

Histogram Clustering for Unsupervised Image Segmentation

(to appear in Proceedings of CVPR'99)

Jan Puzicha⁺, Thomas Hofmann^{*}, and Joachim M. Buhmann⁺

⁺ Institut für Informatik III
University of Bonn, Germany
{jan,jb}@cs.uni-bonn.de

^{*} EECS Department, UC Berkeley and
International Computer Science Institute, Berkeley
hofmann@icsi.berkeley.edu

Abstract

This paper introduces a novel statistical mixture model for probabilistic grouping of distributional (histogram) data. Adopting the Bayesian framework, we propose to perform annealed maximum a posteriori estimation to compute optimal clustering solutions. In order to accelerate the optimization process, an efficient multiscale formulation is developed. We present a prototypical application of this method for the unsupervised segmentation of textured images based on local distributions of Gabor coefficients. Benchmark results indicate superior performance compared to K -means clustering and proximity-based algorithms.

1 Introduction

Grouping, segmentation, coarsening, and quantization are omnipresent topics in image processing and computer vision. In the unsupervised case, these tasks essentially correspond to different instances of the clustering problem, i.e., the general goal is to identify groups of *similar* image primitives such as pixels, local features, segments, or even complete images. Two fundamental steps need to be addressed:

- (i) *Modeling problem*: A precise mathematical notion of similarity between image primitives is required in order to formalize the clustering problem.
- (ii) *Computational problem*: For a given similarity measure, an efficient clustering algorithm has to be derived. The selection of a suitable clustering method is tightly coupled to the chosen similarity measure and its underlying data representation. In this contribution, we focus on the unsupervised segmentation of textured images as a prototypical application in low-level computer vision. Numerous techniques to unsupervised texture segmentation have been proposed over the past decades. In many classical approaches, local features are spatially smoothed and represented as *vectors in a metric space* (e.g., in [5, 6]), thereby characterizing each texture by a specific average fea-

ture vector or *centroid*. The most commonly used distortion measure is a (weighted) squared Euclidean norm which effectively models the data by a Gaussian mixture model with one Gaussian for each texture. The clustering method of choice for vectorial data is the K -means algorithm and its variants. Since the Gaussian mixture assumption turns out to be inadequate in many cases, several alternative approaches have utilized *proximity data*, usually obtained by applying statistical tests to the *local feature distribution* at two image sites [2, 4]. As a major advantage, these methods do not require the specification of a suitable vector-space metric. Instead, similarity is defined directly via the respective feature distributions. Several optimization approaches to graph partitioning [2, 4, 11] have been proposed as clustering techniques, which we refer to as *pairwise dissimilarity clustering* (PDC).

The major contribution of this paper is a general method for grouping *feature distributions*, extending a technique known as distributional clustering in statistical language modeling [7]. In contrast to methods based on feature vectors and proximities, this approach is directly applicable to histogram data. In comparison to K -means clustering, distributional clustering naturally includes component distributions with multiple modes. As a major advantage compared to PDC it requires no external similarity measure, but exclusively relies on the feature occurrence statistics. In addition, distributional clustering provides a generative statistical model that can be utilized in subsequent processing steps such as boundary localization [10]. Distributional clustering also offers computational advantages, because it can be implemented efficiently by multiscale optimization techniques [8]. Since feature histograms are processed directly, it avoids time-consuming stages of data extraction (e.g., pairwise comparisons in PDC), which is crucial in real-time applications like autonomous robotics.

2 Mixture Models for Histogram Data

Model Specification To stress the generality of the proposed model we temporarily detach the presentation from the specific problem of image segmentation and consider the following more abstract setting: Let $\mathbf{X} = \{x_1, \dots, x_N\}$ denote a finite set of abstract objects with arbitrary labeling and let $\mathbf{Y} = \{y_1, \dots, y_M\}$ represent a domain of nominal feature(s). The elementary observations consist of *dyadic* measurements $(x, y) \in \mathbf{X} \times \mathbf{Y}$, i.e., *joint occurrences* of elements $x \in \mathbf{X}$ and $y \in \mathbf{Y}$. All observations are summarized in the co-occurrence matrix \mathbf{n} of counts $n(x, y)$, i.e. $n(x, y)$ denotes how often a feature y has been observed for a particular x . Effectively, this defines for each x an *empirical distribution* or *histogram* over \mathbf{Y} given by $\hat{P}(y|x) \equiv n(x, y)/n(x)$, where $n(x) \equiv \sum_{y \in \mathbf{Y}} n(x, y)$ denotes the number of observations for object x .

