Unsupervised Texture Segmentation in a Deterministic Annealing Framework

Thomas Hofmann, Jan Puzicha, *Student Member, IEEE*, and
Joachim M. Buhmann, *Member, IEEE*

Abstract—We present a novel optimization framework for unsupervised texture segmentation that relies on statistical tests as a measure of homogeneity. Texture segmentation is formulated as a *data clustering* problem based on sparse *proximity data*. Dissimilarities of pairs of textured regions are computed from a multiscale Gabor filter image representation. We discuss and compare a class of clustering objective functions which is systematically derived from invariance principles. As a general optimization framework, we propose deterministic annealing based on a mean-field approximation. The canonical way to derive clustering algorithms within this framework as well as an efficient implementation of mean-field annealing and the closely related Gibbs sampler are presented. We apply both annealing variants to Brodatz-like microtexture mixtures and real-word images.

Index Terms—Image segmentation, pairwise clustering, deterministic annealing, EM algorithm, Gabor filters.

1 Introduction

The *unsupervised segmentation* of textured images is widely recognized as a difficult and challenging computer vision problem. It can be applied to a multitude of important vision tasks, ranging from vision-guided autonomous robotics and remote sensing to medical diagnosis and retrieval in large image databases. While supervised methods rely on labeled data and the strong notion of optimal texture discrimination, the unsupervised approach does not require prior knowledge about the textures present in an image. Therefore, the central topic of unsupervised segmentation is the notion of *texture proximity*, based on a *general similarity measure* which is not class- or texture-specific. The fundamental goal of unsupervised texture segmentation is to solve the clustering problem of how to optimally partition an image into homogeneous regions.

Mimicking the strategy of supervised segmentation, the majority of unsupervised methods have formulated the segmentation problem in a feature-centered fashion, i.e., clustering is performed in a vector space. As a consequence, these approaches have to solve the difficult problem of specifying a *metric* that appropriately represents visual dissimilarities between textures in the chosen feature space [1], [2]. In contrast to this widely appreciated approach, we follow the ideas of Geman et al. [3] to avoid a vector space representation by utilizing *nonparametric statistical tests*. As we will show, statistical tests are reliable measures of local texture similarity which are generally applicable and do not require the usual substantial parameter tuning. Moreover, nonparametric tests are assessable in terms of statistical significance and have the important advantage to be model-free in the sense that the underlying probability distributions are not assumed to belong to a parametric model class.

Following the outlined procedure, there are three main solution steps which have to be distinguished:

1) The data generation stage concerns the *representation* of images and the details of how to apply statistical tests,
2) The modeling stage has to deal with the specification of a suitable *objective function* for proximity-based clustering.
3) We have to develop an *efficient optimization algorithm* in order to address the computational problem.

This paper provides novel contributions to all three challenges: Section 2 deals with the extraction of proximity data from a Gabor image representation. On the basis of empirical performance comparisons, we favor the χ^2-statistic over the Kolmogorov-Smirnov test proposed in [3].

In Section 3, we derive a novel class of clustering objective functions with fundamental invariance properties. As it turns out, the key idea is to choose an appropriate normalization in measuring cluster compactness. The main property, which distinguishes our approach from other graph partitioning schemes [3], [4], [5], is *shift invariance*, i.e., invariance with respect to additive shifts of the proximity scale. In particular, this yields a natural generalization of the *K*-means cost function [6] to proximity data.

Section 4 presents an introduction to the concept of *deterministic annealing*, a general framework to derive efficient heuristic algorithms for a variety of problems in combinatorial optimization and computer vision. Deterministic Annealing has been applied to the traveling salesman problem [7], graph partitioning [8], quadratic assignment and graph matching [9], [10], vector quantization [11], [12], surface

- *T. Hofmann is with the Center for Biological and Computational Learning, Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. E-mail: hofmann@ai.mit.edu.*
- *J. Puzicha and J.M. Buhmann are with Rheinische Friedrich-Wilhelms-Universität, Institut für Informatik III, Römerstrasse 164, D-53117 Bonn, Germany. E-mail: {jan; jb}@cs.uni-bonn.de.*

Manuscript received 27 Jan. 1997; revised 26 May 1998. Recommended for acceptance by R. Chellappa.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number 107097.

reconstruction [13], image restoration [14], [15], and edge detection [16]. A deterministic annealing approach for clustering and visualization of complete proximity data has been presented in [17]. More specifically, we use *mean-field theory* as an approximation principle [18], [19], [20], [21] to obtain computationally tractable algorithms. Astonishingly, deterministic annealing algorithms have only been derived independently for highly specific optimization instances despite these widespread research activities. As a major contribution, we derive a generic algorithm for the complete class of unconstrained partitioning and clustering cost functions. This includes a general convergence proof for asynchronous update schemes and a clarification of the intrinsic relationship between mean-field annealing and simulated annealing by Monte Carlo Gibbs sampling [22].

2 Image Representation and Proximity Evaluation

The differential structure of an image $I(\vec{x})$ is completely extracted by convolving the image with the Gaussian filter family [23]. In many applications, however, it is convenient to use filters, which are tuned to the features of interest, e.g., a particular spatial frequency $\vec{k}$. This tuning operation can be formalized [24] and leads in the case of frequency tuning to the family of complex Gabor filters [23]

$$G\left(\vec{x}, \sigma, \vec{k}\right) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\vec{x}^t\vec{x}/2\sigma^2} e^{i\vec{k}^t\vec{x}}, \tag{1}$$

where σ denotes a scale parameter depending on k. Gabor filters essentially perform a local Fourier analysis and are optimally localized in the sense of the fundamental uncertainty relation [25]. In addition to the theoretical justification in a scale space framework, Gabor filters have empirically proven to possess excellent discrimination properties for a wide range of textures [26], [27]. The multiscale representation of images with a Gabor filter bank is especially useful for unsupervised texture segmentation, where little is known a priori about the characteristic frequencies of occurring textures. In this work, we focus on the computation of a collection of scale space features $\{I_r\}$ which are defined by the modulus of the filter outputs,

$$I_r(\vec{x}) = \left| I(\vec{x}) * G\left(\vec{x}, \sigma_r, \vec{k}_r\right) \right|.$$

The vector $\vec{I}(\vec{x})$ of Gabor coefficients at a position $\vec{x}$ encodes information about the spatial relationship between pixels in the neighborhood of $\vec{x}$, but may not capture the complete characteristics of the texture. To overcome this deficit, we consider the (*weighted*) *empirical distribution* of Gabor coefficients in a window around $\vec{x}$,

$$f_r\left(t, \vec{x}\right) = \sum_{\vec{y}: t_{i-1} \le I_r(\vec{y}) < t_i} W_r\left(\left\|\vec{x} - \vec{y}\right\|\right) \Big/ \sum_{\vec{y}} W_r\left(\left\|\vec{x} - \vec{y}\right\|\right), \quad \text{for } t \in \left[t_{i-1}; t_i\right[\tag{2}$$

W_r is a nonnegative monotone decreasing window function centered at the origin and $t_0 = 0 < t_1 < \ldots < t_L$ is a suitable binning.[1] Simple choices for W_r are circular or squared areas with a constant weight inside and zero weight outside. Alternatively, radial symmetric functions could be used, e.g., Gaussians decaying with σ_r. The autocorrelation function of the Gabor wavelet coefficients varies with a typical length scale which is determined by the scale of the texture and the filter width. Following [1], the window size for each filter is thus chosen proportional to σ_r. Taking $f_r\left(\cdot; \vec{x}\right)$ as the empirical estimate of the density of an underlying texture-specific stationary probability distribution, the dissimilarity between textures at two positions $\vec{x}_i$ and $\vec{x}_j$ is evaluated independently for each Gabor filter channel. We apply a statistical test d based on the distribution of coefficients in either window, i.e.,

$$D_{ij}^{(r)} = d\left(f_r\left(\cdot; \vec{x}_i\right), f_r\left(\cdot; \vec{x}_j\right)\right).$$

Several nonparametric test statistics are available for the *two-sample problem* [28]. We have examined the performance of the mutual information [4], the Kolmogorov–Smirnov statistic [3], [29], tests of the Cramer/von Mises type, and the χ^2-statistic in detail [30], [31]. Empirically, the χ^2-test and the mutual information test have been shown to yield the best results. In the following, we focus on the χ^2-statistic which is defined by

$$D_{ij}^{(r)} = \sum_{k=1}^{L} \frac{\left(f_r\left(t_k, \vec{x}_i\right) - \hat{f}\left(t_k\right)\right)^2}{\hat{f}\left(t_k\right)}, \quad \text{with} \quad \hat{f}\left(t_k\right) = \frac{f_r\left(t_k, \vec{x}_i\right) + f_r\left(t_k, \vec{x}_j\right)}{2} \tag{3}$$

If the coefficients in windows around $\vec{x}_i$ and $\vec{x}_j$ are independent samples drawn from the same underlying distribution, $D_{ij}^{(r)}$ is χ^2-distributed with $L - 1$ degrees of freedom.

While a statistical test is a reliable measure to judge the dissimilarity of two samples of Gabor coefficients in a single channel, the question arises how to combine the independently evaluated comparisons. We have investigated Minkowski norms

$$D_{ij} = \left(\sum_r D_{ij}^{(r)p} \right)^{1/p},$$

including the limiting case of the maximum norm ($p = \infty$) as utilized in [29]. The Minkowski norm for small p is less sensitive to differences in single channels, and the choice of $p = 1$ empirically showed the best performance. Moreover, $p = 1$ is the natural choice for independent channel distributions. Alternatively, the χ^2-statistic can be evaluated for the joint probability distribution. This yields excellent results for large texture patches, but severely suffers from the difficulty to estimate the joint probability distribution for small sample sizes, a deficit which renders this approach inappropriate for image segmentation.

1. Forty equidistant bins adapted to the dynamic range of the respective feature channel are used in the experiments. The results, however, are insensitive w.r.t. binning details.

Since we do not calculate a vector space representation of textures, but directly evaluate proximities between pairs of sites $\vec{x}_i, \vec{x}_j \left(1 \le i, j \le N\right)$ instead, the data clustering problem has to rely on the proximity matrix $\mathbf{D} = \left(D_{ij}\right) \in \mathbb{R}^{N \times N}$. The scaling of $\mathbf{D}$ is the major reason why proximity-based approaches have often been ignored as serious alternatives in image segmentation. While vector-valued data with a fixed number of features scale linearly with the number of pixels N, pairwise comparison results in a scaling with N^2. On the other hand, it is obvious that a complete proximity matrix encodes a significant inherent redundancy. We take advantage of this redundancy and introduce sparseness of the data in two respects:

1) Local histograms $f_r\left(t; \vec{x}\right)$ are only evaluated at positions $\vec{x}_i$ on a regular image sublattice.[2]
2) Dissimilarity values are only computed for a substantially reduced, small fraction of pairs of sites.

A convenient way to represent sparse proximity matrices are weighted loop-free graphs $\mathcal{G} = (V, E, D)$ with vertices $V = \{1, ..., N\}$, edges $E \subset V \times V$ and weights D_{ij} for edges $e = (i, j)$. Following [3], we call the index sets $\mathcal{N}_i = \{j : (i, j) \in E\}$ the neighborhood of site $\vec{x}_i$ and define $\mathcal{N}_i$ to consist of the four-connected neighborhood of $\vec{x}_i$ in the image and a larger number of random neighbors.

