

Discrete Mixture Models for Unsupervised Image Segmentation^{*}

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

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

^{*} Artificial Intelligence Laboratory
Massachusetts Institute of Technology
hofmann@ai.mit.edu

Abstract. This paper introduces a novel statistical mixture model for probabilistic clustering of histogram data and, more generally, for the analysis of discrete co-occurrence data. Adopting the maximum likelihood framework, an alternating maximization algorithm is derived which is combined with annealing techniques to overcome the inherent locality of alternating optimization schemes. We demonstrate an application of this method to the unsupervised segmentation of textured images based on local empirical distributions of Gabor coefficients. In order to accelerate the optimization process an efficient multiscale formulation is utilized. We present benchmark results on a representative set of Brodatz mondrians and real-world images.

1 Introduction

Grouping of homogeneous image regions is an important task in low-level computer vision that is widely pursued to solve the problem of image segmentation, in particular in the context of textured images. Two steps have to be considered in order to address this problem:

- Most fundamentally, a mathematical notion of homogeneity or similarity between image regions is required in order to formalize the segmentation problem. Especially for textured images the similarity measure has to capture the significant variability within a texture, without losing the ability to discriminate between different textures.
- In a second step, after a similarity measure is defined, an efficient algorithm for partitioning or clustering has to be derived to solve the computational problem. The selection of a suitable clustering method of course is tightly coupled to the chosen similarity measure.

A successful approach has to rely on a similarity measure which is powerful enough to discriminate a wide range of textures, while preserving the computational tractability of the overall segmentation algorithm.

^{*} 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 and by a M.I.T. Faculty Sponser's Discretionary Fund. A more detailed report is found in [3].

Numerous approaches to unsupervised texture segmentation have been proposed over the past decades. In the classical approaches, locally extracted features are spatially smoothed and interpreted as *vectors in a metric space* [5, 6], thereby characterizing each texture by a specific average feature vector or *centroid*. The most commonly used distortion measure is the (weighted) squared Euclidean norm which effectively models the data by a Gaussian mixture model, where each Gaussian represents exactly one texture. The method of choice for clustering vectorial data is the K -means algorithm and its variants, which have been exploited for texture segmentation in [5, 6].

Since the Gaussian mixture assumption turns out to be inadequate in many cases, several alternative approaches have utilized *pairwise proximity data*, usually obtained by applying statistical tests to the local feature distribution at two image sites [1, 4, 7]. As a major advantage, these methods do not require the specification of a suitable vector-space metric. Instead, similarity is defined by the *similarity of the respective feature distributions*. For pairwise similarity data agglomerative clustering [7] and, more rigorously, optimization approaches to graph partitioning [1, 4, 12] have been proposed in the texture segmentation context, which we refer to as *pairwise dissimilarity clustering* (PDC). Although these methods are directly applicable to proximity data, they are only tractable in image segmentation problems if they avoid the computation of dissimilarities for all possible pairs of sites [9].

The major contribution of this paper is a general approach to the problem of grouping *feature distributions*, extending a technique known as distributional clustering in statistical language modeling [8]. In contrast to methods based on feature vectors and pairwise dissimilarities this approach is directly applicable to histogram data and empirical distributions. In comparison to K -means clustering, distributional clustering naturally includes component distributions with multiple modes rather than fitting segments with a univariate Gaussian mode. As a major advantage compared to PDC it requires no external similarity measure, but exclusively relies on the *feature occurrence statistics*. Another important consideration for a clustering approach to image segmentation are real-time constraints. Given the respective data (vectors, histograms or proximities) all algorithms require only a few seconds for optimization [9]. While vector-based methods suffer from inferior quality, it is the data extraction process of PDC which is prohibitive for real-time applications like autonomous robotics. Using the histogram data directly thus avoids the necessity for pairwise comparisons altogether while achieving segmentations of similar quality compared to PDC. In addition, the proposed mixture model provides a generative statistical model for the observed features by defining a *texture specific distribution*. This can be utilized in subsequent processing steps such as boundary localization [11].

