

Deterministic Annealing for Unsupervised Texture Segmentation

Thomas Hofmann, Jan Puzicha and Joachim M. Buhmann *

Institut für Informatik III, Römerstraße 164
D-53117 Bonn, Germany
email: {th,jan,jb}@cs.uni-bonn.de, <http://www-dbv.cs.uni-bonn.de>

Abstract. In this paper a rigorous mathematical framework of *deterministic annealing* and *mean-field approximation* is presented for a general class of partitioning, clustering and segmentation problems. We describe the canonical way to derive efficient optimization heuristics, which have a broad range of possible applications in computer vision, pattern recognition and data analysis. In addition, we prove novel convergence results. As a major practical application we present a new approach to the problem of unsupervised texture segmentation which relies on statistical tests as a measure of homogeneity. More specifically, this results in a formulation of texture segmentation as a *pairwise data clustering* problem with a sparse neighborhood structure. We discuss and compare different clustering objective functions, which are systematically derived from invariance principles. The quality of the novel algorithms is empirically evaluated on a large database of Brodatz-like micro-texture mixtures and on a representative set of real-word images.

1 Introduction

The *unsupervised segmentation* of textured images is widely recognized as a difficult and challenging computer vision problem. It possesses a multitude of important applications, ranging from vision-guided autonomous robotics and remote sensing to medical diagnosis and retrieval in large image databases. In addition, object recognition, optical flow and stereopsis algorithms often depend on high quality image segmentations. The segmentation problem can be informally described as partitioning an image into homogeneous regions. For textured images one of the main conceptual difficulties is the definition of a homogeneity measure in mathematical terms. Many explicit texture models have been considered in the last three decades. For example, textures are often represented by feature vectors, by the means of a filter bank output [1], wavelet coefficients [2] or as parameters of an explicit Markov random field model [3]. Feature-based approaches 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 [4], that the segmentation problem can be defined in terms of pairwise dissimilarities between textures without extracting explicit texture features.

Once an appropriate homogeneity measure has been identified, unsupervised texture segmentation can be formulated as a constrained combinatorial optimization problem known as *pairwise data clustering* [5], which is NP-hard in the general case. It is the aim of this paper to develop practicable efficient optimization heuristics.

* Supported by the German Research Foundation (DFG # BU 914/3-1) and by the Federal Ministry for Education, Science and Technology (BMBF # 01 M 3021 A/4).

Our approach to unsupervised texture segmentation is based on four cascaded design decisions, concerning the questions of image representation, texture homogeneity measures, objective functions and optimization procedures.

1. We use a *Gabor wavelet* scale–space representation with frequency–tuned filters as a natural image representation.
2. Homogeneity between pairs of texture patches is measured by a *non–para-metric* statistical test applied to the empirical feature distribution functions of locally sampled Gabor coefficients.
3. Due to the nature of the pairwise proximity data, we systematically derive a family of *pairwise clustering objective functions* based on invariance properties to formalize the segmentation problem. The objective functions are extended to sparse data in order to achieve additional computational efficiency.
4. As an optimization technique we apply *deterministic annealing* to derive heuristic algorithms for efficient minimization of the clustering objective functions.

This novel optimization approach combines advantages of simulated annealing with the efficiency of a deterministic procedure and has been applied successfully to a variety of combinatorial optimization problems [6–10] and computer vision tasks [11,12]. The method is presented in a unifying way for a larger class of partitioning problems and extend the pairwise clustering algorithm derived in [5] to sparse dissimilarity data. To demonstrate the capability of this optimization approach, we present a rigorous mathematical framework for the development of continuation, ‘GNC–like’ [13] algorithms. This also clarifies the intrinsic connection between mean–field approximation and Gibbs sampling [14]. Novel convergence proofs significantly extending [15] are given.

2 Image Representation and Non–parametric Homogeneity Measures

In the following we summarize some of the details specific to our texture segmentation approach [16,17]. The choice of a *scale space* image representation overcomes one of the key difficulties in unsupervised texture segmentation, which is the detection of the characteristic scale of a texture. Since natural textures arise at a wide range of scales, scale space methods are a promising approach for texture segmentation and texture classification. We choose a Gabor filter representation, as their good discrimination properties for textures are well-known [1,18].

The idea of applying statistical tests to compare local feature distributions is due to Geman et al. [4], where the Kolmogorov–Smirnov distance was proposed as a similarity measure. We have intensively investigated additional non–parametric tests with respect to their texture discrimination ability [16]. As a result throughout this work a χ^2 –statistic is used. To apply the test, the image is partitioned into overlapping blocks, which are centered on a regular grid. For each such block \mathbf{B}_i of size n and each Gabor channel $1 \leq r \leq L$ the empirical distribution of Gabor coefficients $(b_s^r)_{1 \leq s \leq n}$, $f_i^r(t) = |\{t_{k-1} \leq b_s^r \leq t_k\}|/n$, $t \in [t_{k-1}, t_k]$, is calculated, where $(t_k)_{0 \leq k \leq M}$ represents an appropriate binning. The dissimilarity between two blocks $(\mathbf{B}_i, \mathbf{B}_j)$ with

respect to channel r is then given by

$$D_{ij}^{(r)} = \sum_{k=1}^M \frac{(f_i^r(t_k) - \hat{f}^r(t_k))^2}{\hat{f}^r(t_k)}, \quad \text{where } \hat{f}^r(t_k) = \frac{f_i^r(t_k) + f_j^r(t_k)}{2}. \quad (1)$$

These values are finally combined to obtain $D_{ij} = \sum_{r=1}^L D_{ij}^{(r)}$. There are several advantages in using a statistical test in this context. First, the result of a statistical test is directly interpretable in terms of statistical confidence and is largely independent of the specific representation and the image domain. Second, the empirical distribution functions of features preserve significantly more information than commonly used moment statistics. Third, the problem of finding the appropriate metric in the feature space is avoided.

To guarantee computational efficiency the evaluation of dissimilarity values for a block \mathbf{B}_i is restricted to a substantially reduced *neighborhood* \mathcal{N}_i , $|\mathcal{N}_i| \ll N$, where a *neighborhood system* $\mathcal{N} = (\mathcal{N}_i)_{i=1, \dots, N}$, $\mathcal{N}_i \subset \{1, \dots, N\}$ is defined as an irreflexive and symmetric binary relation. Notice, that long range interactions are essential to correctly segment unconnected areas of identical textures [4].