3 Clustering of Proximity Data

As we have pointed out, unsupervised segmentation is essentially a partitioning or clustering problem of labeling a set of image sites with group or texture labels l_ν $(1 \le \nu \le K)$. For the derivation of suitable objective functions, we assume that the number of distinct textures K is fixed. The determination of K is achieved by a heuristic model selection criterion (Section 5.1). For notational convenience, $M_{i\nu} \in \{0, 1\}$ denotes an indicator function for the assignment of image site $\vec{x}_i$ to label l_ν. The set of all indicator variables is summarized in terms of a Boolean matrix $\mathbf{M} \in \mathcal{M}$, where

$$\mathcal{M} = \left\{ \mathbf{M} \in \left\{0,1\right\}^{N \times K} : \sum_{\nu=1}^{K} M_{i\nu} = 1, 1 \le i \le N \right\}.$$

The main problem from the modeling perspective is to specify an objective function $\mathcal{H} : \mathcal{M} \to \mathbb{R}$ that assesses the quality of a given image partitioning $\mathbf{M}$ in a semantically meaningful way. In this paper, we focus on objective functions that measure the *intracluster compactness* and depend only on the homogeneity of a cluster. In greater detail, we investigate the following type of additive objective functions, which are linear in the dissimilarities D_{ij},

$$\mathcal{H}(\mathbf{M}) = \sum_{\nu=1}^{K} \sum_{i=1}^{N} M_{i\nu} \frac{\sum_{j \in \mathcal{N}_i} M_{j\nu} D_{ij}}{\eta_{i\nu}\left(\vec{M}_\nu, \mathcal{G}\right)}. \tag{4}$$

2. Typically of size 64×64 or 128×128, thus $N = 4{,}096$ or $16{,}384$.

This type of intracluster measure additively combines contribution for each site $\vec{x}_i$. The site-specific score for $\vec{x}_i$ consists of a sum over all known dissimilarities to sites $\vec{x}_j$ belonging to the same cluster, divided by a data-independent normalizing constant $n_{i\nu}\left(\vec{M}_\nu, \mathcal{G}\right)$. The choice of the normalization constant $n_{i\nu}$ is further limited by the requirement that $\mathcal{H}$ has to be invariant with respect to permutations of data indices and cluster labels. Thus, $n_{i\nu}$ is restricted to the following functional form,[3]

$$n_{i\nu}\left(\vec{M}_\nu, \mathcal{G}\right) = n\left(p_{i\nu}, P_\nu, Q_\nu\right),$$
$$p_{i\nu} = \sum_{j \in \mathcal{N}_i} M_{j\nu},\ P_\nu = \sum_{i=1}^{N} M_{i\nu},\ Q_\nu = \sum_{(i,j) \in E} M_{i\nu} M_{j\nu} \tag{5}$$

While P_ν denotes the cluster size, $p_{i\nu}$ is the number of known comparisons between site $\vec{x}_i$ and sites in the cluster with label l_ν. Q_ν corresponds to the total number of computed comparisons in a cluster. By convention, we set $\mathcal{H}(\mathbf{M}) = +\infty$ if $n_{i\nu} = 0$. The functional form in (5) is still too general to single out particularly interesting candidate cost functions. We will, therefore, put forward an invariance requirement, namely the invariance of $\mathcal{H}$ with respect to affine transformations of the dissimilarity scale,

$$\mathcal{H}(\mathbf{M}; a\mathbf{D} + c) = a\mathcal{H}(\mathbf{M}; \mathbf{D}) + Nc. \tag{6}$$

We believe this invariance to be an important modeling assumption mainly for the following reason: Every noninvariant objective function necessarily makes assumptions about specific properties of the dissimilarity measure, i.e., it gives a meaning to the *scale* and/or *origin* of the data. To avoid dependencies on the units in which dissimilarities are measured seems to be an obvious benefit. In fact, the reader should notice that the linearity of $\mathcal{H}$ in (4) already implies scale-invariance. The advantages of shift invariance are perhaps less evident, since one may wonder why the self-similarity D_{ii} (at least if it is unique) should not be a natural choice of the origin, e.g., $D_{ii} = 0$ for the χ^2-test. However, in the context of statistical tests, one may as well define the average value of D_{ij} for sites x_i and x_j belonging to the same texture as the "natural" origin, which in the present context creates a dependency on binning details and the sample size. The advantage of shift invariance is that the relative quality of two data partitions only depends on differences between proximities and not on their absolute values. Shift invariance, therefore, is of crucial importance. Since every objective function has to weight up dissimilarities of different magnitude, it is important to have a controlled way to solve this trade-off. The following proposition gives a definitive answer about objective functions possessing the shift invariance property. The proof is given in the Appendix.

PROPOSITION 1. *All shift-invariant objective functions of the functional form as given by* (4) *and* (5) *can be expressed as*

3. A rigorous proof requires additional technical conditions. The main idea is that the L_1 norm is the only permutation invariant quantity of a Boolean vector (i.e., all other invariances are functions of the L_1 norm).

linear combinations of four elementary functions with normalizations $n_{iv} = p_{iv}$, $n_{iv} = Q_v / P_v$, $n_{iv} = p_{iv} P_v$, *and* $n_{iv} = Q_v$.

For all cost functions discussed in the sequel, the data dependency is summarized either by averages A_v or B_v, where

$$A_v = \frac{\sum_{i=1}^{N} M_{iv} a_{iv}}{\sum_{i=1}^{N} M_{iv}}, \quad a_{iv} = \frac{\sum_{j \in \mathcal{N}_i} M_{jv} D_{ij}}{\sum_{j \in \mathcal{N}_i} M_{jv}}, \quad \text{and}$$

$$B_v = \frac{\sum_{(i,j) \in E} M_{iv} M_{jv} D_{ij}}{\sum_{(i,j) \in E} M_{iv} M_{jv}} \tag{7}$$

Thus, one of the two aspects which distinguishes the remaining four candidate objective functions is related to the data sparseness. Average dissimilarities are either calculated in a two-stage procedure by first calculating site-specific averages a_{iv}, which are then averaged over all sites belonging to the same cluster, or by directly evaluating average intracluster dissimilarities as in B_v. Both variants of averaging are equivalent in the limit of complete proximity matrices and, therefore, are different generalizations of complete data objective functions to the case of sparse proximity matrices. The more fundamental distinction between objective functions is concerned with the weighting of averages A_v or B_v for different clusters.

1) The first type of invariant cost functions is obtained by weighting every cluster proportional to its size, $\mathcal{H}_A^{\text{pro}}(\mathbf{M}, \mathbf{D}) = \sum_{v=1}^{K} P_v A_v$ and $\mathcal{H}_B^{\text{pro}}(\mathbf{M}, \mathbf{D}) = \sum_{v=1}^{K} P_v B_v$. Thus, the cost contributions for a site do not depend on the size of the cluster.
2) The second type of invariant cost functions combines the contributions with constant weights irrespective of the size of the clusters, $\mathcal{H}_A^{\text{con}}(\mathbf{M}, \mathbf{D}) = \frac{N}{K} \sum_{v=1}^{K} A_v$ and $\mathcal{H}_B^{\text{con}}(\mathbf{M}, \mathbf{D}) = \frac{N}{K} \sum_{v=1}^{K} B_v$. This results in single-site cost contributions which are inversely proportional to the cluster size.
3) A weighting proportional to the number of known dissimilarities in a cluster,

$$\mathcal{H}^{\text{gp}}(\mathbf{M}, \mathbf{D}) = \sum_{v=1}^{K} Q_v B_v = \sum_{v=1}^{K} \sum_{(i,j) \in E} M_{iv} M_{jv} D_{ij},$$

corresponds to the standard cost function for graph-partitioning problems and has been proposed in [3] for texture segmentation, but does not result in a shift-invariant function.
4) For completeness, we mention the proposal made in [5] called *normalized cut* with an objective function given by

$$\mathcal{H}^{\text{nc}}(\mathbf{M}, \mathbf{D}) = \sum_{v=1}^{K} \frac{\sum_{(i,j) \in E} M_{iv} M_{jv} D_{ij}}{\sum_{(i,j) \in E} M_{iv} D_{ij}},$$

which is not of the type defined in (4) and is also not shift invariant.

The principle of shift invariance has been the major guideline for the derivation of normalized clustering objective functions. Let us complete the above argument by pointing out some disadvantages one may encounter with a noninvariant objective function like $\mathcal{H}^{\text{gp}}$. For example, $\mathcal{H}^{\text{gp}}$ applied to graphs with nonnegative weights favors equipartitionings, whereas in the opposite case, the formation of large clusters is advantageous. Indeed, it has been noticed before [3] that the data have to be shifted adequately in order to keep the right balance between negative and positive contributions. However, if a large number of different textures exist in an image, it is often impossible to globally shift the data, such that all textures are well-discriminated by the objective function $\mathcal{H}^{\text{gp}}$. We have empirically verified these arguments in our simulations (cf. Fig. 4).

While the cost functions $\mathcal{H}_A^{\text{con}}$ and $\mathcal{H}_B^{\text{con}}$ are shift invariant, they have the disadvantage to favor unbalanced partitionings. They lack robustness in the sense that even in the limit of $N \to \infty$ globally optimal partitionings may include "nonmacroscopic" small clusters, e.g., "pair-clusters" with $P_v = 2$. This problem can be moderated by adding prior costs to penalize small clusters, $\mathcal{H}^{\text{sz}}(\mathbf{M}) = \lambda_s N^2 \sum_{v=1}^{K} P_v^{-1}$. However, this requires an estimate of the (scale-dependent) parameter λ_s, leading to similar problems as in the case of $\mathcal{H}^{\text{gp}}$. We conclude, that only the cost functions $\mathcal{H}_A^{\text{pro}}$ and $\mathcal{H}_B^{\text{pro}}$ are fully satisfying from a theoretical point of view in that they are robust and possess all important invariance properties.

To further stress this conclusion, notice that if the proximity data were generated from a Euclidean vector space representation by $D_{ij} = \left(\vec{v}_i - \vec{v}_j\right)^2$, then the complete data case would give

$$\mathcal{H}_A^{\text{pro}}(\mathbf{M}, \mathbf{D}) = \mathcal{H}_B^{\text{pro}}(\mathbf{M}, \mathbf{D}) \cong \mathcal{H}^{\text{km}}(\mathbf{M}, \mathbf{D}),$$

with

$$\mathcal{H}^{\text{km}}(\mathbf{M}, \mathbf{D}) = \sum_{v=1}^{K} \sum_{i=1}^{N} M_{iv} \left(\vec{v}_i - \vec{y}_v\right)^2,$$

and the usual centroid definition

$$\vec{y}_v = \sum_{j=1}^{N} M_{jv} \vec{v}_j \Big/ \sum_{j=1}^{N} M_{jv}.^4$$

This is a key argument in favor of $\mathcal{H}^{\text{pro}}$ as the K-means cost function since $\mathcal{H}^{\text{km}}$ is generally considered to be an ade-

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

quate clustering criterion for the case of squared Euclidean distances.

In addition to the data-dependent clustering costs, we propose to include prior costs about plausible image labelings which are simply added to the clustering objective function. Since image segments for natural scenes are expected to form connected components, it is reasonable to assume that sites in the topological neighborhood $\mathcal{T}_i$ of a site $\vec{x}_i$ should have a higher probability to be assigned to the same texture, which is reflected by the choice

$$\mathcal{H}^{\text{top}}(\mathbf{M}) = \lambda_t \sum_{\nu=1}^{K} \sum_{i=1}^{N} M_{i\nu} \sum_{j \in \mathcal{T}_i} \left(1 - M_{j\nu}\right).$$

We have furthermore added some hard constraints about valid image partitionings, excluding very small and thin regions as described in [3]. Since additional hard constraints restrict the development of efficient optimization algorithms and may lead to forbiddingly heavy computational load [32], we enforce these constraints in a separate postprocessing stage which follows the clustering procedure and determines the closest valid partitioning by eliminating components and smoothing borders, if necessary. In practice, the results of the clustering stage, in most cases, are good enough to obtain valid solutions within a few additional sweeps. As a heuristic procedure to determine the correct number of clusters, a contribution which penalizes the model complexity is included in the final cost function. A simple but well-performing penalty term proportional to the number of clusters K is $\mathcal{H}^{\text{cmp}} = \lambda_c NK$. Since $\mathcal{H}^{\text{cmp}}$ is independent of the configuration $\mathbf{M}$, it is used as a criterion for model comparison after clustering solutions with different K have been determined.