2 Mixture Models for Histogram Data

To stress the generality of the proposed model we temporarily detach the presentation from the specific problem of image segmentation. Consider therefore

the following more abstract setting: Denote by $\mathcal{X} = \{x_1, \dots, x_N\}$ a finite set of abstract objects with arbitrary labeling and by $\mathcal{Y} = \{y_1, \dots, y_M\}$ the domain of some nominal scale feature(s). Each $x_i \in \mathcal{X}$ is characterized by a number of observations (x_i, y_j) summarized in the sufficient statistics of counts n_{ij} . Effectively, this defines for each x_i an *empirical distribution* or *histogram* over \mathcal{Y} defined by $n_{j|i} \equiv n_{ij}/n_i$ where $n_i \equiv \sum_j n_{ij}$.

In the context of image segmentation we identify the set of objects \mathcal{X} with the set of image locations or sites and the set \mathcal{Y} with possible values of discrete or discretized texture features computed from the image data. The distributions $n_{j|i}$ then represent a histogram of features occurring in an *image neighborhood* or *window* around some location x_i [1, 4, 7]. The framework is, however, general enough to cover distinctive application domains like information retrieval [3] and natural language modeling [8].

The proposed mixture model, which is referred to as Asymmetric Clustering Model (ACM)¹, explains the observed data by a finite number of component probability distributions on the feature space. The generative model is defined as follows:

1. select an object $x_i \in \mathcal{X}$ with probability p_i ,
2. choose the component (cluster) \mathcal{C}_α according to the cluster membership of x_i ,
3. select $y_j \in \mathcal{Y}$ from a cluster-specific conditional distribution $q_{j|\alpha}$.

In addition to the parameters $\mathbf{p} = (p_i)$ and $\mathbf{q} = (q_{j|\alpha})$ let us introduce indicator variables $M_{i\alpha} \in \{0, 1\}$ for the class membership of x_i ($\sum_\alpha M_{i\alpha} = 1$). The probability of a feature occurrence (x_i, y_j) is then given by

$$P(x_i, y_j | \mathbf{M}, \mathbf{p}, \mathbf{q}) = p_i \sum_{\alpha=1}^K M_{i\alpha} q_{j|\alpha} . \quad (1)$$

The ACM has $K(M-1)$ continuous parameters for the component densities $q_{j|\alpha}$, $N-1$ parameters for the probabilities p_i , and N sets of indicator functions encoding an one-out-of- K choice each. An essential assumption behind the ACM is that observations for x_i are conditionally independent given the continuous parameters and the cluster assignment ($M_{i\alpha}$) of x_i .

Returning to the texture segmentation problem, we see that each class \mathcal{C}_α corresponds to a different texture which is characterized by a specific distribution $q_{j|\alpha}$ of features y_j . Since these component distributions of the mixture model are not constrained, they can virtually model any distribution of features. In particular, no further parametric restrictions on $q_{j|\alpha}$ are imposed. There is also no need to specify an additional noise model or, equivalently, a metric in the feature space.

¹ It is called asymmetric because clustering structure is inferred solely in the \mathcal{X} -space. The ACM is the most suitable model for image segmentation out of a family of novel mixture models developed for general co-occurrence data [3].