3 Deterministic Annealing for Partitioning, Clustering and Segmentation Problems

3.1 Partitioning and Clustering Problems

In this section we consider combinatorial optimization problems, where a set of N objects is assigned to a certain number of K groups or labels. In the texture segmentation problem these ‘objects’ are image blocks \mathbf{B}_i and the labels represent different texture types. If the number of distinctive classes K is known a priori, an assignment is simply given by a total mapping $\Pi : \{1, \dots, N\} \rightarrow \{1, \dots, K\}$, which we represent by indicator functions $M_{i\nu} \in \{0, 1\}$ for each predicate $\Pi(i) = \nu$. All assignments are summarized in terms of a Boolean assignment matrix $\mathbf{M} \in \mathcal{M}$, where

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

Throughout this paper any function $\mathcal{H} : \mathcal{M} \rightarrow \mathbb{R}$ will be called an *partitioning objective function* or *partitioning problem*. In this work we focus on objective functions that measure the intra-cluster compactness and depend only on the homogeneity of a cluster. The simplest choice of a cost function, which corresponds to the one proposed in [4], is the (*unnormalized*) standard graph partitioning cost function:

$$\mathcal{H}^{\text{un}}(\mathbf{M}) = \sum_{\nu=1}^K \sum_{i=1}^N \sum_{j \in \mathcal{N}_i} M_{i\nu} M_{j\nu} D_{ij}. \quad (2)$$

We adopt the principles of *invariance with respect to additive shifts* and *invariance with respect to rescaling* of the dissimilarities as major guidelines to derive alternative *normalized* clustering objective functions. The most obvious advantage of shift invariance

is the independence on the origin of the dissimilarity function. Note that \mathcal{H}^{un} is not shift-invariant. For example \mathcal{H}^{un} applied to non-negative data as is typical for statistical tests favors equipartitionings, because the costs for a cluster \mathcal{C}_ν scale quadratically with the number of assigned blocks. In the opposite case, the formation of large clusters is favored. Indeed, it has been noticed [4], that the data have to be shifted adequately in order to obtain plausible segmentations. However, if a large number of different textures exists 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}^{un} . We have empirically verified these arguments in our simulations.

Under weak additional regularity assumptions it has been shown [16], that only four shift- and scale-invariant cost functions for sparse dissimilarities exist, which measure intra-cluster compactness.

1. Two normalized objective functions which are equivalent for complete neighborhoods combine average homogeneities proportional to the cluster size.

$$\mathcal{H}^{\text{no}}(\mathbf{M}) = \sum_{\nu=1}^K \sum_{i=1}^N M_{i\nu} \frac{\sum_{j \in \mathcal{N}_i} M_{j\nu} D_{ij}}{\sum_{j \in \mathcal{N}_i} M_{j\nu}}, \quad (3)$$

$$\mathcal{H}^{\text{nc}}(\mathbf{M}) = \sum_{\nu=1}^K \left[\sum_{i=1}^N M_{i\nu} \right] \frac{\sum_{i=1}^N \sum_{j \in \mathcal{N}_i} M_{i\nu} M_{j\nu} D_{ij}}{\sum_{i=1}^N \sum_{j \in \mathcal{N}_i} M_{i\nu} M_{j\nu}}. \quad (4)$$

2. In addition, there are two normalized objective functions, which combine homogeneities independent of the cluster size. Again, both versions are equivalent for complete neighborhoods.

$$\mathcal{H}^{\text{sno}}(\mathbf{M}) = \sum_{\nu=1}^K \sum_{i=1}^N \frac{M_{i\nu}}{\sum_{i=1}^N M_{i\nu}} \cdot \frac{\sum_{j \in \mathcal{N}_i} M_{j\nu} D_{ij}}{\sum_{j \in \mathcal{N}_i} M_{j\nu}}, \quad (5)$$

$$\mathcal{H}^{\text{snc}}(\mathbf{M}) = \sum_{\nu=1}^K \frac{\sum_{i=1}^N \sum_{j \in \mathcal{N}_i} M_{i\nu} M_{j\nu} D_{ij}}{\sum_{i=1}^N \sum_{j \in \mathcal{N}_i} M_{i\nu} M_{j\nu}}. \quad (6)$$

There are two properties, which distinguish the four cost functions. The first property is only induced by the sparseness of the neighborhood system and vanishes in the complete data limit. It concerns the question, whether every object in a cluster should have the same influence on the total costs (\mathcal{H}^{no} and \mathcal{H}^{sno}) or whether the contribution should be proportional to the number of known dissimilarities in the assigned cluster (\mathcal{H}^{nc} and \mathcal{H}^{snc}). For the typical neighborhood size used in the segmentation application this difference turns out to be of minor importance. The second property is fundamental, since it concerns the way cluster compactness measured for single clusters is summed up to give the total clustering costs. Cluster homogeneities can either be weighted with the cluster sizes (\mathcal{H}^{no} , \mathcal{H}^{nc}) or combined irrespective of their size (\mathcal{H}^{sno} , \mathcal{H}^{snc}). The later has the tendency to create small clusters, because it is always simpler to find clusters of higher homogeneity with few objects. For this reason, we propose to utilize prior costs to penalize unbalanced data partitionings. In addition, taking advantage of the fact that image segments for natural scenes are expected to form connected components, we also

include a topological prior,

$$\mathcal{H}^{\text{Pr}}(\mathbf{M}) = \lambda_s \sum_{\nu=1}^K \left(\sum_{i=1}^N M_{i\nu} \right)^2 + \lambda_t \sum_{\nu=1}^K \sum_{i=1}^N M_{i\nu} \sum_{j \in \mathcal{T}_i} (1 - M_{j\nu}), \lambda_s, \lambda_t \in \mathbf{R}^+. \quad (7)$$

Here \mathcal{T}_i denotes a topological neighborhood of \mathbf{B}_i , e.g., the four-connected neighborhood of image blocks left, right, above and below of \mathbf{B}_i . More complex topological priors to forbid small and thin regions can be introduced by hard constraints as proposed in [4], but additional constraints restrict the development of efficient optimization algorithms. For this reason, we decided to extend the optimization by a post-processing stage, where the clustering solution is used as an initial configuration for an MRF-model to find a valid image partitioning respecting all constraints.