The proposed mixture model, which is referred to as *one-sided* or *Asymmetric Clustering Model (ACM)*, explains the observed data by a small number of multinomial component probability distributions over the feature space \mathbf{Y} . Introducing a latent structure $\mathbf{c} : x \mapsto \{c_1, \dots, c_K\}$ the generative model is defined as follows:

1. select an object $x \in \mathbf{X}$ with probability $P(x)$,
2. determine the latent class c according to the cluster membership $\mathbf{c}(x)$ of x ,
3. select $y \in \mathbf{Y}$ from the cluster-specific conditional distribution $P(y|c)$.

For notational convenience, the parameters $P(x)$ and $P(y|c)$ are summarized in a parameter vector θ and the discrete assignment variables $\mathbf{c}(x) \in \{1, \dots, K\}$ in a vector \mathbf{c} . According to the generative model, observations are assumed to be independent conditioned on the continuous parameters and the cluster assignments $\mathbf{c}(x)$. Hence, the probability to observe a dyad (x, y) is given by

$$P(x, y|\mathbf{c}, \theta) = P(x)P(y|\mathbf{c}(x)) \quad . \quad (1)$$

Taking a Bayesian perspective, we introduce prior distributions $P(\mathbf{c}, \theta)$ for all quantities which define the data generation process. By assuming prior independence of \mathbf{c} and θ and by putting a non-informative uniform prior on θ , the posterior distribution is – up to a proper normalization – given by

$$P(\mathbf{c}, \theta|\mathbf{n}) \propto P(\mathbf{c}) \prod_{x \in \mathbf{X}} \prod_{y \in \mathbf{Y}} (P(x)P(y|\mathbf{c}(x)))^{n(x, y)} \quad . \quad (2)$$

Returning to the texture segmentation problem, we identify the set of objects \mathbf{X} with the set of image

locations or sites and the set \mathbf{Y} with possible values of discrete or discretized local texture features computed from the image data. The distribution $\hat{P}(y|x)$ then represents a histogram of features occurring in an *image neighborhood* or *window* around some location x (cf. [2, 4]). Each class c corresponds to a different texture which is characterized by a specific distribution $P(y|c)$ of features y . Since these multinomial distributions are not constrained, they can virtually model any distribution of features. In particular, no further parametric restrictions on $P(y|c)$ are imposed. There is also no need to specify an additional noise model or, equivalently, a metric in feature space [7].

Parameter Estimation Taking (2) as a starting point we propose to compute *maximum a posteriori* estimates, i.e., we have to maximize the log-posterior distribution which is equivalent to maximizing

$$\mathbf{L}(\mathbf{c}, \theta; \mathbf{n}) = \sum_{x \in \mathbf{X}} n(x) \left[\sum_{y \in \mathbf{Y}} \hat{P}(y|x) \log P(y|\mathbf{c}(x)) + \log P(x) \right] + \log P(\mathbf{c}) \quad . \quad (3)$$

Stationary equations are derived from (3) by differentiation. Using Lagrange parameters to ensure a proper normalization of the continuous model parameters θ we obtain

$$\hat{P}(x) = \frac{n(x)}{\sum_{x' \in \mathbf{X}} n(x')} \quad , \quad (4)$$

$$\hat{P}(y|c) = \sum_{x: \hat{\mathbf{c}}(x)=c} \frac{n(x)}{\sum_{x': \hat{\mathbf{c}}(x')=c} n(x')} \hat{P}(y|x) \quad , \quad (5)$$

$$\hat{\mathbf{c}}(x) = \arg \min_a \left\{ - \sum_{y \in \mathbf{Y}} \hat{P}(y|x) \log \hat{P}(y|a) - \log P(\hat{\mathbf{c}}_a^x) \right\} \quad (6)$$

