Parametric Distributional Clustering for Image Segmentation

Lothar Hermes, Thomas Zöller, and Joachim M. Buhmann

Rheinische Friedrich Wilhelms Universität
Institut für Informatik III, Römerstr. 164
D-53117 Bonn, Germany {hermes, zoeller, jb}@cs.uni-bonn.de,
http://www-dbv.informatik.uni-bonn.de

Abstract. Unsupervised Image Segmentation is one of the central issues in Computer Vision. From the viewpoint of exploratory data analysis, segmentation can be formulated as a clustering problem in which pixels or small image patches are grouped together based on local feature information. In this contribution, parametrical distributional clustering (PDC) is presented as a novel approach to image segmentation. In contrast to noise sensitive point measurements, local distributions of image features provide a statistically robust description of the local image properties. The segmentation technique is formulated as a generative model in the maximum likelihood framework. Moreover, there exists an insightful connection to the novel information theoretic concept of the *Information Bottleneck* (Tishby et al. [17]), which emphasizes the compromise between efficient coding of an image and preservation of characteristic information in the measured feature distributions.
The search for good grouping solutions is posed as an optimization problem, which is solved by *deterministic annealing* techniques. In order to further increase the computational efficiency of the resulting segmentation algorithm, a multi-scale optimization scheme is developed. Finally, the performance of the novel model is demonstrated by segmentation of color images from the Corel data base.

Keywords: Image Segmentation, Clustering, Maximum Likelihood, Information Theory

1 Introduction

Image understanding and visual object recognition crucially rely on image segmentation as an intermediate level representation of image content. Approaches to image segmentation which lack supervision information are often formulated as data clustering problems. Regardless of the particular nature of the image primitives in question, these methods share as a common trait that they search for a partition of pixels or pixel blocks with a high degree of homogeneity. The specific choice of a clustering algorithm, however, is dependent on the nature of the given image primitives which might be feature vectors, feature relations or feature histograms. In this paper, we advocate to characterize an image site

A. Heyden et al. (Eds.): ECCV 2002, LNCS 2352, pp. 577–591, 2002.

by the empirical color distributions extracted from its neighborhood, which we regard as a robust and statistically reliable descriptor of local color properties.

One way to design a clustering technique for this type of data is to apply a statistical test to the measured histograms. This processing step yields pairwise dissimilarity values, for which a multitude of grouping techniques can be found in the literature (e.g. [7,13,16]). Alternatively, feature histograms can be grouped directly by histogram clustering [10,12]. The histogram clustering approach characterizes each cluster by a prototypical feature distribution, and it assigns feature histograms to the nearest prototype distribution. Closeness is measured by the *Kullback–Leibler–Divergence*. As a consequence, this method retains the efficiency of central clustering approaches like k-means clustering, but it avoids the restrictive assumption that features are vectors in a Euclidean space.

Histogram clustering in its original form is invariant to permutations of histogram bins. In computer vision where the histogramming process is prone to noise induced errors, this invariance neglects information about the order of bins and the distance of bin centers in feature space. We, therefore, suggest to replace the non-parametric density estimation via histograms by a continuous mixture model, which no longer suffers from the shortcomings of a non-adaptive discrete binning process. The resulting statistical model can be interpreted as a generative model for pixel colors, but we can also establish an interesting connection to the novel information theoretic concept of the *Information Bottleneck* principle [17].

The search for a good grouping solution is posed as a combinatorial optimization problem. Due to the fact that the cost landscape may have a very jagged structure, powerful optimization techniques with regularization or smoothing behavior should be applied to avoid poor local minima. We use deterministic annealing which is embedded in a multi-scale framework for additional computational efficiency. To give an impression of the performance of the new parametric distributional clustering algorithm, color segmentations of pictures taken from the Corel gallery are shown in the results section.