3.2 Principles of Deterministic Annealing

In recent years the stochastic optimization strategy *Simulated Annealing* has become popular to solve image processing tasks [14]. The random search is modeled by an inhomogeneous discrete-time Markov chain $(\mathbf{M}^{(t)})_{t \in \mathbf{N}}$, which stochastically samples the solution space. Since the configuration space for partitioning problems naturally decomposes into single site configurations $\mathcal{M} = \bigotimes_i \mathcal{M}_i$, we focus on a restricted class of *local* algorithms, that perform only state transitions between configurations, which differ in the assignment of at most one site. Denote by $\tilde{\mathbf{M}}_\alpha = s_i(\mathbf{M}, \mathbf{e}_\alpha)$ the matrix obtained by substituting the i -th row of \mathbf{M} by the unity vector \mathbf{e}_α . For convenience we introduce a *site visitation schedule* as a map $v : \mathbf{N} \rightarrow \{1, \dots, N\}$ fulfilling $\lim_{U \rightarrow \infty} \# \{t \leq U : v(t) = i\} \rightarrow \infty$ for all i . A sampling scheme known as the *Gibbs sampler* [14] is advantageous, if it is possible to efficiently sample from the conditional distribution at site $v(t)$, given the assignments at all other sites $\{j \neq v(t)\}$. For a given site visitation schedule v the Gibbs sampler is defined by the non-zero transition probabilities

$$S_t(s_i(\mathbf{M}, \mathbf{e}_\alpha), \mathbf{M}) = \frac{\exp[-\mathcal{H}(s_i(\mathbf{M}, \mathbf{e}_\alpha))/T(t)]}{\sum_{\nu=1}^K \exp[-\mathcal{H}(s(\mathbf{M}, \mathbf{e}_\nu))/T(t)]}, i = v(t). \quad (8)$$

The site visitation schedule guarantees the irreducibility of the Markov chain. For a constant temperature $T = T(t)$, the Markov chain defined by (8) will converge towards its equilibrium distribution, which is the *Gibbs distribution*

$$P_{\mathcal{H}}(\mathbf{M}) = \frac{1}{Z_T} \exp(-\mathcal{H}(\mathbf{M})/T), \quad Z_T = \sum_{\mathbf{M} \in \mathcal{M}} \exp(-\mathcal{H}(\mathbf{M})/T). \quad (9)$$

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

$$\mathcal{F}_T(P) = \langle \mathcal{H} \rangle_P - TS(\mathbf{P}) = \sum_{\mathbf{M} \in \mathcal{M}} P(\mathbf{M})\mathcal{H}(\mathbf{M}) + T \sum_{\mathbf{M} \in \mathcal{M}} P(\mathbf{M}) \log P(\mathbf{M}) \quad (10)$$

the *generalized free energy*, which plays the role of an objective function over $\mathcal{P}_{\mathcal{M}}$. For arbitrary partitioning problems \mathcal{H} the Gibbs distribution $P_{\mathcal{H}}$ minimizes the generalized free energy, i.e., $P_{\mathcal{H}} = \arg \min_{P \in \mathcal{P}_{\mathcal{M}}} \mathcal{F}_T(P)$. The basic idea of annealing

is to use Gibbs sampling, but to gradually lower the temperature $T(t)$, on which the transition probabilities depend. For the zero temperature limit a deterministic local optimization algorithm known as *Iterative Conditional Mode* (ICM) [19] is obtained. Both algorithms are used for benchmarking in the segmentation experiments.

After a transient phase, a stochastic search according to a Markov process samples from the canonical Gibbs distribution. Gibbs expectation values for random variables $X(\mathbf{M})$ can thus be approximated by ergodic time-averages. The main disadvantage is the fact that stochastic techniques can be extremely slow. On the other hand, a slow annealing will often produce solutions of a very high quality, while gradient based methods are very sensible to (bad) local minima. 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, analytically. In DA a combinatorial optimization problem with objective function \mathcal{H} over \mathcal{M} is relaxed to a family of stochastic optimization problems with objective functions \mathcal{F}_T over a subspace $\mathcal{Q}_{\mathcal{M}} \subseteq \mathcal{P}_{\mathcal{M}}$. Obviously, the discrete search space \mathcal{M} can be canonically embedded in $\mathcal{P}_{\mathcal{M}}$ by the injective mapping $e: \mathcal{M} \rightarrow \mathcal{P}_{\mathcal{M}}$, where $e(\mathbf{M}) = P_{\mathbf{M}}$ is defined as the Dirac distribution at \mathbf{M} . In order for $\mathcal{Q}_{\mathcal{M}}$ to be a true relaxation we demand $e(\mathcal{M}) \subseteq \mathcal{Q}_{\mathcal{M}}$. The subspace $\mathcal{Q}_{\mathcal{M}}$, which we will discuss in the context of partitioning problems, is the space of all factorial distributions given by

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

$\mathcal{Q}_{\mathcal{M}}$ is distinguished from other subspaces of $\mathcal{P}_{\mathcal{M}}$ in many respects. First, the dimensionality of $\mathcal{Q}_{\mathcal{M}}$ increases only linearly with N . Second, an efficient alternation algorithm exists for a very general class of objective functions, which converges towards a local minimum of \mathcal{F}_T in $\mathcal{Q}_{\mathcal{M}}$. Third, in the limit of $T \rightarrow 0$ solutions to the combinatorial optimization problem can be recovered, which are locally optimal with respect to single site changes [16].

While we recover the original combinatorial problem for $T \rightarrow 0$, the generalized free energy \mathcal{F}_T becomes convex at high temperatures, since the entropy S is convex. Furthermore \mathcal{F}_T also becomes convex over $\mathcal{Q}_{\mathcal{M}}$ for sufficiently large T (cf. Theorem 4). Thus \mathcal{F}_T is an entropy-smoothed version of the original optimization problem, where more and more details of the original objective function appear as T is lowered. In DA a solution is tracked from high temperatures, where \mathcal{F}_T is convex, to low temperatures, where the minimization of \mathcal{F}_T becomes as hard as minimizing \mathcal{H} over \mathcal{M} . This approach relies on the possibility of minimizing the generalized free energy over $\mathcal{Q}_{\mathcal{M}}$. For the unrestricted case of $\mathcal{Q}_{\mathcal{M}} = \mathcal{P}_{\mathcal{M}}$ we know, that the solution is the temperature dependent Gibbs distribution $P_{\mathcal{H}}$. As a consequence, DA will only result in a tractable procedure for $\mathcal{P}_{\mathcal{M}}$, if an explicit summation over \mathcal{M} can be avoided, since the calculation of assignment probabilities would require an exhaustive overall evaluation of \mathcal{H} . In this perspective, the relaxation to factorial distributions is an approximation of a continuation method, which is intractable in $\mathcal{P}_{\mathcal{M}}$. The approximation accuracy can be expressed by the cross entropy $\mathcal{I}(Q||P_{\mathcal{H}})$, which is automatically minimized by minimizing \mathcal{F}_T over $\mathcal{Q}_{\mathcal{M}}$, as $\mathcal{F}_T(Q) = \frac{1}{T} [\mathcal{I}(Q||P_{\mathcal{H}}) - \log \mathcal{Z}_T]$ for all $Q \in \mathcal{P}_{\mathcal{M}}$.