4 Clustering Algorithms

In seminal papers, Kirkpatrick et al. [33] and, independently, Cerny [34] have proposed the stochastic optimization strategy *Simulated Annealing*. Simulated Annealing determines solutions to combinatorial optimization problems by a random search, which is formally modeled by an inhomogeneous discrete-time Markov chain. Representing partitionings with Boolean matrices $\mathbf{M} \in \mathcal{M}$, the Markov chain is a sequence of finite random variables $\left(\mathbf{M}^{(t)}\right)_{t \in \mathbb{N}}$, which is completely specified by state transition probabilities

$$S_t\left(\tilde{\mathbf{M}}, \mathbf{M}\right) = \mathbf{P}\left(\mathbf{M}^{(t+1)} = \tilde{\mathbf{M}} \,\middle|\, \mathbf{M}^{(t)} = \mathbf{M}\right)$$

and the initial probability distribution at $t = 0$. Since the configuration space for partitioning problems decomposes naturally into single-site configurations $\mathcal{M} = \otimes_i \mathcal{M}_i$, we consider a restricted class of *local* Markov chains, which perform only state transitions between configurations, which differ in the assignment of at most one site, i.e.,

$$S_t\left(\tilde{\mathbf{M}}, \mathbf{M}\right) = 0, \text{ if } \left\|\mathbf{M} - \tilde{\mathbf{M}}\right\|_1 > 2.$$

We denote by $s_i\left(\mathbf{M}, \vec{e}_\nu\right)$ a locally modified assignment matrix, which is obtained by substituting the ith row of $\mathbf{M}$ by the unit vector $\vec{e}_\nu$. Moreover, we define a *site visitation schedule* as an infinite sequence of site indices $\nu : \mathbb{N} \to \{1, \ldots, N\}$ with

$$\lim_{U \to \infty} \#\{t \le U : \nu(t) = i\} \to \infty, \ \forall\, i \in \{1, \ldots, N\}.$$

For a given site visitation schedule, the Gibbs sampler [22] is defined by the finite transition probabilities

$$S_t\left(s_i\left(\mathbf{M}, \vec{e}_\nu\right), \mathbf{M}\right) = \frac{\exp\left[-g_{i\nu} \,/\, T(t)\right]}{\sum_{\mu=1}^{K} \exp\left[-g_{i\mu} \,/\, T(t)\right]},$$
$$\text{where } g_{i\nu} = \mathcal{H}\left(s_i\left(\mathbf{M}, \vec{e}_\nu\right)\right) \tag{8}$$

and $i = \nu(t)$. The Gibbs sampler samples from the conditional distribution at site $\nu(t)$ given the assignments at all other sites $\left\{\vec{x}_j : j \neq \nu(t)\right\}$. The Markov chain defined by (8) fulfills the detailed balance condition for constant $T(t)$ and, therefore, converges toward its equilibrium distribution known as the *Gibbs distribution*

$$P_{\mathcal{H}}(\mathbf{M}) = \frac{1}{\mathcal{Z}_T} \exp\left(-\mathcal{H}(\mathbf{M}) \,/\, T\right),$$
$$\mathcal{Z}_T = \sum_{\mathbf{M} \in \mathcal{M}} \exp\left(-\mathcal{H}(\mathbf{M}) \,/\, T\right) \tag{9}$$

The normalization constant $\mathcal{Z}_T$ is called the *partition function*. An important extremal property of the Gibbs distribution is the fact that it maximizes the entropy for fixed expected costs, cf. [35]. More formally, the Gibbs distribution at temperature T minimizes the *generalized free energy*

$$\mathcal{F}_T(P) = \left\langle \mathcal{H} \right\rangle_P - T\, S(P)$$
$$= \sum_{\mathbf{M} \in \mathcal{M}} P(\mathbf{M}) \mathcal{H}(\mathbf{M}) + T \sum_{\mathbf{M} \in \mathcal{M}} P(\mathbf{M}) \log P(\mathbf{M}) \tag{10}$$

over the space of probability distributions

$$\mathcal{P}_{\mathcal{M}} = \{P : \mathcal{M} \to [0, 1] : \textstyle\sum_{\mathbf{M} \in \mathcal{M}} P(\mathbf{M}) = 1\}.$$

The value of $\mathcal{F}_T(P_{\mathcal{H}})$ which is the minimum of the generalized free energy is simply called the *free energy* and is given by $\mathcal{F}_T(P_{\mathcal{H}}) = -T \log \mathcal{Z}_T$. The basic idea of annealing is to use Monte Carlo sampling, but to gradually lower the temperature $T(t)$, on which the transition probabilities depend. It has been proven [22], that for a logarithmic annealing schedule $T(t) = c/(1 + \log t)$ the Gibbs sampler converges in probability to the uniform distribution on the global minima of $\mathcal{H}$. Of course, in practice, annealing schedules always use a decay rate for T, which is too fast to guarantee convergence to a global minimum. For the zero temperature limit, a deterministic greedy optimization algorithm known as *Iterative Conditional Mode* (ICM) [36] is obtained.

While the general convergence results for simulated annealing [37] demonstrate the universality of this optimization principle, the inherently slow convergence of stochastic techniques compared to deterministic algorithms is per-

ceived as a major disadvantage. Therefore, we advocate to use a different, purely deterministic approach known as *deterministic annealing* [11]. Deterministic annealing combines the advantages of a temperature controlled continuation method with a fast, purely deterministic computational scheme. To stress the general ability to canonically derive heuristic algorithms for partitioning problems, we abstract from the details of $\mathcal{H}$ and present results which apply to arbitrary partitioning objective functions. The key idea of deterministic annealing is to analytically calculate the free energy, which completely characterizes the thermodynamic equilibrium in terms of a moment generating function. For the clustering objective functions which couple assignments of different sites, this calculation can only be performed approximately by minimizing the generalized free energy $\mathcal{F}_T$ over a tractable subspace $\mathcal{Q}_{\mathcal{M}} \subseteq \mathcal{P}_{\mathcal{M}}$. In the *mean-field approximation*, $\mathcal{Q}_{\mathcal{M}}$ is chosen to be the space of all factorial distributions

$$\mathcal{Q}_{\mathcal{M}} = \left\{ Q \in \mathcal{P}_{\mathcal{M}} : Q(\mathbf{M}) = \prod_{i=1}^{N} \sum_{\nu=1}^{K} M_{i\nu} q_{i\nu}, \forall \mathbf{M} \in \mathcal{M} \right\}. \quad (11)$$

The $q_{i\nu} \in [0, 1]$ are $N \cdot K$ parameters, which uniquely determine Q. The above approximation yields a procedure which is known as *mean-field annealing*, since it combines the mean-field approximation with the annealing principle [8], [14]. The advantage of annealing in the context of deterministic or mean-field annealing is to track solutions from high temperatures where $\mathcal{F}_T$ is convex, to low temperatures where we can canonically recover a local minimum of $\mathcal{H}$. The approximation quality can be expressed (though not efficiently computed) in terms of the cross entropy,

$$\mathcal{F}_T(Q) - \mathcal{F}_T\left(P_{\mathcal{H}}\right) = \frac{1}{T} \mathcal{I}\left(Q \middle\| P_{\mathcal{H}}\right),$$

which establishes the equivalence of minimizing the generalized free energy and minimizing the cross entropy to the Gibbs distribution [20].

Stationary conditions for (10) yield a system of coupled transcendental, so-called *mean-field equations*, which can be efficiently solved by a convergent iteration scheme. The following statements, which are proven in the Appendix, summarize the most important results for factorial distributions. A more detailed presentation can be found in [30], [38].

THEOREM 1. *Let $\mathcal{H}$ be an arbitrary partitioning cost function, $\mathcal{H} : \mathcal{M} \to \mathbb{R}$. The factorial distribution $Q^* \in \mathcal{Q}_{\mathcal{M}}$, which minimizes the generalized free energy $\mathcal{F}_T$ over $\mathcal{Q}_{\mathcal{M}}$, is characterized by the stationary conditions*

$$q_{i\nu}^* = \frac{\exp\left[-\frac{1}{T} h_{i\nu}\right]}{\sum_{\mu=1}^{K} \exp\left[-\frac{1}{T} h_{i\mu}\right]},$$

$$\text{where } h_{i\nu} = \frac{\partial \left\langle \mathcal{H} \right\rangle_Q}{\partial q_{i\nu}} \Big|_{Q=Q^*} = \frac{1}{q_{i\nu}^*} \left\langle M_{i\nu} \mathcal{H} \right\rangle_{Q^*} \quad (12)$$

Notice, that the right-hand side is independent of $\left\{q_{i1}^*, \ldots, q_{iK}^*\right\}$.

COROLLARY 1. *For any site-visitation schedule ν and arbitrary admissible initial conditions, the following asynchronous update scheme converges to a one-site optimal local minimum of the generalized free energy* (10):

$$q_{i\nu}^{\text{new}} = \frac{\exp\left[-\frac{1}{T} h_{i\nu}\right]}{\sum_{\mu=1}^{K} \exp\left[-\frac{1}{T} h_{i\mu}\right]} \quad (13a)$$

$$\text{where } h_{i\nu} = \frac{\partial \left\langle \mathcal{H} \right\rangle}{\partial q_{i\nu}} \Big|_{Q=Q^{\text{old}}} \text{ and } i = \nu(t). \quad (13b)$$

COROLLARY 2. *A connection to the Gibbs sampler in* (8) *is established by the identity $h_{i\nu} = \left\langle g_{i\nu} \right\rangle_{Q^*}$.*

Theorem 1 is an extension of the results obtained in [8] for the graph partitioning cost function $\mathcal{H}^{\text{gp}}$. In regard of Corollary 1, it is important to distinguish clearly between two numerical methods to find solutions of the stationary equations: a synchronous update of all variational parameters $\{q_{i\nu}\}$ given their previous values and an asynchronous update with respect to a site visitation schedule. Most of the work on mean-field annealing has either favored the synchronous update (e.g., [39], [20], [40]) or has at least been indifferent to this distinction (e.g., [8], [14], [16]). The synchronous scheme has the advantage of being amenable to a parallel implementation as already proposed by [39]. On the other hand, it has commonly been observed that synchronous updating may lead to instabilities, a fact that has recently been investigated more systematically for a simple Ising system based on the contraction mapping theorem [40]. The asynchronous update scheme in contrast is guaranteed to converge for arbitrary partitioning objective functions. Corollary 2 establishes a tight relationship between the quantities $g_{i\nu} = \mathcal{H}\left(s_i\left(\mathbf{M}, \vec{e}_\nu\right)\right)$ involved in implementing the Gibbs sampler in (9) and the mean-field equations. Thus, the mean-field $h_{i\nu}$ is a Q-averaged version of the local costs $g_{i\nu}$.

To obtain an efficient implementation of the Gibbs sampler, we have to find a way to efficiently calculate and update $g_{i\nu}$ under single-site changes. The mean-field algorithm (see Algorithm 1) in addition requires performing the Q-averages $\left\langle g_{i\nu} \right\rangle_Q$. Since Q is factorial, averages of higher-order products of Boolean assignment variables for different sites are easily obtained from products of marginal probabilities, i.e.,

$$\left\langle \prod_{i \in I} M_{i\nu} \right\rangle_Q = \prod_{i \in I} \left\langle M_{i\nu} \right\rangle_Q = \prod_{i \in I} q_{i\nu},$$

with $I \subset \{1, ..., N\}$. A straightforward implementation of the Gibbs sampler would partition the cost function into a sum of clique potentials and recalculate at each step all potentials of cliques to which site $\vec{x}_i$ with $i = \nu(t)$ belongs [22]. This procedure is not efficient for the normalized clustering objective functions, since the assignments enter in the denominator. We, therefore, propose to subtract the costs of the reduced system without site $\vec{x}_i$ from the new state which defines the local update costs

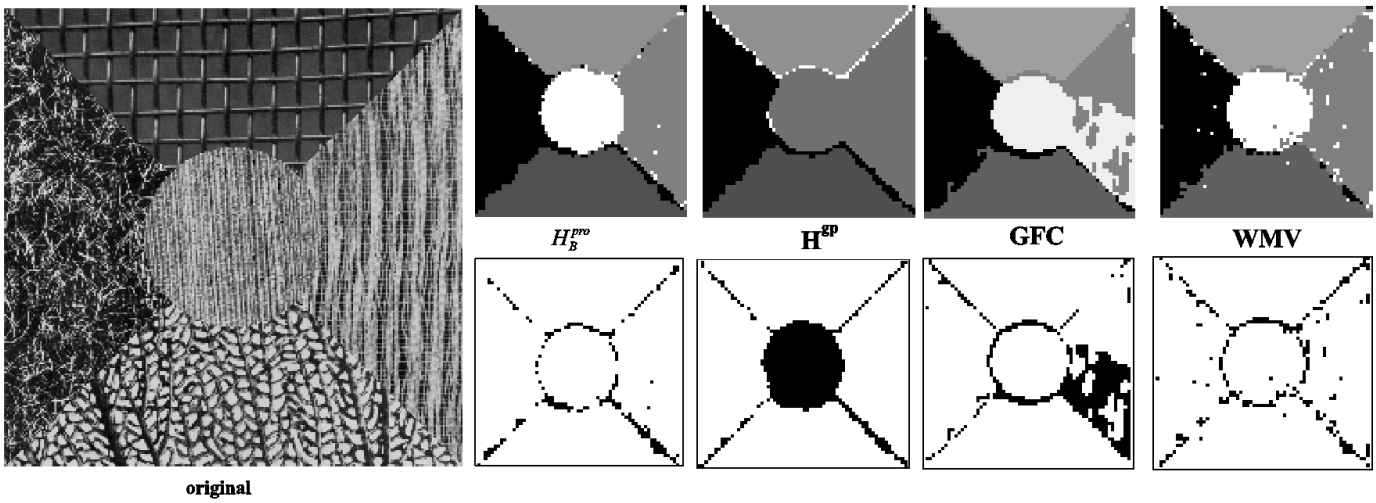


Fig. 1. Typical segmentation results using five clusters for different cost functions. Misclassified sites are depicted in black.