2 The Clustering Model

Notation & Model Definition: To stress the generality of the proposed clustering model, the discussion of the cost function is initially detached from the application domain of image segmentation. Assume a set of objects $\mathbf{o}_i, i = 1, \ldots, n$ to be given. These entities are supposed to be clustered in k groups. The cluster memberships are encoded by Boolean assignment variables $M_{i\nu}, \nu = 1, \ldots, k$ which are summarized in a matrix $\mathbf{M} \in \mathcal{M} = \{0,1\}^{n \times k}$. We set $M_{i\nu} = 1$, if object $\mathbf{o}_i$ is assigned to cluster ν. To avoid multiple group associations, we furthermore enforce $\sum_{\nu \leq k} M_{i\nu} = 1$. Each object $\mathbf{o}_i$ is equipped with a set of n_i observations $\mathcal{X}_i = \{x_{i1}, \ldots, x_{in_i}\}$, $x_{ij} \in \mathbb{R}^d$. These observations are assumed to be drawn according to a particular Gaussian mixture model, which is characteristic for the respective cluster ν of the object. Thus, the generative model for an observation x given the group membership of its associated object is defined as

$$p(x|\,\nu) = \sum_{\alpha=1}^{l} p_{\alpha|\,\nu} g(x|\,\mu_\alpha, \Sigma_\alpha). \tag{1}$$

Here, $g_\alpha(x) = g(x|\,\mu_\alpha, \Sigma_\alpha)$ denotes a multivariate Gaussian distribution with mean μ_α and covariance matrix Σ_α. In order to achieve parsimonious models, the Gaussians g_α are considered to form a common *alphabet* from which the cluster specific distributions are synthesized by a particular choice of mixture coefficients $p_{\alpha|\,\nu}$. In order to further limit the number of free parameters, the covariance matrices $\Sigma_\alpha, \alpha = 1, \ldots, l$ are not altered after being initialized by a preprocessing step, i.e. conventional mixture model estimation. Thus, the remaining free continuous parameters are the means of the Gaussians, the mixture coefficients and the probabilities of the various groups $p_\nu, \nu = 1, \ldots, k$. Gathering these parameters in the set $\Theta = \{p_\nu,\, p_{\alpha|\,\nu},\, \mu_\alpha | \alpha = 1, \ldots, l;\; \nu = 1, \ldots k\}$, the *complete data likelihood* $P(\mathcal{X}, \mathbf{M}|\,\Theta)$ is given by

$$\begin{aligned} p(\mathcal{X}, \mathbf{M}|\,\Theta) = p(\mathcal{X}|\,\mathbf{M}, \Theta) \cdot P(\mathbf{M}|\,\Theta) &= \prod_{i \le n} \sum_{\nu \le k} M_{i\nu} p_\nu p(\mathcal{X}_i|\,\nu, \Theta) \\ &= \prod_{i \le n} \prod_{\nu \le k} \left[p_\nu p\left(\mathcal{X}_i|\,\nu, \Theta\right)\right]^{M_{i\nu}} \quad . \end{aligned} \tag{2}$$

Replacing the sum by a product in eq. (2) is justified since the binary assignment variables $M_{i\nu}$ select one out of k terms.

In the special case of color image segmentation, the abstract objects $\mathbf{o}_i$ can be identified with individual pixel positions or *sites*. The observations $\mathcal{X}$ correspond to locally measured color values. The discrete nature of image data induces a partition of the color space. Computational reasons suggest to further coarsen this partition which leads to a discretization of the color space into regions R_j. Considering the different color channels as being independent, these regions correspond to one dimensional intervals I_j. In practice, the intervals I_j are chosen to cover a coherent set of different color values to alleviate the computational demands. If other image features are available, they can be integrated in this framework as well. For instance, the application of our method to combined color and texture segmentation has been studied and will be discussed in a forthcoming publication. Denote by n_{ij} the number of occurrences that an observation at site i is inside the interval I_j. Inserting in (2) and setting $G_\alpha(j) = \int_{I_j} g_\alpha(x) dx$, the complete data likelihood is given by

$$p(\mathcal{X}, \mathbf{M}|\,\Theta) = \prod_{i \le n} \prod_{\nu \le k} \left[p_\nu \prod_{j \le m} \left(\sum_{\alpha \le l} p_{\alpha|\,\nu} G_\alpha(j) \right)^{n_{ij}} \right]^{M_{i\nu}} \quad . \tag{3}$$