There is a strong motivation for the maximum entropy framework. First, maximizing the entropy yields the least biased inference method being *maximally noncommittal with respect to missing data* [20]. Second, the maximum entropy probability distribution changes the least in terms of the L_2 norm if the expected costs $\langle \mathcal{H} \rangle$ are lowered or raised by changes of the temperature [21], which stresses the robustness of this inference technique. We conclude that a stochastic search heuristic, which starts with a large noise level and which gradually reduces stochasticity to zero, should ideally follow the trajectory defined by the family of Gibbs distributions with decreasing temperature.

3.3 Mean-Field Approximation

Now we concentrate on the space of factorial distribution. The resulting scheme is known as *mean-field approximation*. We will use the more specific term *Mean-Field Annealing* (MFA) instead of DA, if $P_{\mathcal{H}} \notin \mathcal{Q}_{\mathcal{M}}$. A factorial distributions Q can be transformed into the Gibbs normal form.

$$Q(\mathbf{M}) = \exp \left[-\frac{1}{T} \sum_{i=1}^N \sum_{\nu=1}^K M_{i\nu} (-T \log q_{i\nu}) \right]. \quad (12)$$

Thus, factorial distributions could alternatively be defined by Gibbs distributions with linear Hamiltonians of the type $\mathcal{H}^0(\mathbf{M}) = \sum_{i=1}^N \sum_{\nu=1}^K M_{i\nu} h_{i\nu}$, where $h_{i\nu} \in \mathbb{R}$ are $N \cdot K$ variational parameters, which are often called *mean-fields* by physical analogy. The most important relations for factorial distributions are summarized in the following proposition, the proof of which can be found in [16].

Proposition 1. *Let $\mathcal{H}^0(\mathbf{M})$ be a linear partitioning objective function. Denote by $Q = P_{\mathcal{H}^0}$ the associated Gibbs distribution.*

1. *The partition function and the free energy are given by*

$$\mathcal{Z}_T^0 = \prod_{i=1}^N \sum_{\nu=1}^K \exp[-h_{i\nu}/T], \quad \mathcal{F}_T^0 = -T \sum_{i=1}^N \log \sum_{\nu=1}^K \exp[-h_{i\nu}/T].$$

2. *An equivalent reparametrization of Q according to (11) is obtained by*

$$q_{i\nu} = \langle M_{i\nu} \rangle_Q = \frac{\partial \mathcal{F}_T^0}{\partial h_{i\nu}} = \frac{\exp[-\frac{1}{T} h_{i\nu}]}{\sum_{\mu=1}^K \exp[-\frac{1}{T} h_{i\mu}]}.$$

The inverse transformation is only unique up to an additive constant and is given by $h_{i\nu} = -T \log q_{i\nu} + c_i$, where c_i is arbitrary. Thus the parameters $q_{i\nu}$ can be identified with the Q -averages $\langle M_{i\nu} \rangle$.

3. *All correlations w.r.t. Q vanish for assignment variables at different sites, e.g.,*

$$\langle M_{i\nu} M_{j\mu} \rangle_Q = \langle M_{i\nu} \rangle_Q \langle M_{j\mu} \rangle_Q, \nu \neq \mu. \quad (13)$$

Calculating stationary conditions from (10), a system of coupled transcendental, so-called *mean-field equations*, is obtained, which can be efficiently solved by a convergent iteration scheme. For factorial distributions the equation system takes the following general form.

Proposition 2. Let \mathcal{H} be an arbitrary partitioning cost function. 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

$$h_{i\nu}^* = \frac{\partial \langle \mathcal{H} \rangle_{Q^*}}{\partial q_{i\nu}} = \frac{1}{q_{i\nu}^*} \langle M_{i\nu} \mathcal{H} \rangle_{Q^*} . \quad (14)$$

3.4 Mean-Field Equations and Gibbs Sampling

There is a tight relationship between the quantities $g_{i\nu} = \mathcal{H}(s_i(\mathbf{M}, \mathbf{e}_\nu))$ involved in implementing the Gibbs sampler in (8) and the mean-field equations. Rewriting (14) we arrive at

$$h_{i\nu}^* = \sum_{\mathbf{M} \in \mathcal{M}} \frac{M_{i\nu}}{q_{i\nu}^*} \mathcal{H}(\mathbf{M}) Q^*(\mathbf{M}) = \sum_{\mathbf{M} \in \mathcal{M}} \mathcal{H}(s_i(\mathbf{M}, \mathbf{e}_\nu)) Q^*(\mathbf{M}) . \quad (15)$$

This proves the following theorem.

Theorem 3. The mean-fields $h_{i\nu}^*$ are a Q -averaged version of the local costs $g_{i\nu}$. Thus Q^* is characterized by

$$q_{i\nu}^* = \frac{\exp[-\frac{1}{T}h_{i\nu}^*]}{\sum_{\mu=1}^K \exp[-\frac{1}{T}h_{i\mu}^*]}, \quad h_{i\nu}^* = \langle g_{i\nu} \rangle_{Q^*} . \quad (16)$$

This relationship can be further clarified by the Markov blanket identity, also known as the Callen equation [5].

$$\langle M_{i\nu} \rangle_{P_{\mathcal{H}}} = \frac{1}{\mathcal{Z}_T} \sum_{\mathbf{M} \in \mathcal{M}} M_{i\nu} \exp[-\mathcal{H}(\mathbf{M})/T] = \left\langle \frac{\exp[-g_{i\nu}/T]}{\sum_{\mu=1}^K \exp[-g_{i\mu}/T]} \right\rangle_{P_{\mathcal{H}}} . \quad (17)$$

The Markov blanket identity is a relation between Gibbs expectations, corresponding to the probabilistic equation of the Gibbs sampler in (8) at equilibrium. From this perspective, the mean-field approximation is seen to be equivalent to a two step approximation, which interchanges the averages with the non-linearity in (17) and neglects fluctuations in averaging the exponents.

3.5 Convergence of Mean-Field Annealing

Theorem 3 implies an optimization procedure, which converges to a local minimum of the generalized free energy.

Theorem 4. For any site visitation schedule v and arbitrary initial conditions, the following asynchronous update scheme converges to a local minimum of the generalized free energy (10):

$$q_{i\nu}^{\text{new}} = \frac{\exp[-\frac{1}{T}h_{i\nu}]}{\sum_{\mu=1}^K \exp[-\frac{1}{T}h_{i\mu}]}, \quad \text{where } h_{i\nu} = \langle g_{i\nu} \rangle_{Q^{\text{old}}} \quad \text{and } i = v(t) . \quad (18)$$

Furthermore \mathcal{F}_T is strictly convex over $\mathcal{Q}_{\mathcal{M}}$ for T sufficiently large.