where $\hat{\mathbf{c}}_a^x(x) = a$ and $\hat{\mathbf{c}}_a^x(x') = \hat{\mathbf{c}}(x')$ for $x' \neq x$ denotes the class assignments obtained by changing the assignment of x to class a . From (4) we see that the probabilities $P(x)$ are estimated independently of all other parameters. The maximum a posteriori estimates of the class-conditional distributions $\hat{P}(y|c)$ are linear superpositions of all empirical distributions for objects x belonging to cluster c . Eq. (5) thus generalizes the *centroid condition* from K -means clustering to distributional clustering. Notice however, that the components of $\hat{P}(y|c)$ define probabilities for feature values and do not correspond to dimensions in the original feature space; eq. (5) averages over feature *distributions*, not over feature *values*. The formal similarity to K -means clustering is extended

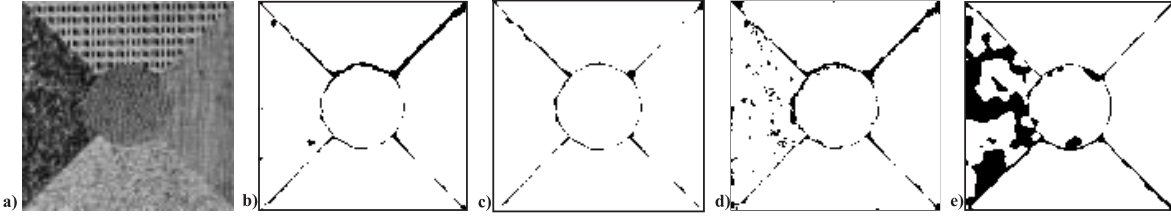


Figure 1: Typical segmentation results with $K = 5$ for the algorithms under examination: (a) original image, (b) ACM with uniform prior, (c) ACM with topological prior (14), (d) PDC and (e) K -means. Misclassified blocks w.r.t. ground truth are depicted in black.

by (6), which is the analogon to the nearest neighbor rule. In the maximum likelihood case, i.e., with a uniform prior $P(\mathbf{c})$, the optimal assignments for given parameters can be estimated in one step. In this case, the analogy between the stationary conditions for the ACM and for K -means clustering also holds for the model fitting algorithm. The likelihood can be maximized by an *alternating maximization* (AM) update scheme which calculates assignments for given centroids according to the nearest neighbor rule (6) and recalculates the centroid distributions (5) in alternation. Both algorithmic steps increase the likelihood. Thus convergence to a (local) maximum of (3) is ensured. For general $P(\mathbf{c})$, however, the assignments $\mathbf{c}(x)$ of different sites are coupled and (6) has to be iterated until convergence in analogy to the iterated conditional mode (ICM) algorithm.

Distributional Clustering In the maximum-likelihood case the ACM is similar to the distributional clustering model formulated in [7] as the minimization of the cost function

$$\mathbf{H}(\mathbf{c}, \theta; \mathbf{n}) = \sum_{x \in \mathbf{X}} D_{\text{KL}} \left[\hat{P}(\cdot|x) \parallel P(\cdot|\mathbf{c}(x)) \right]. \quad (7)$$

Here D_{KL} denotes the cross entropy or Kullback-Leibler (KL) divergence between the empirical and the model distribution over \mathbf{Y} . In distributional clustering, the KL-divergence as a distortion measure for distributions has been motivated by the fact that the centroid equation (5) is satisfied at stationary points. Yet, after dropping the $P(x)$ and $P(\mathbf{c})$ in (3) and a (data dependent) constant we derive the formula

$$\mathbf{L}(\mathbf{c}, \theta; \mathbf{n}) = - \sum_{x \in \mathbf{X}} n(x) D_{\text{KL}} \left[\hat{P}(\cdot|x) \parallel P(\cdot|\mathbf{c}(x)) \right]. \quad (8)$$

This proves that the choice of the KL-divergence as a distortion measure simply follows from the likelihood principle.