Model Identification: Determining the values of the free parameters is the key problem in model identification for a given data set, which is accomplished by maximum likelihood estimation. To simplify the subsequent computations the *log-likelihood* corresponding to equation 2 is considered:

$$\begin{aligned}\mathcal{L}(\theta|\, \mathcal{X}, \mathbf{M}) &= \log p(\mathcal{X}, \mathbf{M}|\, \Theta) \\ &= \sum_i \sum_\nu M_{i\nu} \left[\log p_\nu + \sum_j n_{ij} \log \left(\sum_\alpha p_{\alpha|\,\nu} G_\alpha(j) \right) \right] . \end{aligned} \tag{4}$$

This equation has to be optimized with respect to the following entities: (1) p_ν, (2) $p_{\alpha|\,\nu}$, (3) the means of the Gaussians μ_α and (4) the hidden variables $\mathbf{M}$. The method of choice for these kinds of problems is the well known *Expectation–Maximization–Algorithm* (EM) [4]. It proceeds iteratively by computing posterior probabilities $P(\mathbf{M}|\, \Theta^{\text{old}})$ in the E-step and maximizing the *averaged complete data log–likelihood* $\mathrm{E}[\mathcal{L}(\Theta|\, \mathbf{M})]$ with respect to Θ in the M-step. Extending this interpretation, EM can be viewed as maximizing the following joint function of the parameters Θ and the hidden states $\mathbf{M}$ (see [3,5,9]):

$$\mathcal{F}' = \mathrm{E}\left[\log p(\mathcal{X}, \mathbf{M}|\, \Theta) + \log p(\mathbf{M}) \right] . \tag{5}$$

Apart from a difference in the sign, this equation is identical to the *generalized free energy* $\mathcal{F}$ at temperature $T = 1$ known from statistical physics. Setting the corresponding cost function $\mathcal{C} = -\mathcal{L}$, the free energy for arbitrary temperatures T is given by the following expression:

$$\mathcal{F} = \mathrm{E}[\mathcal{C}] - T \cdot H. \tag{6}$$

Here, H denotes the entropy of the distribution over the states $\mathbf{M}$. This formal equivalence provides an interesting link to another well known optimization paradigm called *Deterministic Annealing* (DA) [14]. The key idea of this approach is to combine the advantages of a temperature controlled stochastic optimization method with the efficiency of a purely deterministic computational scheme. A given combinatorial optimization problem over a discrete state space is relaxed into a family of search problems in the space $\mathcal{P}(\mathcal{M})$ of probability distributions over that space. In this setting, the *generalized free energy* takes the role of the objective function. The temperature parameter T controls the influence of the entropic term, leading to a convex function in the limit of $T \to \infty$. At $T = 0$ the original problem is recovered. The optimization strategy starts at high temperature and it tracks local minima of the objective function while gradually lowering the computational temperature.

Setting $q_{i\nu} = \mathrm{E}[M_{i\nu}] = p(M_{i\nu} = 1)$, the expected costs of a given configuration is given by:

$$\mathrm{E}[\mathcal{C}] = -\sum_i \sum_\nu q_{i\nu} \left[\log p_\nu + \sum_j n_{ij} \log \left(\sum_\alpha p_{\alpha|\,\nu} G_\alpha(j) \right) \right] . \tag{7}$$

E-Step–Equations: Maximizing eq. (7) with respect to $P(\mathcal{M})$, which basically recovers the E-Step of the EM–scheme, requires to evaluate the partial costs of assigning an object $\mathbf{o}_i$ to cluster ν. The additive structure of the objective function allows us to determine these partial costs h as

$$h_{i\nu} = -\log p_\nu - \sum_j n_{ij} \log \left(\sum_\alpha p_{\alpha|\,\nu} G_\alpha(j) \right) . \tag{8}$$