A proof can be found in [16]. Notice, that the variables $h_{i\nu}$ 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. For polynomial \mathcal{H} it is straightforward to compute the expectations of the Gibbs–fields because of Proposition 1.2 and 1.3. The convergent update scheme together with the convexity for large T leads to a GNC–like [13] algorithm, a result which does not extend to non–binary MFA–algorithms in general [22].

MFA Algorithm

```

INITIALIZE  $q_{i\nu}$  randomly, temperature  $T \leftarrow T_0$ ;
WHILE  $T > T_{\text{FINAL}}$ 
  add a small random perturbation to all  $q_{i\nu}$ ;
  REPEAT
    generate a permutation  $\pi \in S_N$ ;
    FOR  $i=1, \dots, N$ 
      update all  $q_{\pi(i)\nu}$  according to (18);
    UNTIL converged ;
   $T \leftarrow \eta \cdot T$ ,  $0 < \eta < 1$ ;
END

```

3.6 Gibbs Sampling and Deterministic Annealing for Pairwise Clustering

To efficiently implement the Gibbs sampler one has to optimize the evaluation of \mathcal{H} for a sequence of locally modified assignment matrices. It is an important observation that the quantities $g_{i\nu}$ only have to be computed up to an additive shift, which may depend on the site index i , but not on ν . The choice of $g_{i\nu}(\mathbf{M}) = \mathcal{H}(s_i(\mathbf{M}, \mathbf{e}_\nu)) - \mathcal{H}(s_i(\mathbf{M}, 0))$ leads to compact analytical expressions, because the contributions of the reduced system without site i are subtracted.

Following Theorem 3 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, to avoid exponential order in the number of conjunctions some approximations have to be made. We do this by independently averaging the numerator and the normalization in the denominator. Using Proposition 1.3 this leads to $h_{i\nu}(\mathbf{M}) = g_{i\nu}(\langle \mathbf{M} \rangle)$, which is exact in the limit of $T \rightarrow 0$, since the susceptibilities $\langle M_{i\nu} \rangle (1 - \langle M_{i\nu} \rangle)$ vanish exponentially fast at low temperatures. The approximation quality as well as an efficient implementation by proper book–keeping is further discussed in [16]. As an example we explicitly display the result for the cost function \mathcal{H}^{nc} . With $Q_{i\nu}^- = \sum_{j \neq i} \sum_{k \in \mathcal{N}_j, k \neq i} \langle M_{j\nu} \rangle \langle M_{k\nu} \rangle$ and $Q_{i\nu}^+ = Q_{i\nu}^- + 2 \sum_{j \in \mathcal{N}_i} \langle M_{j\nu} \rangle$ we obtain

$$h_{i\nu} = \frac{2}{Q_{i\nu}^+} \sum_{j \in \mathcal{N}_i} \langle M_{j\nu} \rangle D_{ij} - \frac{2 \sum_{j \in \mathcal{N}_i} \langle M_{j\nu} \rangle}{Q_{i\nu}^+ Q_{i\nu}^-} \sum_{j \neq i} \sum_{k \in \mathcal{N}_j, k \neq i} \langle M_{j\nu} \rangle \langle M_{k\nu} \rangle D_{jk} \quad (19)$$

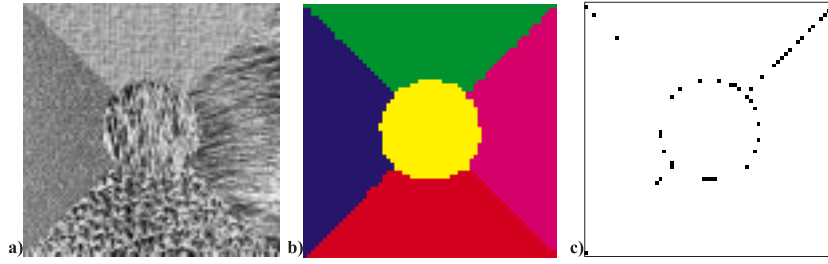


Fig. 1. Typical segmentation result with $K = 5$: (a) Randomly generated image. (b) Segmentation on the basis of the normalized costs \mathcal{H}^{no} . (c) Misclassified blocks (depicted in black).

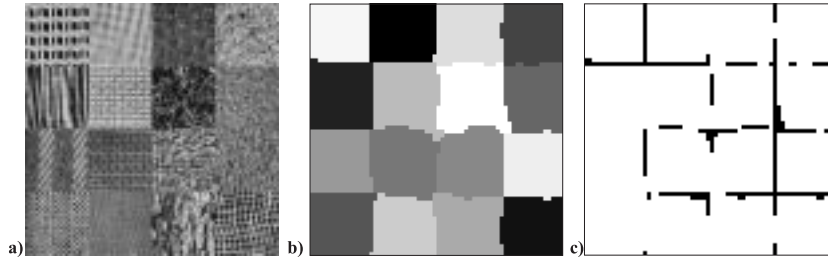


Fig. 2. Typical segmentation result with $K = 16$: (a) Randomly generated image. (b) Segmentation on the basis of the normalized costs \mathcal{H}^{nc} . (c) Misclassified blocks (depicted in black). For the segmentation an average neighborhood size of $|\mathcal{N}_i| = 300$ was used.

4 Results

To empirically test the segmentation algorithms on a wide range of textures we selected a representative set of 40 Brodatz-like micro-patterns. From this collection a database of random mixtures of size 512×512 pixels, containing 100 entities of five textures (as depicted in Fig. 1(a)) was constructed. All segmentations are based on a filter bank of twelve Gabor filters at four orientations and three octave scales. Each image was divided into 64×64 overlapping blocks of size 16×16 pixels each. Dissimilarities have been evaluated for an average neighborhood size of $|\mathcal{N}_i| = 80$, including the four-connected neighborhood in the image. Typical segmentation examples using the normalized cost functions are shown in Fig. 1 and 2. It has been empirically verified that the algorithms are insensitive to variation of parameters such as neighborhood size or cooling schedule, which were chosen conservatively. Typical run-times on a Sun Ultra-Sparc are about 3 minutes for the clustering stage. The used database and additional examples are available via World Wide Web (WWW).