$$\tilde{g}_{i\nu} = \mathcal{H}\left(s_i\left(\mathbf{M}, \vec{e}_\nu\right)\right) - \mathcal{H}\left(s_i\left(\mathbf{M}, \vec{0}\right)\right).$$

Assuming symmetrical dissimilarities for simplicity this implementation yields

$$\begin{aligned}\tilde{g}_{i\nu} &= \sum_{j\in\mathcal{N}_i}\left(\frac{1}{n_{i\nu}} + \frac{1}{n^+_{j\nu}}\right)M_{j\nu}D_{ij} + \\ &\quad \sum_{j\neq i}\left(\frac{1}{n^+_{j\nu}} - \frac{1}{n^-_{j\nu}}\right)\sum_{k\in\mathcal{N}_j, k\neq i} M_{j\nu}M_{k\nu}D_{jk} \\ n^+_{j\nu} &= n_{j\nu}\left(s_i\left(\mathbf{M}, \vec{e}_\nu\right)\right), \quad n^-_{j\nu} = n_{j\nu}\left(s_i\left(\mathbf{M}, 0\right)\right) \qquad (14)\end{aligned}$$

for the general class of objective functions in (4). We have used the quantities P_ν, $p_{i\nu}$, Q_ν, and the sums

$$s_{i\nu} = \sum_{j\in\mathcal{N}_i} M_{j\nu}D_{ij}, \; S_\nu = \sum_{i=1}^{N} M_{i\nu}s_{i\nu}$$

as bookkeeping quantities to achieve a fast evaluation of $\tilde{g}_{i\nu}$ after single-site changes. The remaining technical difficulty in calculating the mean-field equations are the averages of the normalization constants, especially their inversely proportional dependency on functions of sums of assignment variables. Although every Boolean function has a polynomial normal form, which would in principle eliminate the involved denominator, some approximations have to be made to avoid exponential order in the number of conjunctions. Independently averaging the enumerator and the normalization in the denominator in (14),

$$h_{i\nu}(Q) = \left\langle g_{i\nu}(\mathbf{M})\right\rangle_Q \approx g_{i\nu}\left(\left\langle \mathbf{M}\right\rangle_Q\right)$$

achieves this approximation which is exact in the limit of $T \to 0$ for any N and in the thermodynamic limit of $N \to \infty$ for arbitrary T. General bounds as well as higher-order corrections of the approximation error can be obtained by a Taylor expansion around $\langle \mathrm{M}\rangle_Q$. The technical details can be found in [30].

INITIALIZE all $q_{i\nu} = 1/K$, temperature $T \leftarrow T_0$;
WHILE $T > T_{\mathrm{FINAL}}$
 add a small random perturbation to all $q_{i\nu}$
 REPEAT
  generate a permutation $\pi \in S_N$
  FOR i=1,...,N
   update $h_{\pi_{(i)}\nu}$ for all ν according to (13a)
   update $q_{\pi_{(i)}\nu}$ for all ν according to (13b)
 UNTIL converged
 $T \leftarrow \eta \cdot T$, $0 < \eta < 1$
END

Algorithm 1. Algorithm MFA.

5 Results

Several questions are empirically investigated in the following subsections. Section 5.1 addresses data extraction and modeling issues, including performance studies for a large database of textured images with known ground truth. In Section 5.2, we have benchmarked the proposed deterministic annealing optimization method against the ICM-algorithm and the Gibbs sampler. Section 5.3 shows results on representative examples of real-world images.

5.1 Texture Segmentation by Sparse Clustering

To empirically test the segmentation algorithms on a wide range of textures, we selected a representative set of 86 micropatterns from the Brodatz texture album [41].[5] A database of random mixtures (512×512 pixels each) containing 100 entities of five textures each (as depicted in Fig. 1a and Fig. 2a) was constructed from this collection of Brodatz textures.

All segmentations are based on a filter bank of 12 Gabor filters at four orientations and three scales. The χ^2-distance is applied to each channel independently, and a

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 micropattern properties, i.e., all textures are excluded where the texture property is lost when considering small image parts.

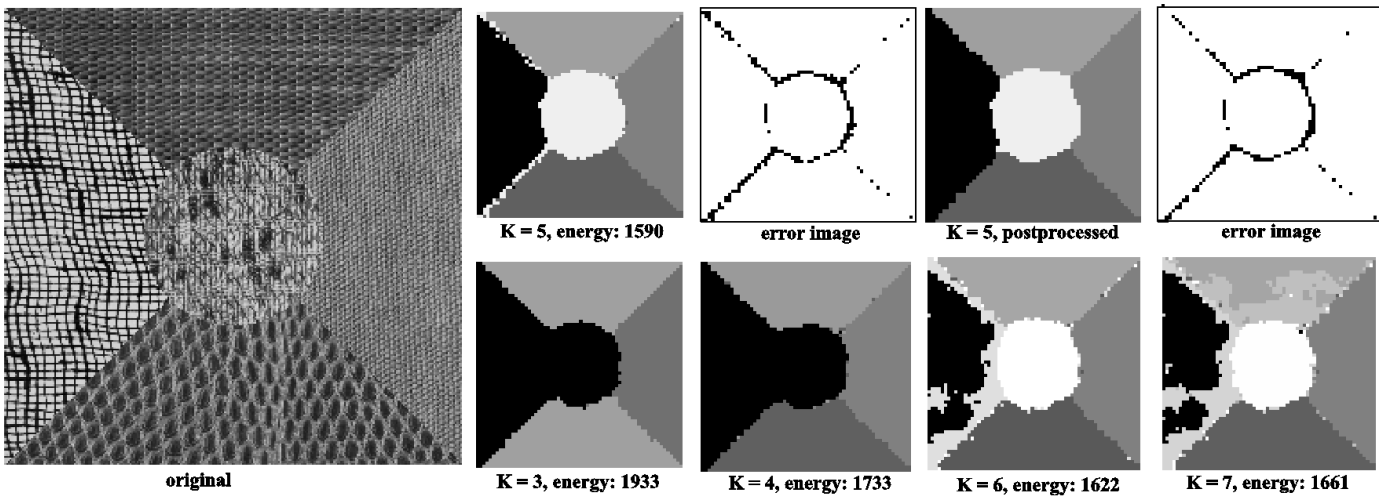


Fig. 2. Typical segmentation results on the basis of the normalized costs $\mathcal{H}_B^{\text{pro}}$ using three to seven clusters and complexity costs with $\lambda_c = 100/N$.

Minkowski norm with $p = 1$ is used to integrate the different channels. Sparse neighborhoods include the four-nearest neighbors and (on average) 80 randomly selected neighbors. For each image, a subset of 64×64 sites is considered using a square window of size 8×8 per site at the smallest scale. The size is increased by a factor of two for each filter octave. Typical segmentation examples are shown in Fig. 1 and Fig. 2. Additional examples are available via the World Wide Web.

The question which is empirically investigated in this section addresses the problem of how adequate texture segmentation is modeled by the extracted data matrices and the presented cost functions. An answer is given by comparing minimal cost configurations with known ground truth. As an independent reference algorithm, we have reimplemented the method of Jain and Farrokhnia [1], which clusters feature vectors extracted from Gabor filter responses (Gabor Feature Clustering, GFC). This method uses the absolute value of the hyperbolic tangens of the real part of the Gabor filters, which are further smoothed by a Gaussian filter. The texture segmentation problem is then formulated as a clustering problem of the resulting normalized feature vectors according to a *K*-means clustering criterion. We have chosen a deterministic annealing algorithm for clustering of vectorial data due to Rose et al. [11], which was empirically found to yield slightly better results than the *K*-means algorithm proposed in [1]. Both algorithms optimize the same cost function, $\mathcal{H}^{\text{km}}(\mathbf{M}, \mathbf{D})$, but the deterministic annealing implementation converges to superior minima. In order to obtain comparable results, we used the same 12 Gabor filters and extracted feature vectors on the same 64×64 regular sublattice of sites. To compare the χ^2-statistic to other state-of-the-art dissimilarity measures based on a vector space representation, we have implemented the parametric Weighted Mean Variance (WMV) measure proposed in [42]. For empirical means μ_i^r, μ_j^r and standard deviations σ_i^r, σ_j^r, the distance is defined by

$$D_{ij}^{(r)} = \frac{\left|\mu_i^r - \mu_j^r\right|}{\left|\sigma\left(\mu^r\right)\right|} + \frac{\left|\sigma_i^r - \sigma_j^r\right|}{\left|\sigma\left(\sigma^r\right)\right|},$$

where $\sigma(\cdot)$ denotes the standard deviation of the respective entity. As reported in [42], this measure based on a Gabor filter image representation outperforms several other parametric models and was chosen because of its competitiveness.

Table 1 summarizes the obtained mean and median values for all cost functions under consideration evaluated on the database of mixture images with five textures each. Fig. 3 shows histograms of misclassified sites for the different cost functions, for the WMV-measure and for the GFC algorithm. For $\mathcal{H}_B^{\text{pro}}$, a mean segmentation error rate as low as 6.0 percent was obtained. The best median error rate of 3.6 percent was obtained by $\mathcal{H}_A^{\text{pro}}$. Both cost functions yield very similar results as expected. With the exception of a few outliers, the distributions are sharply peaked around 3.5 percent. We recommend the use of $\mathcal{H}_B^{\text{pro}}$, since it can be implemented more efficiently. For $\mathcal{H}_B^{\text{con}}$, both mean and median error are larger, although there were a number of examples, where $\mathcal{H}_B^{\text{con}}$ performed slightly better than

TABLE 1
MEAN (MEDIAN) ERROR COMPARED TO GROUND TRUTH FOR SEGMENTING 100 RANDOMLY GENERATED IMAGES WITH $K = 5$ TEXTURES EACH

	$\mathcal{H}_A^{\text{pro}}$	$\mathcal{H}_B^{\text{pro}}$	$\mathcal{H}_B^{\text{con}}$	$\mathcal{H}^{\text{gp}}$	GFC
ICM	9.4 (4.6)	11.9 (6.7)	11.0 (6.1)	18.6 (18.0)	10.8 (6.7)
MFA	6.3 (3.6)	6.0 (3.7)	7.8 (4.8)	7.9 (4.0)	

The first and second rows show the results for the ICM algorithm and for Mean-Field Annealing (MFA), respectively. The columns correspond to different cost functions $\mathcal{H}$. For $\mathcal{H}_B^{\text{con}}$ a prior with $\lambda_s = \frac{150}{N} \cdot \mathbf{E}\left[D_{ij}\right]$ was used, while the data were shifted by $\triangle D = 0.1 - \mathbf{E}[D_{ij}]$ for $\mathcal{H}^{\text{gp}}$.

