial n_\nu} = \frac{1}{K}\sum_{\mu=1}^{K}\frac{\partial f(n_\mu)}{\partial n_\mu}.$$

Since this has to hold for an arbitrary cluster index $\nu$, all the derivatives have in fact to be equal: $\partial f(n_\nu)/\partial n_\nu = \partial f(n_\mu)/\partial n_\mu$. The ansatz $f(n_\nu) = a \quad n_\nu + b$ yields

$$f(n_\nu) = \lambda\ n_\nu + (1-\lambda)\,\frac{N}{K} \text{ for } \lambda \in \mathbb{R}. \qquad \square$$

**Proof of Proposition 3.** The decomposition property of $\psi^{(2)}$ allows us to apply Lemma A.1 resulting in

$$f(n_\nu) = \lambda n_\nu + (1-\lambda)\,\frac{N}{K} \quad \text{with}$$

$$f(n_\nu) = n_\nu \sum_{\mu=1,\ \mu\neq\nu}^{K} n_\mu \psi^{(2)}(n_\nu, n_\mu)$$

and symmetric solutions obtained from interchanging the first two arguments of $\psi^{(2)}$.

Setting $\lambda = 1$ we obtain $\sum_{\mu\neq\nu} n_\mu \psi^{(2)}\ (n_\nu, n_\mu) = 1$. To consider the dependency on the second argument we calculate directional derivatives with $w_\mu = 1$, $w_\alpha = -1$, and $w_\gamma = 0$, otherwise, where $\alpha \neq \mu$. This yields

$$\frac{\partial n_\alpha \psi^{(2)}(n_\nu, n_\alpha)}{\partial n_\alpha} = \frac{\partial n_\mu \psi^{(2)}(n_\nu, n_\mu)}{\partial n_\mu}$$

$$\Rightarrow \psi^{(2)}(n_\nu, n_\mu) = a(n_\nu) + \frac{b(n_\nu)}{n_\mu}.$$

Inserting this back into the original condition in order to determine the functions $a$ and $b$ results in

$$(N - n_\nu)a(n_\nu) + (K-1)b(n_\nu) = 1$$

$$\Rightarrow a(n_\nu) = \frac{1}{N - n_\nu} \wedge b(n_\nu) = \frac{1}{K-1}.$$

A similar calculation is carried out for the case of $\lambda = 0$. The resulting functions $a$ and $b$ are given by

$$a(n_\nu) = \frac{N}{K}\,\frac{1}{(N-n_\nu)n_\nu}$$

and

$$b(n_\nu) = \frac{N}{K(K-1)}\,\frac{1}{n_\nu n_\mu}.$$

From these and the symmetric conditions for interchanging the first and second argument, we obtain 7 elementary weighting functions

$$\psi_1 = \frac{1}{N - n_\nu}, \quad \psi_5 = \frac{1}{N - n_\mu},$$

$$\psi_2 = \frac{1}{(K-1)n_\nu}, \quad \psi_6 = \frac{1}{(K-1)n_\mu},$$

$$\psi_3 = \frac{N}{K(N-n_\nu)n_\nu}, \quad \psi_7 = \frac{N}{K(N-n_\mu)n_\mu},$$

$$\psi_4 = \frac{N}{K(K-1)n_\nu n_\mu}. \qquad \square$$

## References

[1] A. Jain, R. Dubes, Algorithms for Clustering Data, Prentice-Hall, Englewood Cliffs, NJ 07632, 1988.

[2] J. MacQueen, Some methods for classification and analysis of multivariate observations, in Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 1967, pp. 281–297.

[3] G. Lance, W. Williams, A general theory of classification sorting strategies: II. Clustering systems, Comput. J. 10 (1969) 271–277.

[4] G. McLachlan, K. Basford, Mixture Models, Marcel Dekker, New York, Basel, 1988.

[5] J. Dunn, A fuzzy relative of the ISODATA process and its use in detecting compact, well-sperated clusters, J. Cybernet. 3 (1974) 32–57.

[6] S. Ahalt, P. Chen, D. Melton, Competitive learning algorithms for vector quantization, Neural Networks 3 (3) (1990) 277–290.

[7] J. Buhmann, H. Kühnel, Complexity optimized data clustering by competitive neural networks, Neural Comput. 5 (1) (1993) 75–88.

[8] P. Brucker, On the complexity of clustering problems, in: R. Henn, B. Korte, W. Oletti (Eds.), Optimierung und Operations Research, Lecture Notes in Economics and Mathematical Systems, Springer, Berlin, 1978, pp. 45–55.

[9] T. Hofmann, J.M. Buhmann, Pairwise data clustering by deterministic annealing, IEEE Trans. Pattern Anal. Mach. Intell. 19 (1) (1997) 1–14.

[10] S. Geman, D. Geman, Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images, IEEE Trans. Pattern Anal. Mach. Intell. 6 (6) (1984) 721–741.

[11] G. Bilbro, W. Snyder, Mean field approximation minimizes relative entropy, J. Opt. Soc. Amer. 8 (2) (1991) 290–294.

[12] K. Rose, E. Gurewitz, G. Fox, A deterministic annealing approach to clustering, Pattern Recognition Lett. 11 (11) (1990) 589–594.