Deterministic Annealing *Deterministic annealing* (DA) is an optimization technique which allows us to improve the presented AM procedure by avoiding unfavorable local minima. The key idea is to introduce a temperature parameter T and to replace the minimization of a combinatorial objective function by a substitute known as the *generalized* or *variational free energy*. Details on this topic in the context of data clustering can be found in [9, 7, 4]. Minimization of the free energy corresponding to (8) yields the following equations for probabilistic assignments:

$$P(\mathbf{c}(x) = a | \hat{\theta}) = \frac{\exp(-h(a, x; \hat{\theta})/T)}{\sum_{b=1}^K \exp(-h(b, x; \hat{\theta})/T)}, \quad (9)$$

$$h(a, x; \hat{\theta}) = n(x) D_{\text{KL}} \left[\hat{P}(\cdot|x) \parallel \hat{P}(\cdot|a) \right]. \quad (10)$$

This partition of unity is an intuitive generalization of the nearest neighbor rule in (6). For $T \rightarrow 0$ the argmin operation performed in the nearest neighbor rule is recovered. Since solutions in DA are tracked from high to low temperatures, we finally maximize the log-likelihood with respect to both, θ and \mathbf{c} , at $T = 0$. Notice, that the DA procedure also generalizes the Expectation Maximization (EM) algorithm obtained for $T = 1$, in which \mathbf{c} is treated as an unobserved variable. In the latter case (9) corresponds to the computation of posterior probabilities in the E-step. For the general case in (3), a more complex coupled system of transcendental equations for the probabilistic assignments is obtained. This system of equations is solved by a convergent iterative scheme [4]. As an additional advantage, it has been demonstrated in [1], that deterministic annealing with finite stopping temperature $T > 0$ can be utilized to efficiently avoid data over-fitting.

Multiscale Annealing It is a natural assumption, that adjacent image sites contain identical textures with high probability. This fact can be exploited to significantly accelerate the optimization of the likeli-

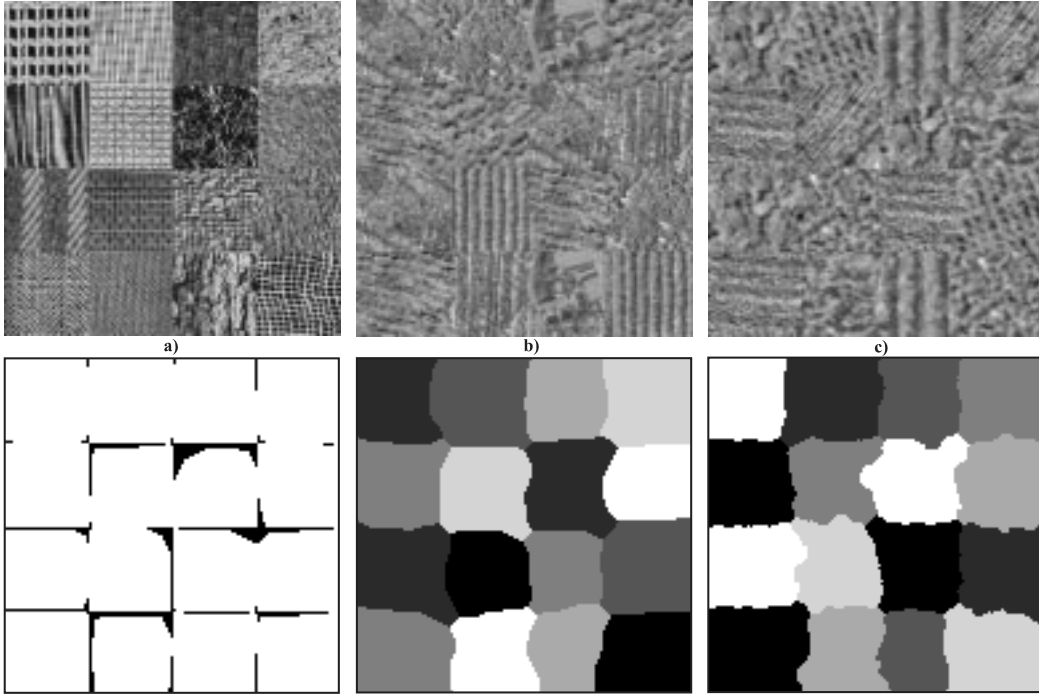


Figure 2: Typical segmentation results obtained by ACM with topological prior: (a) on a mondrian of 16 different Brodatz textures, (b) and (c) mondrians of 7 different textures taken from aerial images (ground truth unavailable).