Utilizing the well known fact from statistical physics that the generalized free energy at a certain temperature is minimized by the corresponding Gibbs distribution, one arrives at the update equations for the various $q_{i\nu}$:

$$q_{i\nu} \propto \exp(-\frac{1}{T} h_{i\nu}) \;=\; \exp\left(\frac{1}{T} \left(\log p_\nu + \sum_j n_{ij} \log \left(\sum_\alpha p_{\alpha|\,\nu} G_\alpha(j) \right) \right) \right) . \tag{9}$$

M-Step–Equations: In accordance with [9], the estimates for the class probabilities p_ν must satisfy

$$\frac{\partial}{\partial p_\nu} \mathcal{F} - \lambda \cdot \left(\sum_{\mu=1}^{k} p_\mu - 1 \right) = 0 \ , \tag{10}$$

where λ is a Lagrange parameter enforcing a proper normalization of p_ν. Expanding $\mathcal{F}$ and solving for p_ν leads to the M-step formulae

$$p_\nu = \frac{1}{n} \sum_{i=1}^{n} q_{i\nu} \ \ , \nu = 1, \ldots, k. \tag{11}$$

While lacking a closed-form solution for the second set of parameters $p_{\alpha|\,\nu}$, their optimal values can be found by an iterated numerical optimization. Instead of directly solving

$$\frac{\partial}{\partial p_{\alpha|\,\nu}} \mathcal{F} - \lambda \cdot \left(\sum_{\gamma=1}^{L} p_{\gamma|\,\nu} - 1 \right) = 0 \ , \tag{12}$$

which would be the analog to eq. (10), we repeatedly select two Gaussian components α_1 and α_2. Keeping $p_{\gamma|\,\nu}$ fixed for $\gamma \notin \{\alpha_1, \alpha_2\}$, $p_{\alpha_2|\,\nu}$ is directly coupled to $p_{\alpha_1|\,\nu}$ via

$$p_{\alpha_2|\,\nu} = 1 - \sum_{\gamma \notin \{\alpha_1, \alpha_2\}} p_{\gamma|\,\nu} - p_{\alpha_1|\,\nu} \ , \tag{13}$$

so that only one free parameter remains. Inserting (13) into (12), we obtain

$$\frac{\partial}{\partial p_{\alpha_1|\,\nu}} \mathcal{F}\left(\alpha_1, \alpha_2\right) = -\sum_{j=1}^{m} \left(\sum_{i=1}^{n} q_{i\nu} n_{ij} \right) \frac{G_{\alpha_1}(j) - G_{\alpha_2}(j)}{\sum_{\gamma=1}^{L} p_{\gamma|\,\nu} G_\gamma(j)} \tag{14}$$

and

$$\frac{\partial^2}{\partial p^2_{\alpha_1|\,\nu}} \mathcal{F}\left(\alpha_1, \alpha_2\right) = \sum_{j=1}^{m} \left(\sum_{i=1}^{n} q_{i\nu} n_{ij} \right) \frac{\left(G_{\alpha_1}(j) - G_{\alpha_2}(j)\right)^2}{\left(\sum_{\gamma=1}^{L} p_{\gamma|\,\nu} G_{\gamma}(j)\right)^2} \geq 0 \ . \tag{15}$$

The joint optimization of α_1 and α_2, therefore, amounts to solving a one-dimensional convex optimization problem. The optimal value of α_1 is either located on the boundary of the interval $\left[0; 1 - \sum_{\gamma \notin \{\alpha_1, \alpha_2\}} p_{\gamma|\,\nu}\right]$, or is equal to the zero-crossing of (14). In the latter case, it can be determined by the Newton method or by an interval bisection algorithm, which were both found to achieve sufficient precision after few optimization steps. The computational demands of this algorithm are dominated by the evaluation of $\sum_{i=1}^{n} q_{i\nu} n_{ij}$, which is linear in the number of sites, n. The computation of the remaining parts of (14) scales with the number of clusters, k, and the number of bins, m, and can thus be done efficiently.