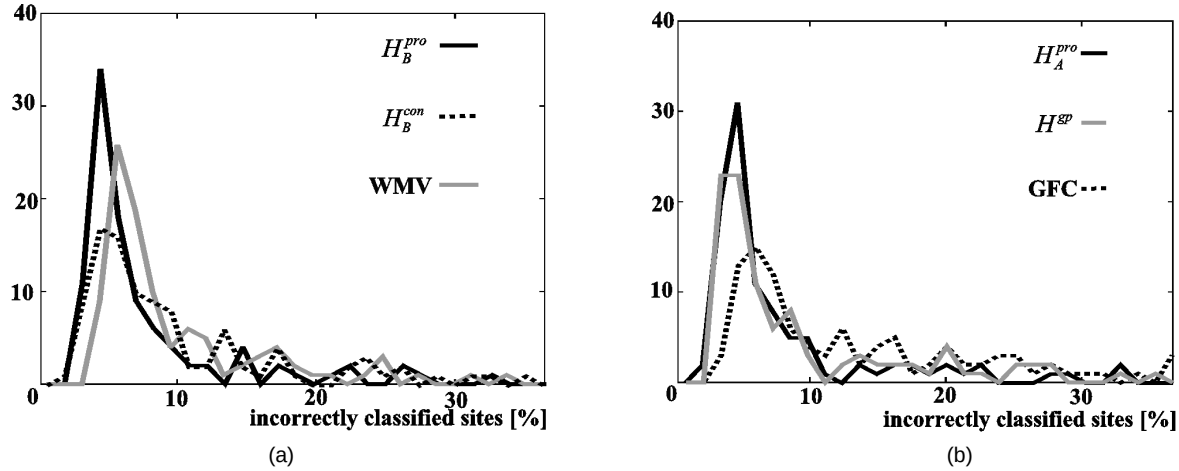


Fig. 3. Empirical density of the percentage of misclassified sites for the database with five textures each before postprocessing: (a) Normalized cost function $\mathcal{H}_B^{\text{pro}}$, the graph partitioning cost function $\mathcal{H}^{\text{gp}}$ and the GFC algorithm. (b) Normalized cost function $\mathcal{H}_A^{\text{pro}}$, the normalized cost function $\mathcal{H}_B^{\text{con}}$ and the normalized cost function $\mathcal{H}_B^{\text{pro}}$ using the WMV-distance measure instead of χ^2. Similar results are obtained after postprocessing.

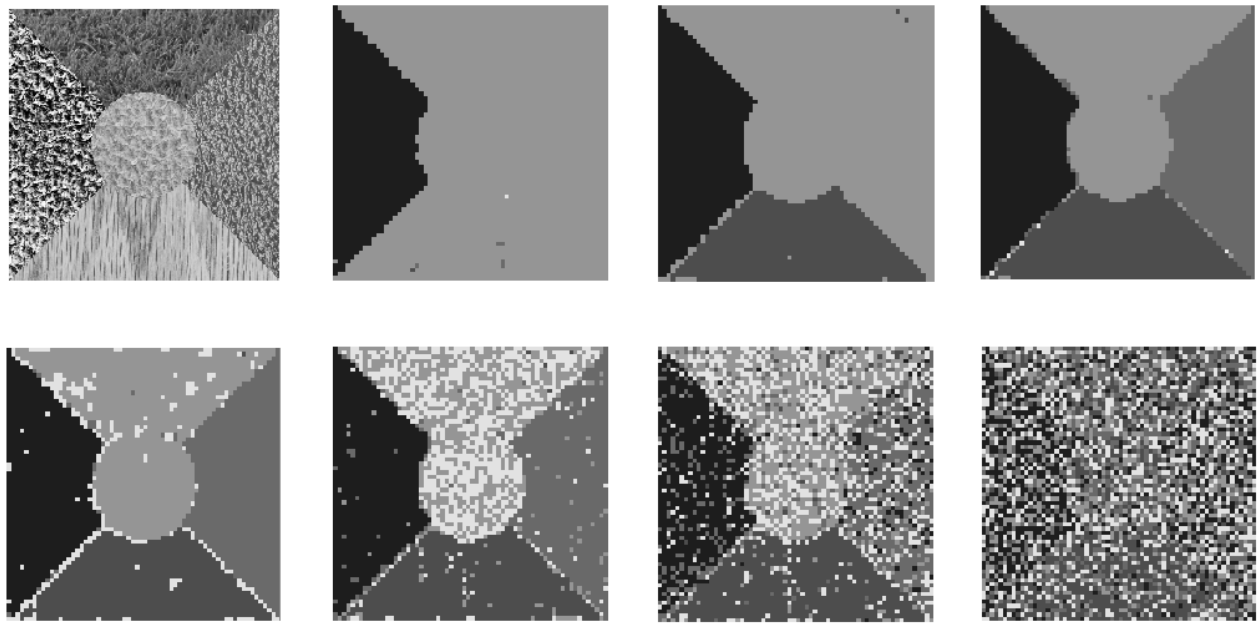


Fig. 4. Segmentations obtained by $\mathcal{H}^{\text{gp}}$ for several data shifts. In rows from the upper left: The original image and segmentations with a mean dissimilarity of –0.05, 0, 0.05, 0.1, 0.15, 0.2, and 0.25 are depicted. Segments collapse for negative shifts. For large positive shifts, the obtained segmentations become random, because the sampling noise induced by the random neighborhood system dominates the data contributions.

$\mathcal{H}_B^{\text{pro}}$. The structural deficiencies can be compensated by choosing an appropriate prior, but at the cost of empirically fixing an additional data-dependent parameter. Similar results were obtained by $\mathcal{H}_A^{\text{con}}$, the results of which are not explicitly reported here. We conclude that the invariant cost functions based on a pairwise data clustering formalization capture the true structure of the image in most cases. Furthermore, a weighting of cluster homogeneities proportional to the cluster size as in $\mathcal{H}^{\text{pro}}$ has proven to be advantageous. As can be seen in Fig. 2, the misclassified sites mainly correspond to errors at texture borders, which contain more than one texture. Misclassifications at the boundary are unavoidable due to the support of Gabor filters [43], as statistics from different textures are mixed. The postprocessing step improves the segmentations by a significant noise reduction.

To compare the quality of the different cost functions in detail, the distribution of the differences

$$\left(\mathcal{H}_B^{\text{pro}}\right) - \text{per}\left(\mathcal{H}^{other}\right)$$

are depicted in Fig. 5, where per(·) denotes the percentage of misclassified sites. Positive differences correspond to a better performance of $\mathcal{H}^{other}$. The value at zero corresponds to the percentage of instances, where $\mathcal{H}_B^{\text{pro}}$ performed better. It can be seen that the unnormalized cost function $\mathcal{H}^{\text{gp}}$ severely suffers from the missing shift-

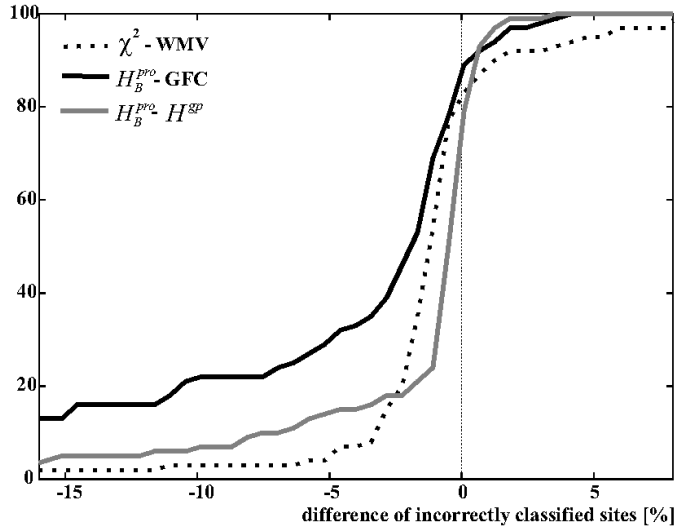


Fig. 5. The empirical density of the difference in segmentation quality for $\text{per}(\chi^2) - \text{per}(WMV)$, for $\text{per}(\mathcal{H}_B^{\text{pro}}) - \text{per}(GFC)$, and for $\text{per}(\mathcal{H}_B^{\text{pro}}) - \text{per}(\mathcal{H}^{\text{gp}})$ evaluated over 100 images containing five clusters each. Deterministic annealing algorithms with postprocessing were used.

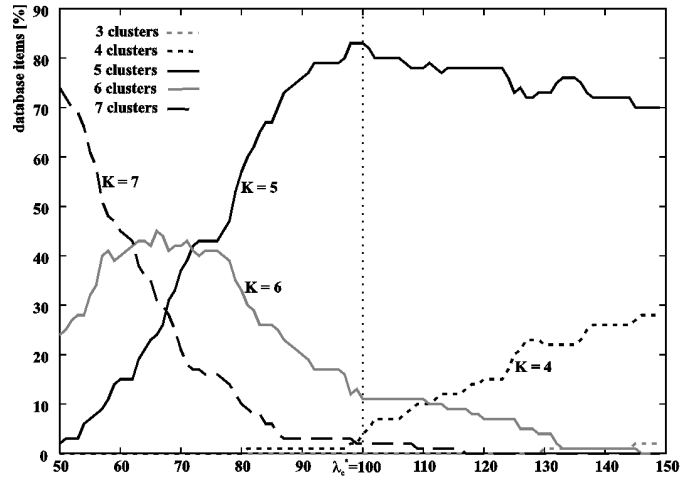


Fig. 6. The dependency of the chosen number K of segments depending on the prior weight λ_c for 100 images, where $\lambda_c N$ is depicted on the x-axis. We empirically found $\lambda_c^* = 100 / N$ as an optimal value.

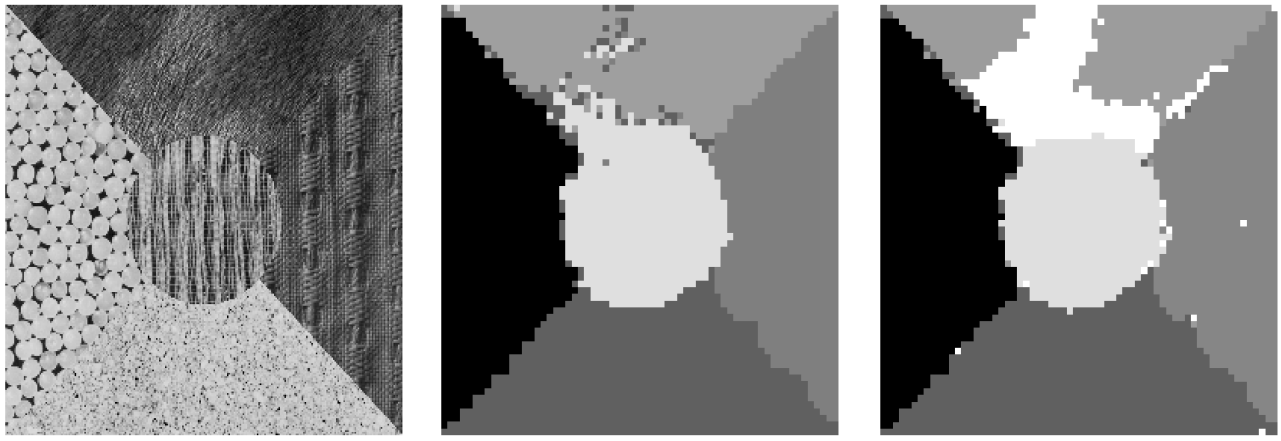
invariance property. This is further illustrated by Fig. 4. Depending on the shift, the unnormalized cost function often completely misses several texture classes (Fig. 1). As seen in Fig. 4, there may not even exist a parameter value to find all five textures. Even worse, the optimal value depends on the data at hand and varies for different images. With $\mathcal{H}^{\text{gp}}$, we achieved a mean error rate of 7.9 percent after extensive tuning to find the appropriate data shift. The results are worse than $\mathcal{H}_B^{\text{pro}}$, although the five textures are approximately of the same size. A further deterioration on images with largely varying texture sizes was observed. Fig. 3 contains the statistic of $\mathcal{H}^{\text{gp}}$ on the database with five textures. Note that a larger number of rather poor segmentations were obtained. We thus conclude that shift invariance is an important property to avoid parameter fine tuning of sensitive parameters and that the increased computational complexity for additional normalizations in $\mathcal{H}^{\text{pro}}$ is well-spent.

To empirically evaluate the performance of the dissimilarity measures, we compare in Fig. 5 the segmentation quality achieved by the χ^2-statistic with the parametric WMV measure. In the majority of examples better segmentations are obtained by χ^2. The correct structure is found by WMV in most cases, but the obtained segmentations are less accurate as illustrated by the example in Fig. 1. As can be seen from the histogram in Fig. 3, severe outliers are produced more frequently. We conclude that a nonparametric approach for similarity extraction is more powerful than parametric feature-based methods, since they are largely independent of the underlying feature distribution.

In Fig. 5, $\mathcal{H}_B^{\text{pro}}$ is benchmarked against the GFC-algorithm which is clearly outperformed. GFC yields a significant amount of structurally incorrect segmentations as can be seen from the histogram in Fig. 3. Typically, as in the presented example in Fig. 1, large image parts are missclassified, resulting in a mean error rate of 10.8 percent, which was worse than the results obtained by mean-field annealing for all tested proximity-based methods. Thus, a significant gain in segmentation quality is observed by using proximity data based on statistical tests instead of feature vectors.