hood by maximizing over a suitable nested sequence of subspaces in a coarse-to-fine manner, where each subspace has a greatly reduced number of class assignment variables. This strategy is formalized by the concept of *multiscale optimization* [3, 8] which in essence leads to cost functions redefined on a coarse version of the original image. In contrast to most multi-resolution optimization schemes, the *original* cost function is optimized at all grids, only the configuration space is reduced by variable tying. We first sketch the general theory and then derive multiscale equations for histogram clustering.

Formally, we denote the original set of sites by $\mathbf{X}^0 = \mathbf{X}$ and assume sets of sites $\mathbf{X}^l = \{x_1^l, \dots, x_{N^l}^l\}$ at coarse grid levels l . Typically, $\mathbf{X}^0 = \mathbf{X}$ corresponds to the set of pixel sites and \mathbf{X}^{l+1} is obtained by subsampling \mathbf{X}^l by a factor of 2 in each direction. A coarsening map γ_l links each fine grid point to a single coarse grid point,

$$\gamma_l : \mathbf{X}^l \rightarrow \mathbf{X}^{l+1}, \quad x^l \mapsto x^{l+1} = \gamma_l(x^l). \quad (11)$$

The inverse map γ_l^{-1} defines a subset of the fine grid sites, $\gamma_l^{-1}(x^l) \subset \mathbf{X}^l$. Multiscale optimization proceeds not by coarsening the image, but by *coarsening the variable space*. Each coarse grid is associated with a reduced set of assignment variables \mathbf{c}^l . Thus, a single variable $c^l(x^l)$ is attached to each grid point x^l coding

the texture class of the set of respective pixels. Coarsened cost functions at level $l+1$ are defined by proper restriction of the optimization space at level l . Restricting the consideration for notational convenience to the simplified model (8), the following coarse grid log-likelihood functions are obtained:

$$\mathbf{L}^l(\mathbf{c}^l, \theta; \mathbf{n}) = \sum_{x^l \in \mathbf{X}^l} \sum_{y \in \mathbf{Y}} n^l(x^l, y) \log P(y | \mathbf{c}^l(x^l)), \quad (12)$$

with the recursive definition

$$n^{l+1}(x^{l+1}, y) = \sum_{x^l \in \gamma_l^{-1}(x^{l+1})} n^l(x^l, y). \quad (13)$$

Note, that \mathbf{L}^l has the same functional form as $\mathbf{L}^0 = \mathbf{L}$ and, therefore, an optimization algorithm developed for \mathbf{L} is applicable to any coarse grid cost function \mathbf{L}^l .

Traditionally, clustering algorithms like K -means are used in conjunction with *splitting techniques* to obtain successive solutions for a growing number of clusters. We adopt this idea by successively splitting clusters with high distortion. Since the number of data objects is drastically reduced at coarser resolution levels, the splitting strategy and the coarse-to-fine optimization are interleaved. The question of choosing the maximal number of clusters for a given resolution has been addressed in a statistical learning theory context in [1]. We adopt these results by choosing