Some care should also be spent on the selection of α_1 and α_2. Although the free energy will monotonously decrease even if α_1 and α_2 are randomly drawn, the convergence can be enhanced by choosing, in each iteration, α_1 and α_2 such that $\left\| \frac{\partial}{\partial p_{\alpha_1|\,\nu}} \mathcal{F}\left(\alpha_1, \alpha_2\right) \right\|$ is maximum. To adjust the mixture distribution $p_{\alpha|\,\nu}$ for a fixed cluster ν, it is usually sufficient to repeat the selection and subsequent optimization of pairs (α_1, α_2) for $c \cdot L$ times, where c is a small constant (e.g. $c = 3$). Although the optimization process might not have found the exact position of the global cost minimum at this time (*incomplete M-step*), any further optimization is unlikely to substantially influence the M-step result, and can thus be skipped.

Finally it is possible to adapt the means μ_α. To improve the readability, we restrict our calculations to one-dimensional data (when operating in d dimensions, we assume diagonal covariance matrices, so that the estimation of the d-dimensional vector μ_α reduces to d one-dimensional optimization problems). Denote by $x_j^\ominus$ and $x_j^\oplus$ the boundaries of the interval $I_j = \left[x_j^\ominus;\, x_j^\oplus\right]$, so that $G_\alpha(j) = \int_{x_j^\ominus}^{x_j^\oplus} g_\alpha(x)dx$. μ_α can then be determined by gradient or Newton descent, the first derivative of $\mathcal{F}$ being given by

$$\frac{\partial}{\partial \mu_\alpha} \mathcal{F} = -\sum_{\nu=1}^{k} \sum_{j=1}^{m} \left(\sum_{i=1}^{n} q_{i\nu} n_{ij} \right) p_{\alpha|\,\nu} \frac{g_\alpha\left(x_j^\ominus\right) - g_\alpha\left(x_j^\oplus\right)}{\sum_{\gamma=1}^{L} p_{\gamma|\,\nu} G_\gamma(j)} \ . \tag{16}$$

We observed in color segmentation experiments, that fixed means μ_α, initialized by a conventional mixture model procedure, produced satisfactory segmentation results. Adapting the means, however, can improve the generative performance of the PDC model.

Multi-Scale Techniques: If the number of objects is large, e.g. in the case of large images, the proposed approach is computationally demanding, even if comparatively efficient optimization techniques like DA are used. In order to arrive at improved running times for the PDC algorithm, a multi-scale optimization scheme [6] [11] is applied. The idea of multi-scale optimization is to lower the computational complexity by decreasing the number of considered entities in the object space. In most application domains for image segmentation it is a natural assumption, that neighboring image sites contain identical, or at least similar, feature histograms. This domain–inherent structure is exploited to create a pyramid of coarsened data and configuration spaces by tying neighboring assignment variables.

It is a well known fact that the reliable estimation of a given number of clusters requires a sufficient amount of data. Since the multi-scale optimization greatly reduces the cardinality of the configuration spaces at coarser levels, the splitting strategy and the coarse to fine optimization have to be brought in line. The inherent splitting behavior of DA optimization supports the coarse to fine hierarchy. Clusters degenerate at high temperatures, leaving only a reduced *effective* number k_T of groups visible. While the computational temperature is continously lowered during optimization, clusters successively split at *phase transitions* [14]. Therefore, a scheme known as *multi-scale annealing*[11] is applied, which couples the splitting strategy and the annealing process.

3 Relation to the Information Bottleneck Framework

The *Information Bottleneck* principle has recently been proposed as a general information theoretical framework for describing clustering problems [17]. Essentially, it formalizes the idea that a given input signal X has to be efficiently encoded by a cluster variable $\tilde{X}$, and that on the other hand the relevant information about a context variable Y should be preserved as well as possible. This tradeoff is made explicit by the difference between two mutual information terms

$$I\left(X;\tilde{X}\right) - \lambda I\left(\tilde{X};Y\right) \quad , \tag{17}$$

which has to be minimized to determine the optimal cluster variables $\tilde{X}$. The quantity $I\left(A;B\right) := H(A) - H(A|B)$ is the mutual information between two random variables A, B [2]. $H(A)$ and $H(A|B)$ are the entropy of A and the conditional entropy of A given B, respectively. $\lambda > 0$ is a control parameter that adjusts the tradeoff between good compression on the one hand and the level of information preservation on the other hand.