3 Maximum Likelihood Estimation for the ACM

To fit the model specified by (1) we apply the maximum likelihood principle and determine the parameter values with the highest probability to generate the observed data. We start with the log-likelihood function which is given by

$$\mathcal{L} = \sum_{i=1}^N \sum_{j=1}^M n_{ij} \sum_{\alpha=1}^K M_{i\alpha} \log q_{j|\alpha} + \sum_{i=1}^N n_i \log p_i . \quad (2)$$

Maximum likelihood equations are derived from (2) by differentiation, using Lagrange parameters to ensure a proper normalization of the continuous model parameters \mathbf{p} and \mathbf{q} . The resulting stationary equations are given by

$$\hat{p}_i = \frac{n_i}{\sum_k n_k}, \quad (3)$$

$$\hat{q}_{j|\alpha} = \frac{\sum_{i=1}^N \hat{M}_{i\alpha} n_{ij}}{\sum_{i=1}^N \hat{M}_{i\alpha} n_i} = \sum_{i=1}^N \frac{\hat{M}_{i\alpha} n_i}{\sum_{k=1}^N \hat{M}_{k\alpha} n_k} n_{j|i}, \quad (4)$$

$$\hat{M}_{i\alpha} = \begin{cases} 1 & \text{if } \alpha = \arg \min_{\nu} \left\{ -\sum_{j=1}^M n_{j|i} \log \hat{q}_{j|\nu} \right\} \\ 0 & \text{else} \end{cases} \quad (5)$$

From (3) we see that the probabilities p_i are estimated independently of all other parameters. The maximum likelihood estimates of the class-conditional distributions $\hat{q}_{j|\alpha}$ are linear superpositions of all empirical distributions for objects x_i belonging to cluster \mathcal{C}_α . Eq. (4) thus generalizes the *centroid condition* from K -means clustering. Notice however, that the components of $\hat{q}_{j|\alpha}$ define probabilities for feature values and do not correspond to dimensions in the original feature space. Eq. (4) averages over feature *distributions*, not over feature *values*. The formal similarity to K -means clustering is extended by (5), which is the analogon to the nearest neighbor rule.

The ACM is similar to the distributional clustering model formulated in [8] as the minimization of the cost function

$$\mathcal{H} = \sum_{i=1}^N \sum_{\alpha=1}^K M_{i\alpha} D[n_{j|i}|q_{j|\alpha}] . \quad (6)$$

Here D denotes the cross entropy or Kullback-Leibler (KL) divergence. In distributional clustering, the KL-divergence as a distortion measure for distributions has been motivated by the fact that the centroid equation (4) is satisfied at stationary points². Yet, after dropping the p_i parameters in (2) and a (data dependent) constant we derive the formula

$$\mathcal{L} = - \sum_{i=1}^N n_i \sum_{\alpha=1}^K M_{i\alpha} D[n_{j|i}|q_{j|\alpha}] . \quad (7)$$

² Note that this is not a unique property of the KL-divergence as it is also satisfied for the Euclidean distance.

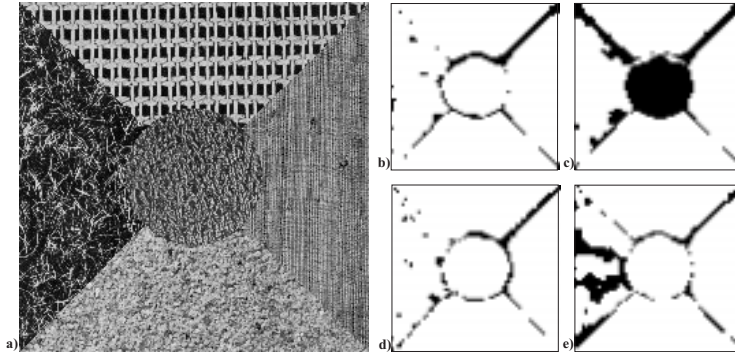


Fig. 1. Typical segmentation results with $K = 5$ for the algorithms under examination: (a) original image, (b) annealed ACM, (c) AM for ACM, (d) PDC and (e) K -means. Misclassified blocks w.r.t. ground truth are depicted in black.

This proves that the choice of the KL-divergence as a distortion measure simply follows from the likelihood principle. 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 (5) and recalculates the centroid distributions (4) in alternation. Both algorithmic steps increase the likelihood and convergence to a (local) maximum of (7) is thus ensured.