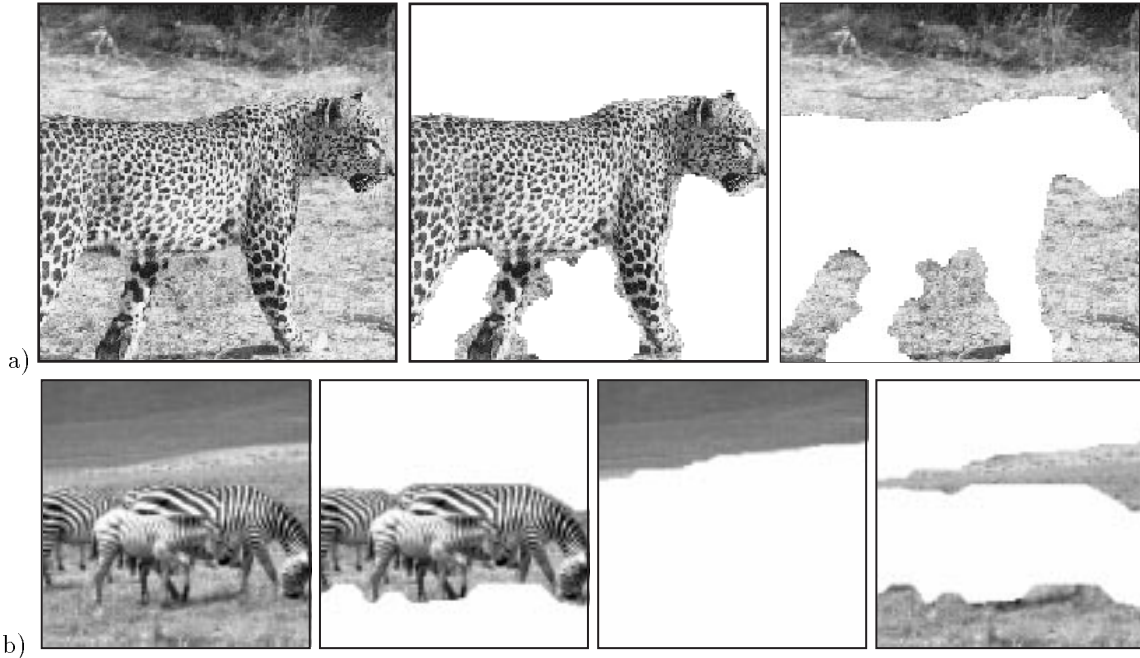


Figure 3: Typical segmentation result on real-world images with (a) $K = 2$ and (b) $K = 3$ segments obtained by ACM with topological prior.

$K_{\max}^l \sim N^l / \log N^l$ and selecting the proportionality factor on an empirical basis.

One of the key advantages of the DA approach is the *inherent splitting behaviour*. Clusters degenerate at high temperature and they successively split at *bifurcations* or *phase transitions* when T is lowered [9]. Therefore, at an *effective number* of $K_T \leq K$ of clusters is distinguishable at each temperature T . For a given resolution level l , we anneal until K_T exceeds the predefined maximal number of clusters K_{\max}^l at a certain temperature level T^* . After prolongation to level $l - 1$, the optimization is continued at temperature T^* . This scheme is known as *multiscale annealing* [8].

3 Results

Implementation Details We applied the asymmetric clustering model (ACM) to the unsupervised segmentation of textured images. Since the number of observed features is identical for all image sites x , one can simply set $P(x) = 1/N$. In the experiments, we have adopted the framework of [5, 4, 8] and utilized an image representation based on the modulus of complex Gabor filters. For each site the empirical distribution of coefficients in a surrounding (filter-specific) window is determined. All reported segmentation results are based on a filter bank of twelve Gabor filters with four orientations and three scales. Each filter output was discretized into 16 equally sized bins resulting in

a feature space \mathbf{Y} of size $M = 192$. For each channel Gabor coefficients were sampled in a local window of a size proportional to the scale of the filter was utilized (a 16×16 window at the finest scale). The approach taken in [5, 4, 8] is based upon the marginal filter distribution and neglects statistical dependencies between different Gabor channels. The main reason for the restriction to marginals is the notorious difficulty to estimate multi-dimensional densities from a limited amount of data. But the statistical correlation of filter responses carries important information which might be necessary to discriminate certain types of textures. We, therefore, propose to use *adaptive multivariate histograms* as an alternative approach, where an image-specific binning is estimated. The K -means algorithm was utilized to identify a representative set of prototypes ($K = 64$ bins in the experiments). The adaptive histogram binning is defined as the corresponding Voronoi tessellation. A histogram entry at a site is given as the number of data points in a local window of size 32×32 with a vector of Gabor responses falling in the corresponding Voronoi cell.