The application of this general framework to our generative model is depicted in fig. 1. In our case, the signal X can be identified with the decision to select a single object $i \in \{1, \ldots, n\}$. We can assume that objects are drawn according to a uniform distribution, i.e. $p_i = 1/n$. The object i is then encoded by mapping it to a cluster $\nu \in \{1, \ldots, k\}$, which corresponds to a cluster variable $\tilde{X}$ in the Information Bottleneck framework. As we assume deterministic, unique assignment of objects to clusters, the probability of cluster ν given the object i is a Boolean variable $p\left(\nu|\, i\right) = M_{i\nu}$. Accordingly, the conditional entropy

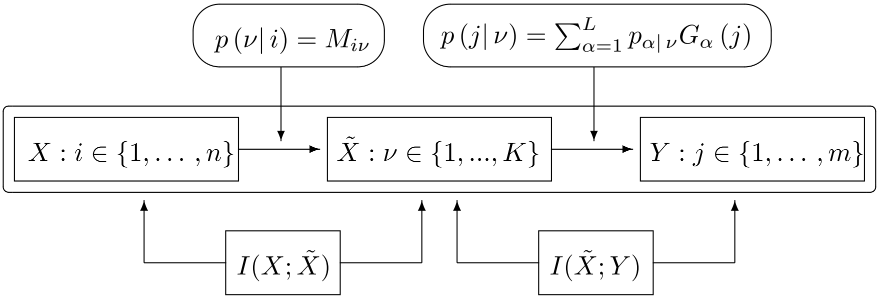

$p(\nu|\,i) = M_{i\nu}$
$p(j|\,\nu) = \sum_{\alpha=1}^{L} p_{\alpha|\,\nu} G_\alpha(j)$
$X : i \in \{1,\dots,n\}$
$\tilde X : \nu \in \{1,...,K\}$
$Y : j \in \{1,\dots,m\}$
$I(X;\tilde X)$
$I(\tilde X;Y)$

Fig. 1. The generative model and its relation to the Information Bottleneck principle.

$H(\tilde X|\,X) = -\sum_i p_i \sum_\nu M_{i\nu} \log M_{i\nu} = 0$ vanishes, which implies

$$I(X;\tilde X) = H(\tilde X) = -\sum_{\nu \le k} p_\nu \log p_\nu \ . \tag{18}$$

As the next step, it is necessary to define the context variable Y, which in the Information Bottleneck framework is used to measure the relevant information preserved by $\tilde X$. As it is desirable to retain the information by which typical observations of an object i is characterized, it is the natural choice to let Y encode the observed bin indices j. n_i denotes the number of observations for object i, so that the relative frequencies n_{ij}/n_i, $j \in \{1,\dots,m\}$, form an object-specific normalized histogram. Furthermore, let p_j denote the marginal probability that an observation is attributed to bin j. The conditional entropy between Y and $\tilde X$ can be rewritten using the Markov dependency between X, $\tilde X$ and Y, i.e. $p(\tilde X|X,Y) = p(\tilde X|X)$:

$$\begin{aligned} H(Y|\tilde X) &= -\sum_{\tilde X, Y} p(Y,\tilde X) \log p(Y|\tilde X) \ = \ -\sum_{\tilde X, Y} \sum_X p(Y,\tilde X, X) \log p(Y|\tilde X) \\ &= -\sum_{X,\tilde X,Y} p(\tilde X|X) P(Y|X) \frac{1}{n} \log p(Y|\tilde X) \ . \end{aligned} \tag{19}$$