Another important question concerns the choice of the number of clusters K. As illustrated by the example in Fig. 2, the energy value of the final configuration is a rather good indicator for the correct number of clusters. The rapid decrease for underestimated cluster numbers nearly stops after reaching the correct number. The final cluster number can be determined by adding a penalty term proportional to the number of clusters. In Fig. 6, the obtained number of clusters depending on the weighting factor λ_c is depicted.[6] The few errors made are visually plausible, as illustrated by the example in Fig. 7, where an inhomogeneous texture is segmented into two homogeneous parts in a satisfying way. Note that a specific choice of λ_c merely selects a certain *segmentation resolution* by weighting complexity costs against the data term. The results demonstrate that the exact value of λ is not critical, as for a large range of values pleasing solutions are found for a large set of images. In our opinion, there does not exist a unique *true* number of clusters in unsupervised texture segmentation (at least not for natural images). For example, it is impossible to decide whether a segmentation into four or eight regions is better for the aerial image of San Francisco in Fig. 10. We believe a hierarchical clustering model to be more appropriate for many purposes and have extended our work in that direction in more recent publications [44].

6. For the chosen value of $\lambda_c^* = 100 / N$, we obtained five clusters in 83, four clusters in four, six clusters in 11, and seven clusters in two cases. The results are similar for a broad range of values for λ_c.



(a) (b) (c)

Fig. 7. In this example, the proposed algorithm selects $K = 6$ instead of $K = 5$ as the optimal number of clusters. (a) Original. (b) K = 5, energy: 1,568. (c) K = 6, energy: 1,536.

5.2 Mean-Field Approximation and Gibbs Sampling

Another important question is concerned with the quality of deterministic annealing algorithms compared to stochastic procedures. The quality of the proposed clustering algorithm was evaluated by comparing the costs of the achieved segmentation with the deterministic, greedy ICM algorithm and with a stochastic Gibbs sampling method.

The distribution of the differences of costs were chosen for a graphical representation. Exemplarily for the normalized cost functions the cost differences for $\mathcal{H}_B^{\text{pro}}$ using deterministic annealing versus ICM and deterministic annealing versus the Gibbs sampler are depicted in Fig. 8. A substantial improvement over the ICM algorithm can be reported, since ICM gets frequently stuck in poor local minima, as one might expect from a greedy technique. The comparison with the Gibbs sampler is more difficult, because the Gibbs sampler can be improved by slow cooling rates. We decided to use a comparable running time for both, deterministic annealing and Gibbs sampler, in our implementation[7] with a conservative annealing schedule.

7. About four minutes on a Sun UltraSparc.

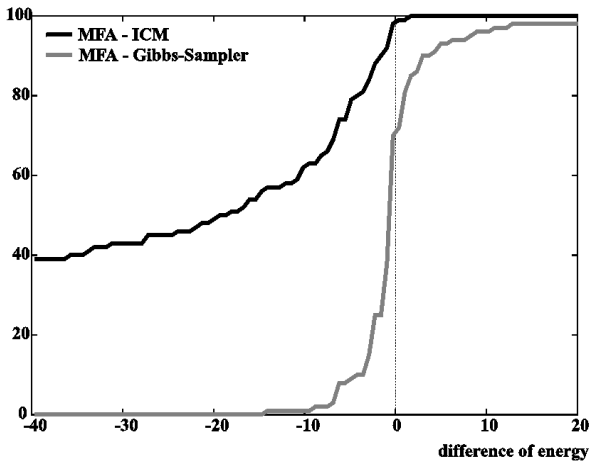


Fig. 8. The empirical density of the cost difference of mean-field annealing (MFA) versus the greedy ICM algorithm and versus the Gibbs sampler evaluated over 100 images containing five textures each.

The ICM algorithm runs notably faster than the other algorithms. Deterministic annealing and Gibbs sampling yield similar results with slight advantages for deterministic annealing; but in all cases, where one of the algorithms yields superior solutions, the improvement is marginal. This detailed analysis shows that deterministic annealing yields optimal or near optimal solutions in most experiments. Furthermore, we advocate deterministic annealing algorithm as a good choice for an efficient computation of near-optimal solutions, especially since the loss in segmentation quality caused by fast annealing schedules is substantially lower for deterministic annealing than for Gibbs sampling. In a followup study, both deterministic annealing and ICM have been substantially accelerated using the concept of *multiscale optimization*. The resulting optimization times are in the range of less than five seconds[8] for the examples shown here without loss in performance quality [45].

8. For deterministic annealing on a Sun UltraSparc, less than two seconds for ICM. The multiscale annealing scheme is not directly applicable for stochastic Gibbs sampling.

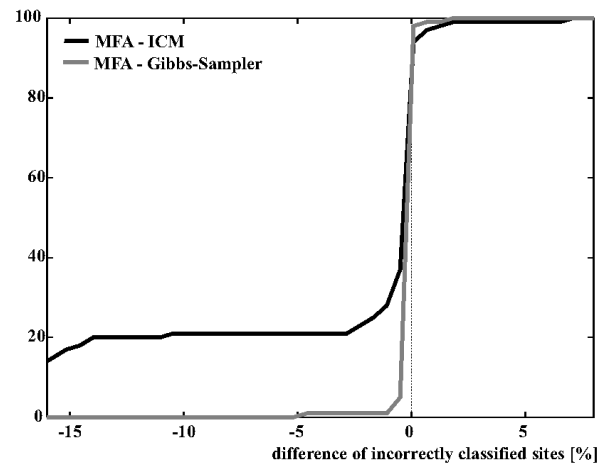


Fig. 9. The empirical density of the misclassification error difference between mean-field annealing and either the ICM algorithm or the Gibbs sampler. Differences are evaluated on 100 images containing five textures each.

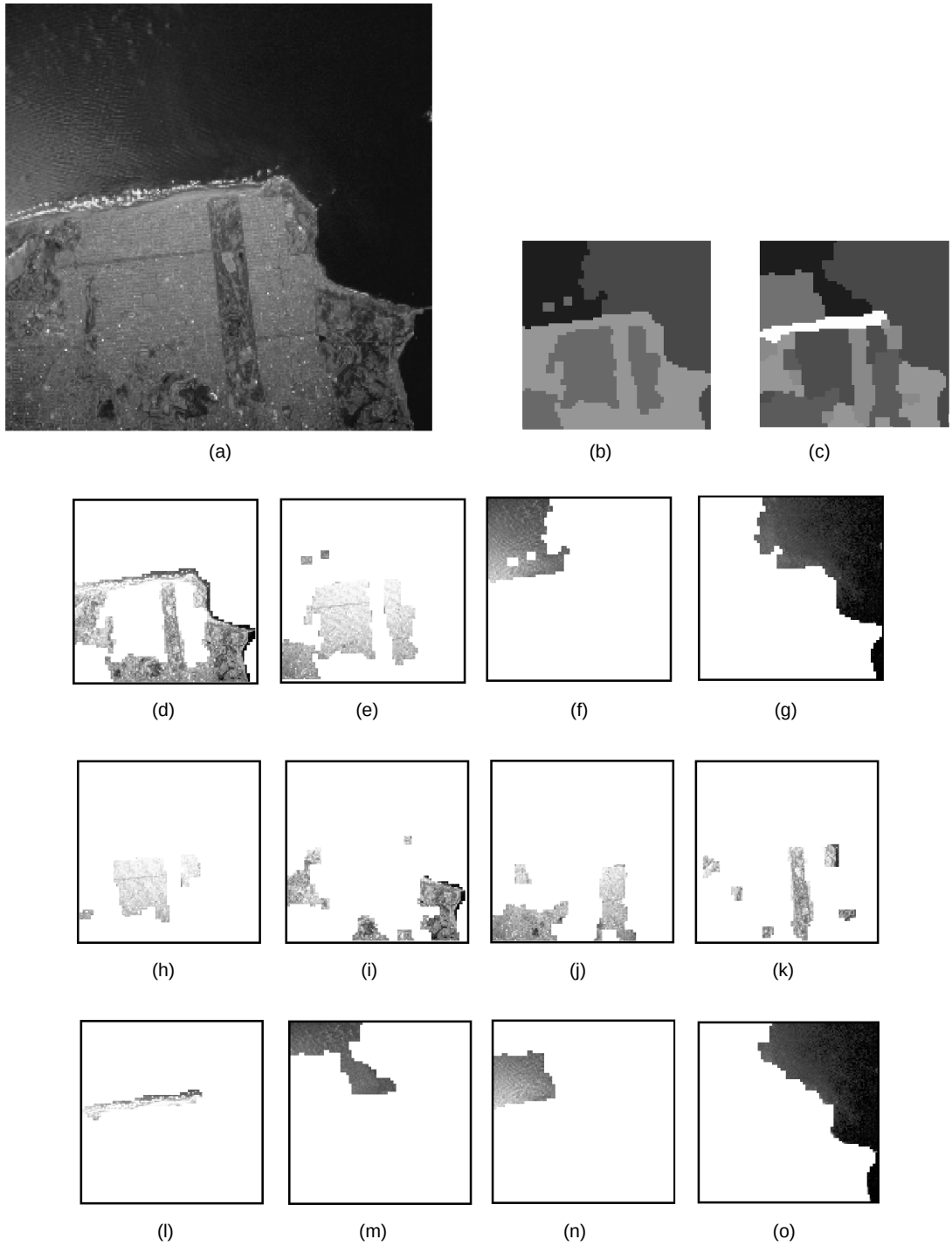
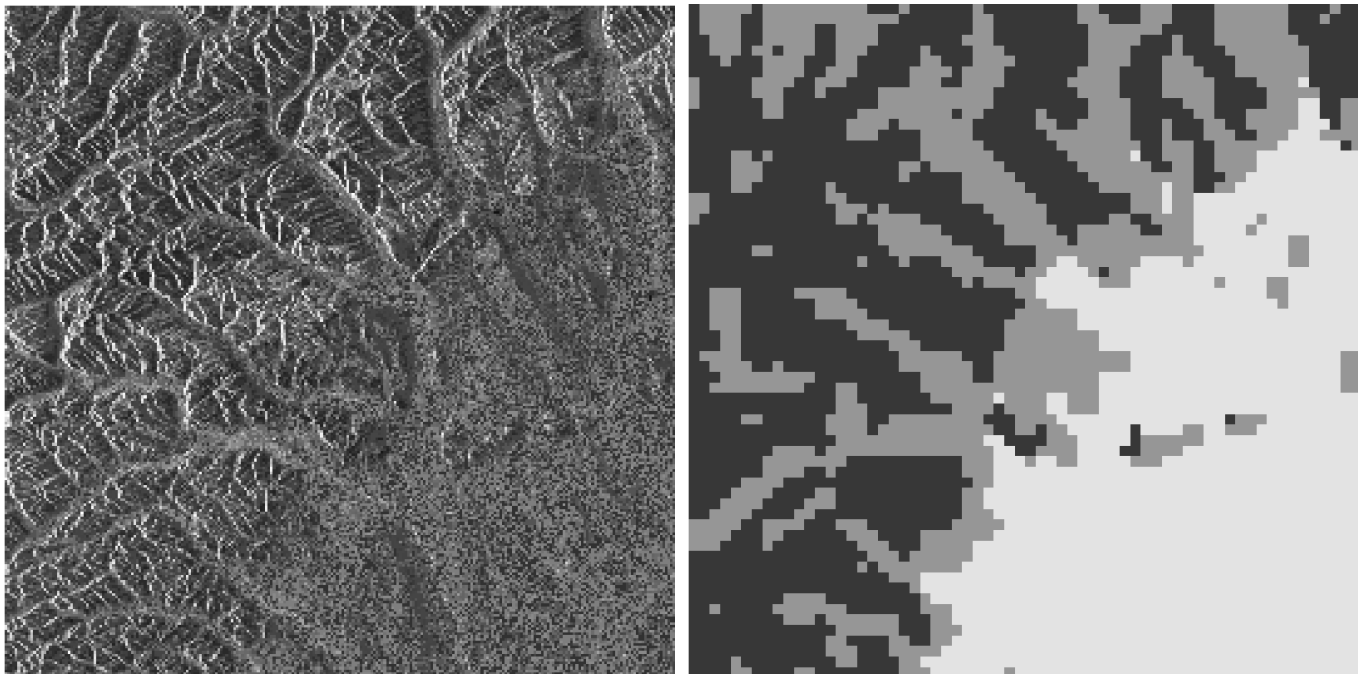


Fig. 10. Segmentation of an aerial image of San Francisco: (a) Original gray-scale image. (b) Segmentation into four clusters. (c) Segmentation into eight clusters. (d)-(g) Visualization of the four-cluster segmentation. (h)-(o) Visualization of the eight-cluster segmentation. The segmentations are obtained for a resolution of 64×64 sites, a topological prior with $\lambda_t = 0.01$ added to the cost function $\mathcal{H}_B^{\text{pro}}$ and the deterministic annealing optimization procedure.