[13] T. Hofmann, J. Puzicha, J. Buhmann, Deterministic annealing for unsupervised texture segmentation, in: Proceedings of the EMMCVPR'97, Lecture Notes in Computer Science, vol. 1223, Springer, Berlin, 1997, pp. 213–228.

[14] T. Hofmann, J. Puzicha, J. Buhmann, Unsupervised texture segmentation in a deterministic annealing framework,

IEEE Trans. Pattern Anal. Mach. Intell. 20 (8) (1998) 803–818.

[15] J. Ward, Hierarchical grouping to optimize an objective function, J. Amer. Statist. Assoc. 58 (1963) 236–244.

[16] M. Grötschel, Y. Wakabayashi, A cutting plane algorithm for a clustering problem, Math. Programm. Ser. B 45 (1989) 59–96.

[17] D. Geman, S. Geman, C. Graffigne, P. Dong, Boundary detection by constrained optimization, IEEE Trans. Pattern Anal. Mach. Intell. 12 (7) (1990) 609–628.

[18] J. Shi, J. Malik, Normalized cuts and image segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'97), 1997, pp. 731–737.

[19] T. Hofmann, Data clustering and beyond: a deterministic annealing framework for exploratory data analysis, Shaker Verlag, Ph.D. Thesis, 1997.

[20] T. Hofmann, J. Puzicha, J. Buhmann, An optimization approach to unsupervised hierarchical texture segmentation, in Proceedings of the IEEE International Conference on Image Processing (ICIP'97), 1997.

[21] S. Kirkpatrick, C. Gelatt, M. Vecchi, Optimization by simulated annealing, Science 220 (4598) (1983) 671–680.

[22] J. Besag, On the statistical analysis of dirty pictures, J. Roy. Statist. Soc. Ser. B 48 (1986) 25–37.

[23] A. Blake, A. Zisserman, Visual Reconstruction, MIT Press, Cambridge, MA, 1987.

[24] D. Miller, K. Rose, Hierarchical, unsupervised learning with growing via phase transitions, Neural Comput. 8 (8) (1996) 425–450.

[25] A. Jain, F. Farrokhnia, Unsupervised texture segmentation using Gabor filters, Pattern Recognition 24 (12) (1991) 1167–1186.

[26] O. Pichler, A. Teuner, B. Hosticka, A comparison of texture feature extraction using adaptive Gabor filtering, pyramidal and tree-structured wavelet transforms, Pattern Recognition 29 (5) (1996) 733–742.

[27] J. Mao, A. Jain, Texture classification and segmentation using multiresolution simultaneous autoregressive models, Pattern Recognition 25 (1992) 173–188.

[28] J. Daugman, Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters, J. Opt. Soc. Amer. A 2 (7) (1985) 1160–1169.

[29] J. Puzicha, T. Hofmann, J. Buhmann, Non-parametric similarity measures for unsupervised texture segmentation and image retrieval, in Proceedings of the Conference on Computer Vision and Pattern Recognition, 1997.

[30] P. Brodatz, Textures: A Photographic Album for Artists and Designers, Dover Publications, New York, 1966.

[31] K. Rose, E. Gurewitz, G. Fox, Vector quantization by deterministic annealing, IEEE Trans. Inform. Theory 38 (4) (1992) 1249–1257.

[32] J. Puzicha, J. Buhmann, Multiscale annealing for real-time unsupervised texture segmentation, Technical Report IAI-97-4, Institut für Informatik, Universität Bonn (a short version appeared in: Proceedings of ICCV'98, pp. 267–273), 1997.

[33] B. Hajek, Cooling schedules for optimal annealing, Math. Oper. Res. 13 (1988) 311–324.

[34] C. Van Rijsbergen, Information Retrieval, Butterworths, London, Boston, 1979.

[35] P. Willett, Recent trends in hierarchic document clustering: a critical review, Inform. Process. Manage. 24 (5) (1988) 577–597.

[36] P. Sneath, R. Sokal, Numerical Taxonomy, W.H. Freeman and Company, San Francisco, CA, 1973.

**About the Author**—JAN PUZICHA received the Diploma degree in Computer Science from the University of Bonn, Germany, in 1995. In November 1995, he joined the Computer Vision and Pattern Recognition group at the University of Bonn, where he is currently completing his Ph.D. Thesis on optimization methods for grouping and segmentation. His research interests include image processing, remote sensing, autonomous robots, data analysis, and data mining.

**About the Author**—THOMAS HOFMANN received the Diploma and Ph.D. degrees in Computer Science from the University of Bonn, in 1993 and 1997, respectively. His Ph.D. research was on statistical methods for exploratory data analysis. In April 1997 he joined the Center for Biological and Computational Learning at the Massachusetts Institute of Technology as a postdoctoral fellow. His research interests are in the areas of pattern recognition, neural networks, graphical models, natural language processing, information retrieval, computer vision, and machine learning.

**About the Author**—JOACHIM M. BUHMANN received a Ph.D. degree in theoretical physics from the Technical University of Munich in 1988. He held postdoctoral positions at the University of Southern California and at the Lawrence Livermore National Laboratory. Currently, he heads the research group on Computer Vision and Pattern Recognition at the Computer Science department of the University of Bonn, Germany. His current research interests cover statistical learning theory and its applications to image understanding and signal processing. Special research topics include exploratory data analysis, stochastic optimization, and computer vision.