Exploiting prior knowledge, segmentation results can be improved by suppressing small and highly fragmented regions. As a quality criterion, we propose to count for each image site, how many sites of the same class label are found in a small topological neighborhood. If the number of identically labeled pixels drops

	Median	20% quantile	Run-time
ACM with uniform prior	5.1%	5%	3.5s
ACM with topological prior (14)	2.8%	5%	27.4s
ACM on multivariate histograms	3.7%	1%	5.4s
Pairwise Clustering	5.8%	8%	3.6s
Normalized Cut	5.9%	8%	2.5s
K -means Clustering	6.9%	6%	0.94s

Table 1: Errors by comparison with ground truth over 100 randomly generated images with $K = 5$ textures, 512×512 pixels and 128×128 sites. For all algorithms multiscale annealing has been used for optimization. The median run-time over 100 images has been measured on a Pentium II 300 MHz.

below a threshold, the label configuration is considered to be less likely, the log-likelihood being proportional to the difference to the threshold. The probability distribution

$$P(\mathbf{c}) = \frac{1}{Z} \exp(-\lambda H^p(\mathbf{c})) \quad , \quad (14)$$

$$H^p(\mathbf{c}) = \sum_{x \in \mathbf{X}} (B - A(x))_+ \quad (15)$$

makes this considerations mathematically precise, where B denotes a threshold, $\mathbf{N}(x)$ denotes the topological neighborhood of site x , $A(x) = |\{x' \in \mathbf{N}(x) : \mathbf{c}(x') = \mathbf{c}(x)\}|$ denotes the number of neighboring sites assigned to the same class, λ defines a weighting parameter. In the experiments, we used a large value for λ to suppress small regions completely where small regions have been defined via $B = 10$ and a 7×7 topological neighborhood. It is worth to note, that for this topological prior, multiscale techniques are mandatory for optimization as (14) effectively erects barriers in the search space which can not be traversed solely by single-site changes.

The benchmark results are obtained on images which were generated from a representative set of 86 micro-patterns taken from the Brodatz texture album. A database of random mixtures (512×512 pixels each) containing 100 entities of five textures each (as depicted in Fig. 1) was constructed. For comparison, we utilized K -means and two proximity-based clustering algorithms. For the K -means algorithm a spatial smoothing step was applied to the Gabor coefficients before clustering (cf. [5]). The *pairwise dissimilarity clustering* algorithm (PDC) is based on a normalized cost function, which is invariant to linear transformation of the proximity data [4]. The *normalized cut* has been proposed only for $K = 2$ [11]. The corresponding normalized cost function is equivalent to the normalized association which generalizes to $K > 2$ and which has been used in the experiments. The χ^2 test statistic applied to

the Gabor channel histograms was utilized to obtain proximity data, where we have computed approximately 80 randomly selected dissimilarity scores for each image site [11, 4]. The multiscale annealing technique was utilized for all clustering cost functions.

Segmentation Results The question examined in detail is concerned with the benefits of the ACM in comparison to other clustering schemes. A typical example with $K = 5$ clusters is given in Fig. 1. The error plots demonstrate, that the segmentations achieved by ACM have the highest quality. Most errors occur at texture boundaries where texture information is mixed due to the spatial support of the Gabor filters and the spatial extent of the computed local feature statistics. The K -means clustering cost function exhibits substantial deficits to correctly model the segmentation task. These observations are confirmed by the benchmark results in Tab. 1. We report the median, since the distributions of the empirical errors are highly asymmetric. In addition, the percentage of segmentations with an error rate larger than 20% is given, which we define as the percentage of segmentations where the essential structure has not been detected. For the ACM a median error of 5.1% has been achieved compared to 5.7% for PDC and 5.8% for the normalized cut. Moreover, using the topological prior (14) further improves the segmentation results, which leads to a median error as low as 2.8%. The K -means model yields significantly worse results with a median error of 6.9%. The percentage of segmentations, where the essential structure has been detected, is highest for the ACM with 95%. This is further improved using multi-dimensional adaptive histograms, where the essential structure has been detected in all but one case. While the optimization time for multivariate histograms is lower due to the smaller number of bins used, one has to keep in mind that the data extraction process is substantially more time-consuming. The excellent segmentation quality