The first question, which is empirically investigated, addresses the problem of how adequate texture segmentation is modeled by the extracted proximity data and the presented cost functions. Figure 3 shows the distribution of misclassified blocks. The distributions for the other normalized cost functions under examination are similar. For \mathcal{H}^{no} a median segmentation error rate as low as 2.83% (6.84% before post-processing)

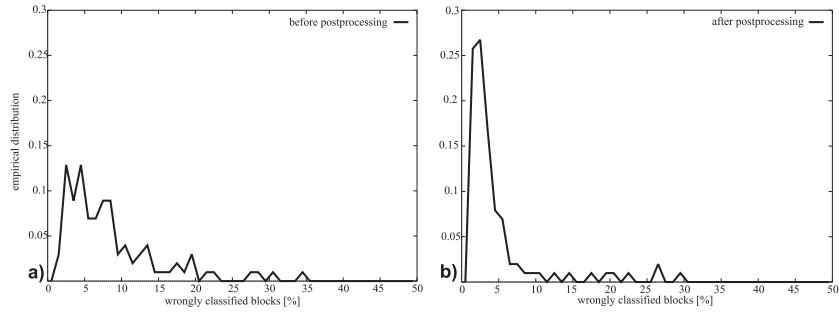


Fig. 3. Empirical density of the percentage of misclassified blocks for the database with five textures each: before (a) and after (b) post-processing. The diagram depicts the results achieved for the normalized cost function \mathcal{H}^{no} with the MFA algorithm.

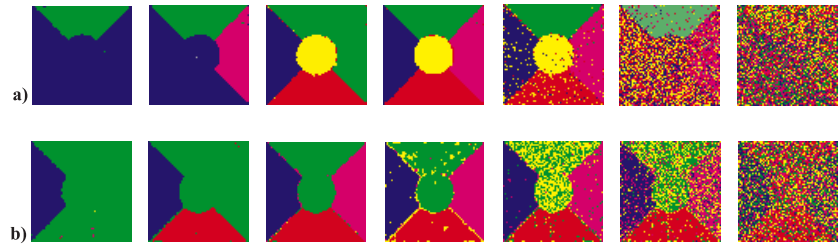


Fig. 4. Segmentations for two example images obtained by \mathcal{H}^{un} for several data shifts. From the left segmentations with a mean dissimilarity of $-0.05, 0, 0.05, 0.1, 0.15, 0.2$ and 0.25 are depicted. Segments start collapsing for negative shifts. For large positive shifts the obtained segmentations become random, because the sampling noise induced by the random neighborhood system dominates the data contributions.

was obtained, which was only beaten by the \mathcal{H}^{nc} cost function with a median error of 2.65% (7.12% before post-processing). For \mathcal{H}^{snc} the error was 3.56% (7.50%). \mathcal{H}^{sno} was excluded from the empirical investigations, because the MFA and Gibbs sampling implementation is inefficient compared to \mathcal{H}^{snc} and the quality differences are expected to be marginal. We conclude, that in most cases the normalized cost functions based on a pairwise data clustering formalization capture the true structure. As can be seen in Fig. 1 (c) the misclassified blocks mainly correspond to errors at texture borders. The post-processing step improves the segmentations by a significant noise reduction. The unnormalized cost function \mathcal{H}^{un} severely suffers from the missing shift-invariance property as shown in Fig. 4. Depending on the shift the unnormalized cost function often misses several full texture classes. As seen in Fig. 4 (b) there may not even exist any parameter value to find all five textures. Even worse the optimal value depends on the data at hand and varies for different images. With the unnormalized cost function \mathcal{H}^{un} we achieved a median error rate of 3.86% (9.50% before post-processing) after ex-

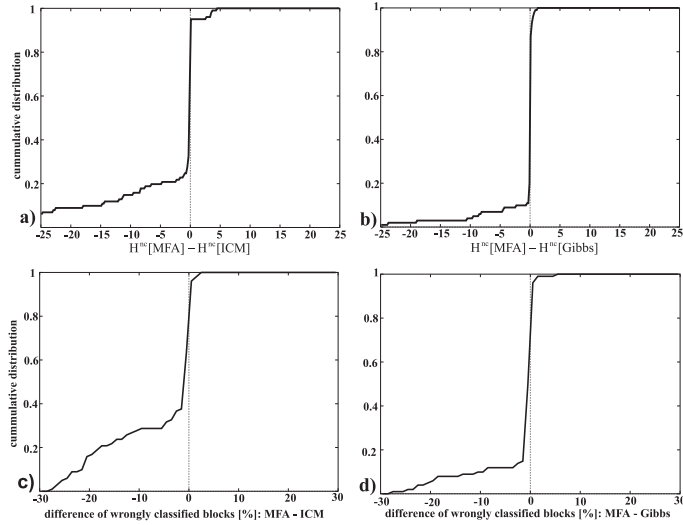


Fig. 5. The empirical density of the cost difference/misclassification rate of deterministic annealing versus the ICM algorithm (a)/(c) and versus the Gibbs sampler (b)/(d).

tensive tuning to find the appropriate data shift. A further deterioration on images with largely varying texture sizes was observed.

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 of the achieved segmentation with the deterministic ICM algorithm and with Gibbs sampling. Exemplary for the normalized cost functions the cost differences for \mathcal{H}^{nc} using MFA versus ICM (Fig. 5(a)) and MFA versus Gibbs sampler (Fig. 5(b)) are depicted. Compared with the ICM algorithm a substantial improvement has to be noted, since the ICM algorithm gets frequently stuck in bad local minima. For the comparison with the Gibbs sampler we decided to use the same number of updates for both MFA and Gibbs sampling, although the running time for the Gibbs sampler is slightly superior. As depicted in Fig. 5 MFA yields much better results. In the few cases where the other algorithms yield better solutions the achievements are insignificantly small. We have also compared the algorithms w.r.t. the percentage of misclassifications instead of energy, see 5 (c),(d). The better optimization procedure leads to substantial improvements in the segmentation quality, which is not trivial, as the global optimum of the cost function does not necessarily correspond to the ground truth segmentation.

To visualize the annealing process, we have taken the example displayed in Fig. 1 and show solutions with different number of effective clusters [7] at different temperatures. Obviously we obtain more information than just a single image segmentation. For $K = 5$ effective clusters the decrease of the mean energy starts to flatten and no phase transition occurred for the maximal range of ≈ 0.02 units. This information can be used to decide upon the question of the optimal number of clusters.