A technique which allows us to improve the presented AM procedure by avoiding unfavorable local minima is known as *deterministic annealing* (DA). 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 free energy*. Details on this topic in the context of data clustering can be found in [10, 8, 4]. Minimization of the free energy corresponding to (7) yields the following equations for probabilistic assignments:

$$\mathbf{P}(M_{i\alpha} = 1 | \mathbf{p}, \mathbf{q}) = \frac{\exp(-n_i D[n_{j|i} | \hat{q}_{j|\alpha}] / T)}{\sum_{\nu=1}^K \exp(-n_i D[n_{j|i} | \hat{q}_{j|\nu}] / T)}. \quad (8)$$

This partition of unity is a very intuitive generalization of the nearest neighbor rule in (5). For $T \rightarrow 0$ the arg-min operation performed in the nearest neighbor rule is recovered. Since in DA solutions are tracked from high to low temperatures, we finally maximize the log-likelihood at $T = 0$. Notice that the DA procedure also generalizes the Expectation Maximization (EM) algorithm which is obtained for $T = 1$. In this case (8) corresponds to the computation of posterior probabilities for hidden variables $M_{i\alpha}$ in the E-step.

4 Unsupervised Segmentation of Textured Images

It is a natural assumption that adjacent image sites contain identical texture with high probability. This fact can be exploited to significantly accelerate the

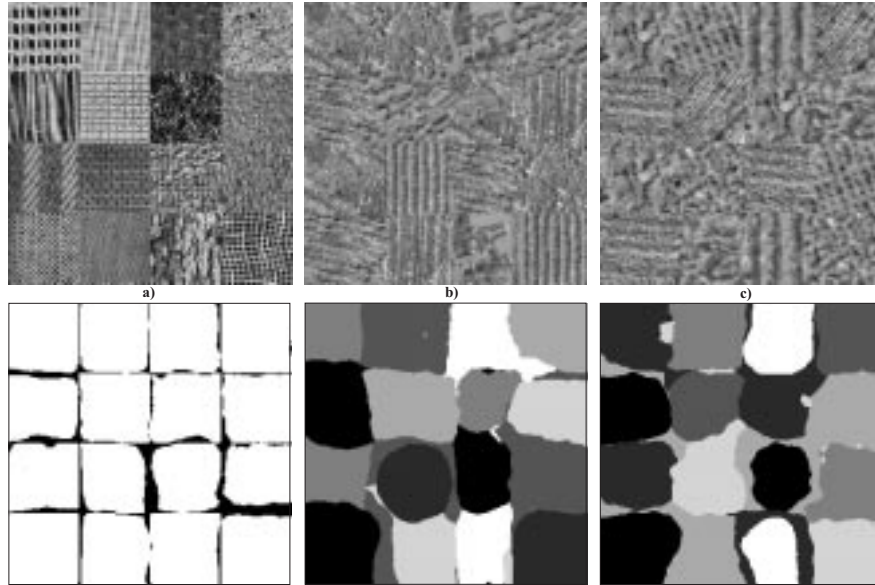


Fig. 2. Typical segmentation results: (a) on a mondrian of 16 different Brodatz textures (misclassified blocks w.r.t. ground truth are depicted in black), (b) and (c) mondrians of 7 different textures taken from aerial images (no ground truth available).

optimization of the likelihood by maximizing over a suitable nested sequence of subspaces in a coarse-to-fine manner, where each of these subspaces is spanned by a greatly reduced number of indicator variables. This strategy is formalized by the concept of *multiscale optimization* [2] and it essentially leads to cost functions redefined on a coarsened version of the original image. In contrast to most multiresolution optimization schemes the *original* log-likelihood is optimized at all grids, only the variable configuration space is reduced. For the ACM log-likelihood cost-functions of identical algebraic structure are obtained at all levels. Deterministic annealing and multiscale optimization are combined in the concept of multiscale annealing. The resulting algorithm provides an acceleration factor of 5 – 500 compared to single scale optimization. For details we refer to [9, 3].