In Fig. 9, deterministic annealing is compared with the ICM algorithm and the Gibbs sampler, but now with respect to the percentage of misclassifications instead of energy. The results are very similar to Fig. 8, thus the better optimization procedure leads to substantial improvements in the segmentation quality. This result, although mandatory for optimization approaches in computer vision, is by no means obvious, since the global optimum of the cost function does not necessarily coincide with the ground truth segmentation. Indeed, the ground truth segmentation has higher energy than the minima found in most cases. This deficit is reasonably explained by the fact that border areas are often incorrectly modeled by the cost functions.



(a) (b)

Fig. 11. Segmentation of an SAR image of a mountain landscape. (a) Original image. (b) Segmentation into three clusters. The segmentation was obtained using a resolution of 64×64 sites, a topological prior with $\lambda_t = 0.01$ added to the cost function $\mathcal{H}^{\text{nc}}$, and the deterministic annealing optimization procedure.

5.3 Real-World Images

The presented algorithms are applicable to real-world images without any restrictions, and we demonstrate their robustness on three types of images: aerial images, SAR images, and indoor scenes.

The segmentation of aerial images is an important application, as many aerial images contain texture-like structures. Fig. 10 shows two segmentations of an aerial image of San Francisco as an example. Both segmentations are visually satisfying, as is evident from the single-cluster representations in Fig. 10d to Fig. 10o. Furthermore, up to small errors such as the classification of water as tilled area in the four-cluster segmentation, the solution obtained is semantically correct, e.g., tilled area and parks as well as water are well-discriminated in both segmentations.[9]

A second important class of textured images are Synthetic Aperture Radar (SAR) images. The dramatically increasing quantity of available SAR imagery requires unsupervised processing, e.g., to automatically detect environmental changes. In Fig. 11, the segmentation into three texture classes of an SAR image is depicted. The achieved segmentation is both visually and semantically correct, since mountains, valleys, and the plain are well-separated. Even small valley structures are detected. Note that the segmentation was obtained by introducing a topological prior which renders an additional postprocessing step superfluous.

A third class of applications for texture segmentation are indoor and outdoor images, which contain textured objects. Unsupervised segmentation can be beneficially applied in autonomous robotics, and the presented algorithms have been implemented on the autonomous robot Rhino [46]. An example image of a typical office environment is presented in Fig. 12. The achieved segmentation is both visually and semantically satisfying. Untextured parts of the image are grouped together irrespective of their absolute luminance value, and the discrimination of the remaining three textures is plausible.

6 Conclusion

We have presented a novel approach to segment textured images on the basis of four, mutually independent building blocks. First, a scale space approach for data representation based on Gabor filters has been suggested, which evolves naturally from theoretical concepts and exhibits excellent discrimination properties for a wide range of natural textures.

Second, we have suggested to use nonparametric statistical tests for texture comparison, and we have investigated the discriminative power of these tests. There is no need to adjust any parameters or thresholds to obtain the proximity data for the data clustering stage apart from general system parameters like bin sizes for histograms or filter size and filter orientation.

Third, unsupervised texture segmentation was formulated as a pairwise data clustering problem based on dissimilarities between texture blocks with a sparse neighborhood structure. Four new cost functions have been derived

9. To our surprise, it turned out that segments g and o are known as the "potato patch," a part of the open ocean with very rough water and strong currents, which explains the different texture.

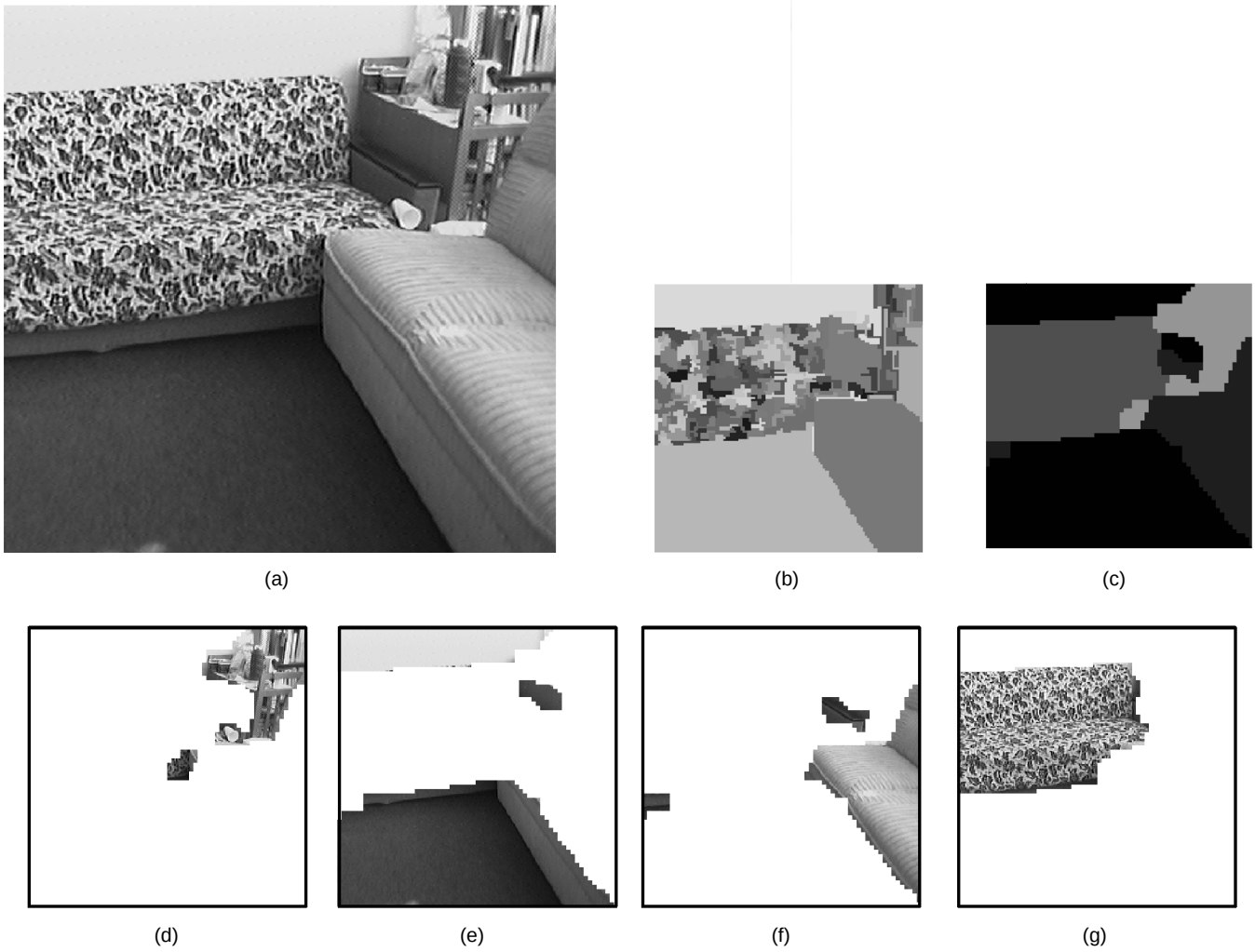


Fig. 12. (a) An indoor image of a typical office environment containing an old-style sofa. (b) A contrast-based image segmentation with a region-merging algorithm. (c) A texture segmentation with $K = 4$. The image partitioning is visualized in (d)-(g).

from the principles of scale- and shift-invariance. These objective functions as well as the unnormalized graph partitioning cost function proposed in [22] have been empirically compared on a large dataset of textured images to evaluate their advantages and disadvantages. The new objective functions have been demonstrated to be substantially superior compared to the unnormalized objective function and the GFC algorithm. The new cost function $\mathcal{H}_B^{\text{pro}}$ can be interpreted as a natural generalization of the K–means criterion. $\mathcal{H}_B^{\text{pro}}$ possesses the desired invariance properties and the necessary robustness, and it demonstrates excellent segmentation quality combined with computational efficiency.

Fourth, we have developed a general mathematical framework for applying the optimization principle of deterministic annealing to arbitrary partitioning problems. The framework has been developed from a purely algorithmic perspective to construct efficient continuation methods with convergence proofs. The mean-field equations as well as an efficient implementation of Gibbs sampling for the proposed objective functions have been presented. The deterministic annealing algorithm has been benchmarked against the ICM algorithm and the Gibbs sampler, yielding clearly superior results.

The segmentation algorithms have been tested and validated on a large database of Brodatz-like microtexture mixtures and on a representative set of textured real-world images. Note that in all simulations which covered a wide range of image domains, the same set of parameters were used to obtain satisfactory results. We, therefore, conclude that our approach constitutes a truly *unsupervised* method for texture segmentation.

APPENDIX

PROOF OF PROPOSITION 1. We have to show, that

$$\sum_{\nu=1}^{K}\sum_{i=1}^{N}\frac{\sum_{j\in\mathcal{N}_i} M_{i\nu}M_{j\nu}\Delta D}{n_{i\nu}\left(p_{i\nu},P_\nu,Q_\nu\right)} = \text{const}$$

for arbitrary $\mathbf{M} \in \mathcal{M}$, where $\triangle D \in \mathbb{R}$ is a global data shift. This is obviously equivalent to

$$\sum_{\nu=1}^{K}\sum_{i=1}^{N} M_{i\nu} r_{i\nu} = \text{const},$$

with $r_{i\nu} = \frac{p_{i\nu}}{n_{i\nu}}$. To proceed, we state two Lemmas which can be proven by elementary mathematics. □

LEMMA 1. *If for an arbitrary, but fixed graph $\mathcal{G}$ and a function*

$$f : \mathbb{R}_+^2 \to \mathbb{R}_+,$$
$$\forall \mathbf{M} \in \mathcal{M} : \sum_{\nu=1}^{K} f(P_\nu, Q_\nu) = const$$

then $f(P_\nu, Q_\nu) = aP_\nu + c$ for some $a, c \in \mathbb{R}$.

LEMMA 2. *If for an arbitrary, but fixed graph $\mathcal{G}$ and a function*

$$\hat{f} : \mathbb{R}_+^3 \to \mathbb{R}_+,$$
$$\forall \mathbf{M} \in \mathcal{M} : \sum_{i=1}^{N} M_{i\nu} \hat{f}(p_{i\nu}, P_\nu, Q_\nu) = const$$

then $\hat{f}(p_{i\nu}, P_\nu, Q_\nu) = aP_\nu^{-1} + bp_{i\nu}Q_\nu^{-1}$ for some $a, b \in \mathbb{R}$.

Applying Lemma 1, it suffices to show

$$\sum_{i=1}^{N} M_{i\nu} r_{i\nu} = 1$$

or

$$\sum_{i=1}^{N} M_{i\nu} r_{i\nu} = P_\nu,$$

since all other solutions are linear combinations of these two elementary solutions. Applying Lemma 2 to the first case results in elementary solutions $n_{i\nu} = p_{i\nu}P_\nu$ and $n_{i\nu} = Q_\nu$. In the second case, we simply multiply the equation in Lemma 2 by P_ν and obtain $n_{i\nu} = p_{i\nu}$ and $n_{i\nu} = Q_\nu / P_\nu$.