To demonstrate the applicability of the presented MFA algorithm for \mathcal{H}^{nc} to real-world images we performed tests on three types of images. An important application is the

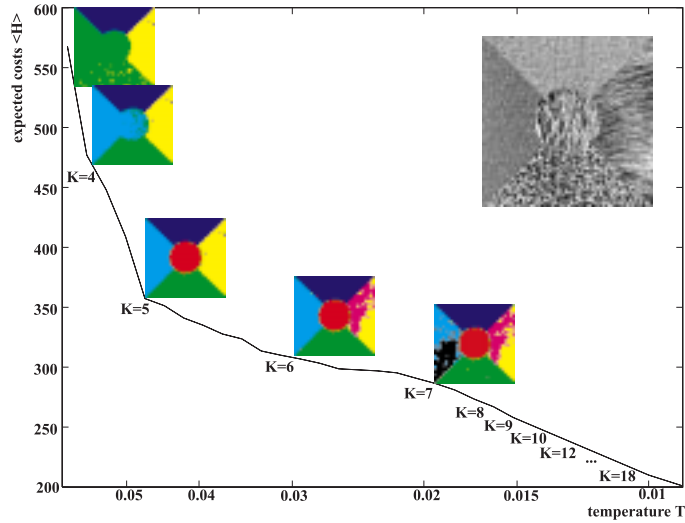


Fig. 6. Mean energy $\langle \mathcal{H}^{nc} \rangle$ and effective number of clusters as a function of T .

segmentation of Synthetic Aperture Radar (SAR) images. In Fig. 7 the segmentation into three texture classes of a SAR image is depicted. The achieved segmentation is both visually and semantically correct. Mountains, valleys and plain areas are well-separated. Even small valley structures are detected. A second interesting class of images are aerial images, as many aerial images contain texture-like structures as for example seen in Fig. 8, where an aerial image of San Francisco is segmented. The solution for $K = 8$ is visually satisfying. Tilled area and parks as well as water are well-discriminated. A third class of applications for texture segmentation are indoor and outdoor images, which contain textured objects. Unsupervised segmentation has important applications in autonomous robotics and the presented algorithms are currently implemented on the autonomous robot RHINO [23]. An example image of a typical office environment is presented in Fig. 9. Untextured parts of the image are grouped together irrespective of their absolute luminance and the discrimination of the remaining three textures is very plausible.

5 Discussion

The main contribution of the paper with respect to the development of efficient and scalable optimization algorithms is the rigorous mathematical framework for applying the optimization principle of deterministic annealing to arbitrary partitioning problems. Also guided by physical insight the framework has been developed from a purely algorithmic perspective to construct efficient GNC-like continuation methods with guaranteed convergence. The canonical way to obtain mean-field equations as well as an efficient implementation of Gibbs sampling for the proposed objective functions have been presented. The intrinsic connection with the Gibbs-sampler has been exploited

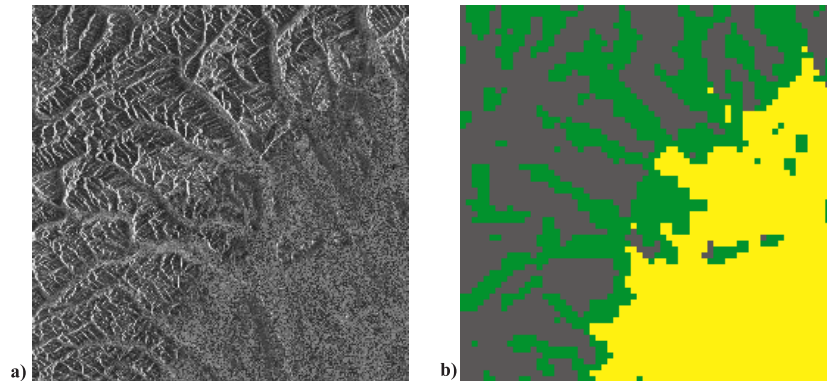


Fig. 7. SAR image of a mountain landscape: (a) original image, (b) segmentation into three clusters ($\lambda_t = 0.01$) without post-processing.

to derive efficient MFA-algorithms. The framework is general enough to adopt to a broad range of other possible clustering and partitioning applications. For the problem of unsupervised texture segmentation we have demonstrated that statistical tests are a powerful technique for texture discrimination without the need of parametric assumptions or explicit texture models. Moreover, objective functions were derived which have proven to be in very good agreement with ground truth data for an image database generated from a large collection of textures. In all these simulations covering a wide range of image domains, we have been using the same algorithms without the need of parameter tuning or learning. This is the reason, why we consider our approach to be *unsupervised* in a strict sense. As is generally true for optimization approaches to computer vision problems, this would nevertheless not be of practical use without the existence of efficient optimization algorithms. Our derivation of a deterministic annealing algorithm provides a solution to this problem. We strongly believe that deterministic annealing algorithms for related computer vision problems can be derived along the same lines.

References

1. A. Jain and F. Farrokhnia, "Unsupervised texture segmentation using Gabor filters," *Pattern Recognition*, vol. 24, no. 12, pp. 1167–1186, 1991.
2. O. Pichler, A. Teuner, and B. Hosticka, "A comparison of texture feature extraction using adaptive Gabor filtering, pyramidal and tree-structured wavelet transforms," *Pattern Recognition*, vol. 29, no. 5, pp. 733–742, 1996.
3. J. Mao and A. Jain, "Texture classification and segmentation using multiresolution simultaneous autoregressive models," *Pattern Recognition*, vol. 25, pp. 173–188, 1992.
4. D. Geman, S. Geman, C. Graffigne, and P. Dong, "Boundary detection by constrained optimization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 12, pp. 609–628, July 1990.
5. T. Hofmann and J. Buhmann, "Pairwise data clustering by deterministic annealing," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, 1997.
6. C. Peterson and B. Söderberg, "A new method for mapping optimization problems onto neural networks," *International Journal of Neural Systems*, vol. 1, no. 1, pp. 3–22, 1989.