Algorithms based on distributions of features have been successfully used in texture analysis [1, 7, 4]. In the experiments we have followed [5, 4, 9] 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 segmentations are based on a filter bank of twelve Gabor filters with four orientations and three scales. 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 Fig. 1) was constructed. For the K -means algorithm a spatial smoothing step was applied before clustering as described in [5]. For all

	Median	20% quantile
AM for ACM	8.9%	18%
annealed ACM	6.7%	6%
annealed PDC	6.0%	6%
annealed K -means	11.7%	28%

Table 1. Errors by comparison with ground truth over 100 randomly generated images with $K = 5$ textures. 512x512 pixels and 64x64 assignments.

cost functions the multiscale annealing optimization scheme was applied, where coarse grids up to a resolution of 8x8 grid points have been used.

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. It can be seen that the segmentations achieved by ACM and PDC are highly similar. Most errors occur at texture boundaries as expected, where texture information is mixed due to the spatial support of the Gabor filters and the extent of the neighborhood used for computing the 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 reported, which we define as the percentage of segmentations where the essential structure is not detected. ACM and PDC yield similar results with a statistically insignificant difference. For the ACM a median error of 6.7% was achieved compared to 6.0% for PDC. The percentage of segmentations, where the essential structure has been detected, is in both cases as high as 94%. The K -means model yields significantly worse results with a median error of 11.7%. Moreover, in 28% of the cases the essential structure was not detected. The quality of the ACM model 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 7.9%. The errors basically correspond to boundary sites. The results obtained for the mondrians of aerial images are satisfactory. Disconnected texture regions of the same type have been identified correctly, while problems again occur at texture boundaries.

We conclude, that (annealed) ACM combines the expressive power of pairwise similarity 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 the time-consuming algorithmic step of computing pairwise dissimilarities between objects has been avoided. Moreover, statistical group information is provided for subsequent processing steps making the ACM an interesting alternative to PDC in the segmentation context. The advantages and disadvantages of all three clustering methods are summarized in Tab. 2. It has been confirmed, that global optimization algorithms like multiscale

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

Table 2. Advantages and disadvantages of the clustering algorithms.

annealing are essential for reliable computation of high quality segmentations. As a general clustering scheme this model can be extended to color and motion segmentation, region grouping and even to combinations of these simply by choosing appropriate features.

References

1. D. Geman, S. Geman, C. Graffigne, and P. Dong. Boundary detection by constrained optimization. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 12(7):609–628, 1990.
2. F. Heitz, P. Perez, and P. Bouthemy. Multiscale minimization of global energy functions in some visual recovery problems. *Computer Vision and Image Understanding*, 59(1):125–134, 1994.
3. T. Hofmann and J. Puzicha. Statistical models for co-occurrence data. AI-Memo 1625, MIT, 1998.
4. T. Hofmann, J. Puzicha, and J. Buhmann. Deterministic annealing for unsupervised texture segmentation. In *Proc. EMMCVPR'97, LNCS 1223*, pages 213–228, 1997.
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. T. Ojala and M. Pietikäinen. Unsupervised texture segmentation using feature distributions. Tech. Rep. CAR-TR-837, Center for Robotics and Automation, University Maryland, 1996.
8. F. Pereira, N. Tishby, and L. Lee. Distributional clustering of English words. In *Proc. Association for Computational Linguistics*, pages 181–190, 1993.
9. 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.
10. K. Rose, E. Gurewitz, and G. Fox. A deterministic annealing approach to clustering. *Pattern Recognition Letters*, 11:589–594, 1990.
11. P. Schroeter and J. Bigun. Hierarchical image segmentation by multi-dimensional clustering and orientation-adaptive boundary refinement. *Pattern Recognition*, 28(5):695–709, 1995.
12. J. Shi and J. Malik. Normalized cuts and image segmentation. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR'97)*, pages 731–737, 1997.