	K -means	ACM	PDC / NC
Underlying data type	vector	histogram	proximity
Computational complexity of data extraction	lowest	medium	highest
Computational complexity of optimization	lowest	medium	medium
Segmentation quality	lowest	highest	medium
Generative statistical model provided	yes	yes	no
Implementation effort	low	low	high

Table 2: Advantages and disadvantages of the clustering algorithms: K -means, histogram clustering, pairwise dissimilarity clustering/normalized cut.

obtained by the ACM histogram clustering algorithm is confirmed by the results on more difficult segmentation tasks in Fig. 2. The mixture of $K = 16$ different Brodatz textures has been partitioned accurately with an error rate of 4.7%. The errors basically correspond to boundary sites. The results obtained for the mondrians of aerial images are satisfactory but due to missing ground truth the quality could not be quantified. Disconnected texture regions of the same type have been identified correctly, while problems again occur at texture boundaries. The segmentation quality achieved on outdoor images in Fig. 3 are both visually and semantically satisfying.

4 Conclusion

The ACM histogram clustering model combines the expressive power of pairwise dissimilarity clustering with the efficiency of the conventional K -means clustering and provides a fast, accurate and reliable algorithm for unsupervised texture segmentation. Compared to PDC and the normalized cut the time-consuming algorithmic step of computing pairwise dissimilarities between objects has been avoided. Yet, the benchmark results indicate that the ACM provides improved segmentation quality, especially when combined with a topological prior. Moreover, statistical group information is provided for subsequent processing steps making the ACM a highly interesting alternative to PDC and the normalized cut in the segmentation context. The advantages and disadvantages of all three clustering methods are summarized in Tab. 2. As a general clustering scheme, this model can be extended to color and motion segmentation, region grouping and even integrated sensor segmentation simply by choosing appropriate features.

Acknowledgments It is a pleasure to thank Hans du Buf for providing the aerial image mixtures in Fig. 2. This work has been supported by the German Research Foundation (DFG) under grant #BU 914/3-1, by a M.I.T. Faculty Sponser’s Discretionary Fund and

by an ICSI Postdoctoral Fellowship.

References

- [1] J. Buhmann and J. Puzicha. Unsupervised learning for robust texture segmentation. In *Proc. Dagstuhl Seminar on Empirical Evaluation of Computer Vision Algorithms (to appear)*, 1999.
- [2] D. Geman, S. Geman, C. Graffigne, and P. Dong. Boundary detection by constrained optimization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(7):609–628, 1990.
- [3] F. Heitz, P. Perez, and P. Boutheimy. Multiscale minimization of global energy functions in some visual recovery problems. *CVGIP: Image Understanding*, 59(1):125–134, 1994.
- [4] T. Hofmann, J. Puzicha, and J. Buhmann. Unsupervised texture segmentation in a deterministic annealing framework. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(8):803–818, 1998.
- [5] A. Jain and F. Farrokhnia. Unsupervised texture segmentation using Gabor filters. *Pattern Recognition*, 24(12):1167–1186, 1991.
- [6] J. Mao and A. Jain. Texture classification and segmentation using multiresolution simultaneous autoregressive models. *Pattern Recognition*, 25:173–188, 1992.
- [7] F. Pereira, N. Tishby, and L. Lee. Distributional clustering of English words. In *30th Annual Meeting of the Association for Computational Linguistics, Columbus, Ohio*, pages 183–190, 1993.
- [8] J. Puzicha and J. Buhmann. Multi-scale annealing for real-time unsupervised texture segmentation. In *Proc. International Conference on Computer Vision (ICCV’98)*, pages 267–273, 1998.
- [9] K. Rose, E. Gurewitz, and G. Fox. A deterministic annealing approach to clustering. *Pattern Recognition Letters*, 11:589–594, 1990.
- [10] P. Schroeter and J. Bigun. Hierarchical image segmentation by multi-dimensional clustering and orientation-adaptive boundary refinement. *Pattern Recognition*, 28(5):695–709, 1995.
- [11] J. Shi and J. Malik. Normalized cuts and image segmentation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR’97)*, pages 731–737, 1997.