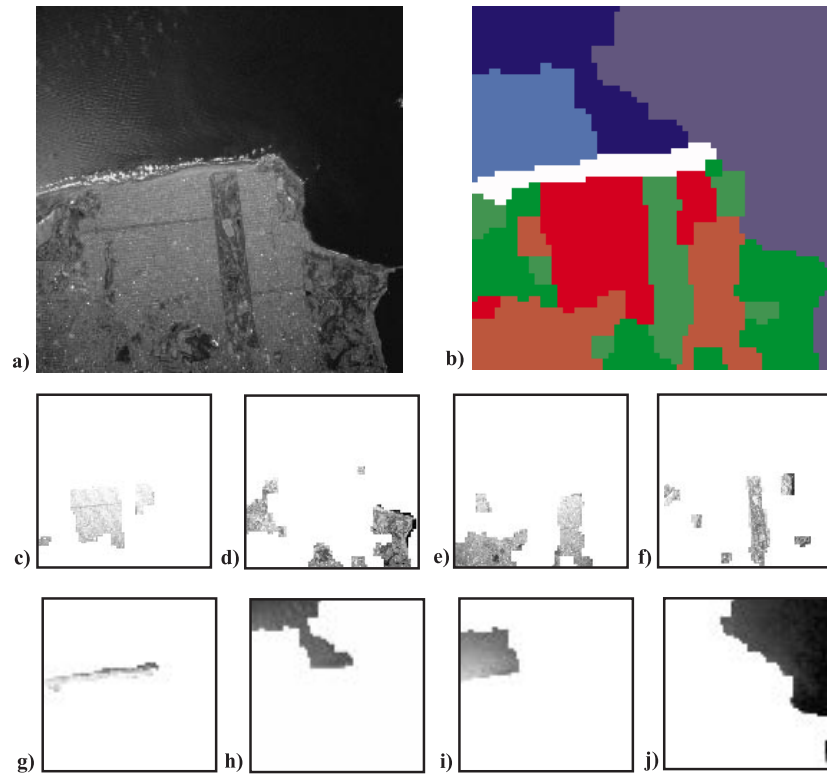


Fig. 8. Aerial image of San Francisco: (a) original grey-scale image, (b) segmentation for $K = 8$ ($\lambda_t = 0.01$) after post-processing, (c) - (j) visualization of the solution.

7. K. Rose, E. Gurewitz, and G. Fox, "Statistical mechanics and phase transition in clustering," *Physical Review Letters*, vol. 65, no. 8, pp. 945–948, 1990.
8. J. Buhmann and H. Kühnel, "Vector quantization with complexity costs," *IEEE Transactions on Information Theory*, vol. 39, pp. 1133–1145, 1993.
9. J. Kosowsky and A. Yuille, "The invisible hand algorithm: Solving the assignment problem with statistical physics," *Neural Networks*, vol. 7, no. 3, pp. 477–490, 1994.
10. S. Gold and A. Rangarajan, "A graduated assignment algorithm for graph matching," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1996.
11. D. Geiger and F. Girosi, "Parallel and deterministic algorithms from MRF's: Surface reconstruction," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 401–412, 1991.
12. J. Zerubia and R. Chellappa, "Mean field annealing using compound Gauss-Markov random fields for edge detection and image estimation," *IEEE Transactions on Neural Networks*, vol. 4, no. 4, pp. 703–709, 1993.
13. A. Blake and A. Zisserman, *Visual Reconstruction*. MIT Press, 1987.
14. S. Geman and D. Geman, "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 6, no. 6, pp. 721–741, 1984.

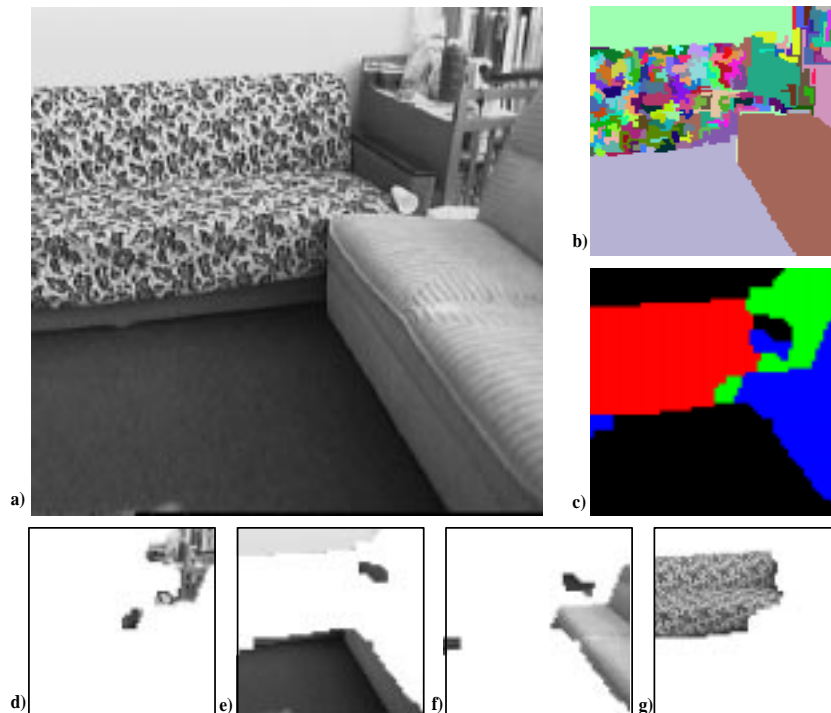


Fig. 9. (a) Indoor image of a typical office environment containing an old-fashioned sofa, (b) contrast based image segmentation with a region merging algorithms, (c) a texture segmentation with $K = 4$ ($\lambda_t = 0.01$). The image partitioning is visualized in (d) - (g).

15. J. Zhang, "The convergence of mean field procedures for MRF's," *IEEE Transactions on Image Processing*, vol. 5, no. 12, pp. 1662–1665, 1996.
16. T. Hofmann, J. Puzicha, and J. Buhmann, "A deterministic annealing framework for textured image segmentation," Tech. Rep. IAI-TR-96-2, Institut für Informatik III, 1996.
17. T. Hofmann, J. Puzicha, and J. Buhmann, "Unsupervised segmentation of textured images by pairwise data clustering," in *Proceedings of the IEEE International Conference on Image Processing, Lausanne*, 1996.
18. J. Puzicha, T. Hofmann, and J. Buhmann, "Unsupervised texture segmentation on the basis of scale space features.," in *Proceedings of the Workshop on Classical Scale-Space Theory, TR DIKU 96/19, University of Copenhagen*, 1996.
19. J. Besag, "On the statistical analysis of dirty pictures," *Journal of the Royal Statistical Society, Series B*, vol. 48, pp. 25–37, 1986.
20. E. Jaynes, "Information theory and statistical mechanics," *Physical Review*, vol. 106, no. 4, pp. 620–630, 1957.
21. Y. Tikhonchinsky, N. Tishby, and R. Levine, "Alternative approach to maximum-entropy inference," *Physical Review A*, vol. 30, no. 5, pp. 2638–2644, 1984.
22. M. Nielsen, "Surface reconstruction: GNCs and MFA," in *Proceedings of the International Congress on Computer Vision*, 1995.
23. J. Buhmann, W. Burgard, A. Cremers, D. Fox, T. Hofmann, F. Schneider, I. Strikos, and S. Thrun, "The mobile robot RHINO," *AI Magazin*, vol. 16, no. 1, 1995.