PROOF OF THEOREM 1. Introducing Lagrange parameters λ_i to enforce the normalization $\sum_{\nu=1}^{K} q_{i\nu} = 1$ and taking derivatives of the Lagrangian of the generalized free energy results in

$$\begin{aligned}
&\frac{\partial}{\partial q_{i\nu}}\left[\langle \mathcal{H} \rangle_Q - TS(Q) + \sum_{k=1}^{N}\lambda_k \sum_{\mu=1}^{K} q_{k\mu}\right] \\
&= \frac{\partial \langle \mathcal{H} \rangle_Q}{\partial q_{i\nu}} + T\frac{\partial}{\partial q_{i\nu}}\sum_{k=1}^{N}\sum_{\mu=1}^{K} q_{k\mu}\log q_{k\mu} + \lambda_i \\
&= \frac{\partial \langle \mathcal{H} \rangle_Q}{\partial q_{i\nu}} + T\log q_{i\nu} + \lambda_i + 1 \qquad (15)
\end{aligned}$$

Setting equal to zero proves the first part of the theorem. Performing the derivatives gives

$$\begin{aligned}
\frac{\partial \langle \mathcal{H} \rangle_Q}{\partial q_{i\nu}} &= \sum_{\mathbf{M}\in\mathcal{M}} \mathcal{H}(\mathbf{M})\frac{\partial Q(\mathbf{M})}{\partial q_{i\nu}} \\
&= \sum_{\mathbf{M}\in\mathcal{M}} \frac{M_{i\nu}}{q_{i\nu}}\mathcal{H}(\mathbf{M})Q(\mathbf{M}) \qquad (16)
\end{aligned}$$

□

PROOF OF COROLLARY 1. By differentiating (15), the Hessian of the generalized free energy can be expressed as

$$\begin{aligned}
\frac{\partial^2}{\partial q_{i\nu}\partial q_{j\mu}}\mathcal{F}_T(Q^*) &= \frac{\partial^2 \langle \mathcal{H} \rangle}{\partial q_{i\nu}\partial q_{j\mu}} + \frac{T}{q_{i\nu}^*}\delta_{ij}\delta_{\nu\mu} \\
&= \frac{\langle M_{i\nu}M_{j\mu}\mathcal{H} \rangle}{q_{i\nu}^* q_{j\mu}^*} + \left[\frac{T}{q_{i\nu}^*} - \frac{\langle M_{i\nu}\mathcal{H} \rangle}{(q_{i\nu}^*)^2}\right]\delta_{ij}\delta_{\nu\mu} \\
&= \begin{cases} T / q_{i\nu}^* & \text{for } i = j \wedge \nu = \mu \\ 0 & \text{for } i = j \wedge \nu \neq \mu \\ \langle M_{i\nu}M_{j\mu}\mathcal{H} \rangle / (q_{i\nu}^* q_{j\mu}^*) & \text{otherwise} \end{cases}
\end{aligned}$$

which is positive definite in the subspace spanned by one site *i*. Thus, an asynchronous update step (13) minimizes $\mathcal{F}_T$ with respect to a single site subspace. As $\mathcal{F}_T$ is bounded from below for a fixed temperature by definition, this ensures convergence to a local minimum. □

PROOF OF COROLLARY 2. Rewriting (12), we arrive at

$$\begin{aligned}
h_{i\nu} &= \sum_{\mathbf{M}\in\mathcal{M}} \frac{M_{i\nu}}{q_{i\nu}}\mathcal{H}(\mathbf{M})Q(\mathbf{M}) \\
&= \sum_{\mathbf{M}\in\mathcal{M}} \mathcal{H}(\mathbf{s}_i(\mathbf{M}, \vec{e}_\nu))Q(\mathbf{M}) = \langle g_{i\nu} \rangle_Q
\end{aligned}$$

□

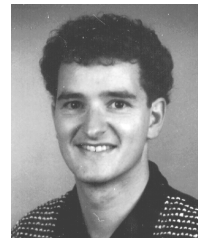
ACKNOWLEDGMENTS

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

REFERENCES

[1] A. Jain and F. Farrokhnia, "Unsupervised Texture Segmentation Using Gabor Filters," *Pattern Recognition*, vol. 24, no. 12, pp. 1,167–1,186, 1991.
[2] J. Mao and A. Jain, "Texture Classification and Segmentation Using Multiresolution Simultaneous Autoregressive Models," *Pattern Recognition*, vol. 25, pp. 173–188, 1992.
[3] D. Geman, S. Geman, C. Graffigne, and P. Dong, "Boundary Detection by Constrained Optimization," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 12, no. 7, pp. 609–628, July 1990.
[4] T. Ojala, M. Pietikäinen, and D. Harwood, "A Comparative Study of Texture Measures With Classification Based Feature Distributions," *Pattern Recognition*, vol. 29, no. 1, pp. 51–59, 1996.
[5] J. Shi and J. Malik, "Normalized Cuts and Image Segmentation," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 731–737, 1997.
[6] J. MacQueen, "Some Methods for Classification and Analysis of Multivariate Observations," *Proc. Fifth Berkeley Symp. Mathematical Statistics and Probability*, pp. 281–297, 1967.
[7] C. Peterson and B. Söderberg, "A New Method for Mapping Optimisation Problems Onto Neural Networks," *Int'l J. Neural Systems*, vol. 1, no. 1, pp. 3–22, 1989.
[8] D. van den Bout and T. Miller, "Graph Partitioning Using Annealed Neural Networks," *IEEE Trans. Neural Networks*, vol. 1, no. 2, pp. 192–203, 1990.
[9] A. Yuille, "Generalized Deformable Models, Statistical Physics, and Matching Problems," *Neural Computation*, vol. 2, pp. 1–24, 1990.
[10] S. Gold and A. Rangarajan, "A Graduated Assignment Algorithm for Graph Matching," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 18, no. 4, pp. 377–388, Apr. 1996.
[11] K. Rose, E. Gurewitz, and G. Fox, "Vector Quantization by Deterministic Annealing," *IEEE Trans. Information Theory*, vol. 38, no. 4, pp. 1,249–1,257, 1992.

[12] J. Buhmann and H. Kühnel, "Vector Quantization With Complexity Costs," *IEEE Trans. Information Theory*, vol. 39, pp. 1,133–1,145, 1993.
[13] D. Geiger and F. Girosi, "Parallel and Deterministic Algorithms From MRF's: Surface Reconstruction," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 13, pp. 401–412, 1991.
[14] G. Bilbro, W. Snyder, S. Garnier, and J. Gault, "Mean Field Annealing: A Formalism for Constructing GNC–Like Algorithms," *IEEE Trans. Neural Networks*, vol. 3, no. 1, 1992.
[15] J. Zhang, "The Mean Field Theory in EM Procedures for Blind Markov Random Fields," *IEEE Trans. Image Processing*, vol. 2, no. 1, pp. 27–40, 1993.
[16] J. Zerubia and R. Chellappa, "Mean Field Annealing Using Compound Gauss-Markov Random Fields for Edge Detection and Image Estimation," *IEEE Trans. Neural Networks*, vol. 4, no. 4, pp. 703–709, 1993.
[17] T. Hofmann and J. Buhmann, "Pairwise Data Clustering by Deterministic Annealing," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 19, no. 1, pp. 1–14, Jan. 1997.
[18] C. Peterson and J. Anderson, "A Mean Field Theory Learning Algorithm for Neural Networks," *Complex Systems*, vol. 1, pp. 995–1,019, 1987.
[19] D. Geiger and F. Girosi, "Coupled Markov Random Fields and Mean Field Theory," *Advances in Neural Information Processing Systems 2*, pp. 660–667, 1990.
[20] G. Bilbro and W. Snyder, "Mean Field Approximation Minimizes Relative Entropy," *J. Optical Soc. Am.*, vol. 8, no. 2, 1989.
[21] J. Zhang, "The Application of the Gibbs–Bogoliubov–Feynman Inequality in Mean-Field Calculations for Markov Random Fields," *IEEE Trans. Image Processing*, vol. 5, no. 7, pp. 1,208–1,214, 1996.
[22] S. Geman and D. Geman, "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 6, no. 6, pp. 721–741, Nov. 1984.
[23] H. Romeny, ed., *Geometry-Driven Diffusion in Computer Vision*. Kluwer Academic Publishers, 1994.
[24] L. Florack, B.t. Haar Romeny, J. Koenderink, and M. Viergever, "Families of Tuned Scale-Space Kernels," G. Sandini, ed., *Proc. Second European Conf. Computer Vision*, pp. 19–23. Berlin: Springer-Verlag, 1992.
[25] J. Daugman, "Uncertainty Relation for Resolution in Space, Spatial Frequency, and Orientation Optimized by Two-Dimensional Visual Cortical Filters," *J. Optical Soc. Am. A*, vol. 2, no. 7, pp. 1,160–1,169, 1985.
[26] A. Bovik, M. Clark, and W. Geisler, "Multichannel Texture Analysis Using Localized Spatial Filters," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 12, no. 1, pp. 55–73, Jan. 1990.
[27] I. Fogel and D. Sagi, "Gabor Filters as Texture Discriminators," *Biological Cybernetics*, vol. 61, pp. 103–113, 1989.
[28] B.W. Lindgren, *Statistical Theory*, 3rd ed. New York: Macmillan Publishing, 1976.
[29] T. Hofmann, J. Puzicha, and J. Buhmann, "Unsupervised Segmentation of Textured Images by Pairwise Data Clustering," *Proc. IEEE Int'l Conf. Image Processing*, pp. III:137–140, 1996.
[30] T. Hofmann, J. Puzicha, and J. Buhmann, "A Deterministic Annealing Framework for Unsupervised Texture Segmentation," Technical Report IAI-TR-96-2, Institut für Informatik III, 1996.
[31] J. Puzicha, T. Hofmann, and J. Buhmann, "Non-Parametric Similarity Measures for Unsupervised Texture Segmentation and Image Retrieval," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 267–272, 1997.
[32] T.-S. Lee, D. Mumford, and A. Yuille, "Texture Segmentation by Minimizing Vector-Valued Energy-Functionals: The Coupled Membrane Model," *Proc. Second European Conf. Computer Vision*, G. Sandini, ed., pp. 165–183, 1992.
[33] S. Kirkpatrick, C. Gelatt, and M. Vecchi, "Optimization by Simulated Annealing," *Science*, vol. 220, no. 4,598, pp. 671–680, 1983.
[34] V. Cerny, "Thermodynamical Approach to the Travelling Salesman Problem," *J. Optimization Theory and Applications*, vol. 45, pp. 41–51, 1985.
[35] E. Jaynes, "Information Theory and Statistical Mechanics," *Physical Rev.*, vol. 106, no. 4, pp. 620–630, 1957.
[36] J. Besag, "On the Statistical Analysis of Dirty Pictures," *J. Royal Statistical Soc.*, Series B, vol. 48, pp. 25–37, 1986.
[37] B. Hajek, "Cooling Schedules for Optimal Annealing," *Mathematics of Operation Research*, vol. 13, pp. 311–324, 1988.
[38] J. Puzicha, T. Hofmann, and J. Buhmann, "Deterministic Annealing: Fast Physical Heuristics for Real Time Optimization of Large Systems," *Proc. 15th IMACS World Congress on Scientific Computation, Modelling and Applied Mathematics*, 1997.
[39] J. Hopfield and D. Tank, "Neural Computation of Decisions in Optimisation Problems," *Biological Cybernetics*, vol. 52, pp. 141–152, 1985.
[40] J. Zhang, "The Convergence of Mean Field Procedures for MRF's," *IEEE Trans. Image Processing*, vol. 5, no. 12, pp. 1,662–1,665, 1991.
[41] P. Brodatz, *Textures: A Photographic Album for Artists and Designers*. New York: Dover Publications, 1966.
[42] B. Manjunath and W. Ma, "Texture Features for Browsing and Retrieval of Image Data," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 18, no. 8, pp. 837-842, Aug. 1996.
[43] R. Navarro, O. Nestares, J. Portilla, and A. Tabernero, "Several Experiments on Texture Analysis, Coding and Synthesis by Gabor Wavelets," Tech. Rep. 52, Instituto de Optica Daza de Valdes de Madrid, 1994.
[44] T. Hofmann, J. Puzicha, and J. Buhmann, "An Optimization Approach to Unsupervised Hierarchical Texture Segmentation," *Proc. IEEE Int'l Conf. Image Processing*, pp. 213–217, 1997.
[45] J. Puzicha and J. Buhmann, "Multiscale Annealing for Real-Time Unsupervised Texture Segmentation," Technical Report IAI–97–4, Institut für Informatik III (a short version appeared in *Proc. ICCV '98*, pp. 267–273), 1997.
[46] J. Buhmann, W. Burgard, A. Cremers, D. Fox, T. Hofmann, F. Schneider, I. Strikos, and S. Thrun, "The Mobile Robot RHINO," *AI Magazine*, vol. 16, no. 1, 1995.



Thomas Hofmann received the Diploma and PhD degrees in computer science from the University of Bonn in 1993 and 1997, respectively. His PhD 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.

Jan Puzicha received the Diploma degree in computer science from the University of Bonn, Germany, in 1995. In November 1993, he joined the Computer Vision and Pattern Recognition group at the University of Bonn, where he is currently completing his PhD thesis on optimization methods for grouping and segmentation. His research interests include image processing, remote sensing, autonomous robots, data analysis, and data mining.

Joachim M. Buhmann received a PhD 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.