Inserting these terms and replacing $I\left(\tilde X;Y\right) = H\left(Y\right) - H\left(Y|\,\tilde X\right)$ yields the bottleneck functional

$$\begin{aligned} & I\left(X;\tilde X\right) - \lambda I\left(\tilde X;Y\right) \\ &= H(\tilde X) - \lambda \sum_{X,\tilde X,Y} p(\tilde X|X) P(Y|X) \frac{1}{n} \log p(Y|\tilde X) - \lambda H(Y) \\ &= -\frac{1}{n} \sum_{i=1}^{n} \sum_{\nu=1}^{k} M_{i\nu} \left[\log p_\nu + \lambda \sum_{j \le m} \frac{n_{ij}}{n_i} \log \left(\sum_{\alpha \le L} p_{\alpha|\,\nu} G_\alpha(j) \right) \right] + \lambda H(Y) \ . \end{aligned} \tag{20}$$

The entropy of Y is a constant and does not influence the search for optimal $\tilde x$ parameters. We can, therefore, subtract $\lambda H(Y)$ from (20) and multiply the equation by n without changing the minimum w.r.t. $M_{i\nu}$and$p_{\alpha|\,\nu}$. This operation yields the function

$$\mathcal{C}^{IB} = -\sum_{i=1}^{n}\sum_{\nu=1}^{k} M_{i\nu}\left[\log p_\nu + \frac{\lambda}{n_i}\sum_{j=1}^{m} n_{ij}\log\left(\sum_{\alpha=1}^{L} p_{\alpha|\,\nu} G_\alpha\left(j\right)\right)\right] \quad . \tag{21}$$

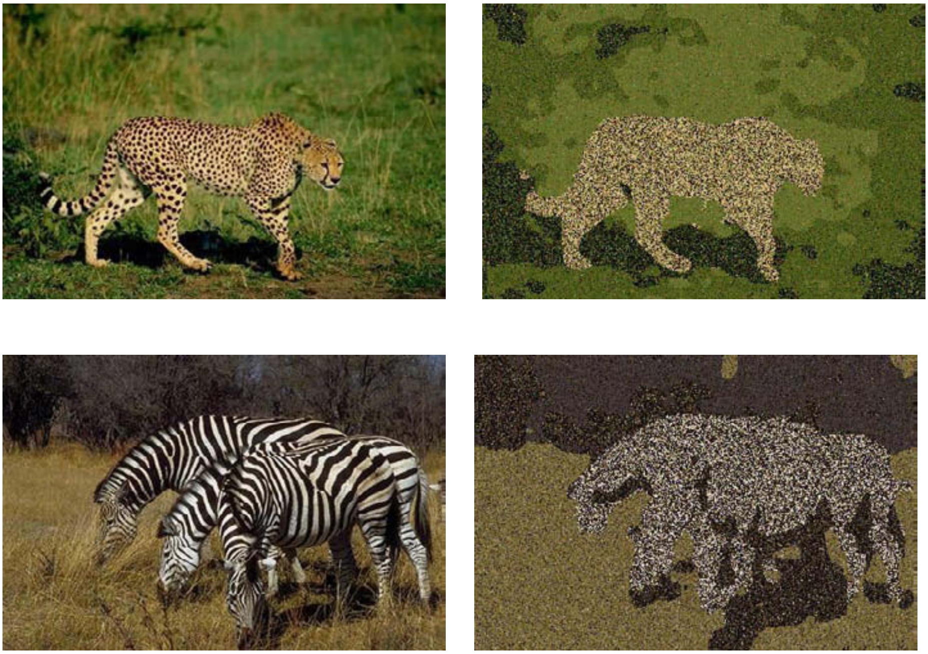
Compared to (4), it is equipped with an additional weighting factor λ, that explicitly controls the influence of the cluster probabilities p_ν. If the number of observations is identical for each site i, i.e. $n_i = \text{const}\ \forall i \in \{1,\ldots,n\}$, we can set $\lambda = n_i$ to obtain our original cost function $\mathcal{C} = -\log \mathcal{L}$. This calculation proves that the generative model introduced in this paper is equivalent to the Information Bottleneck framework suggested in [17].

4 Experimental Results

Implementation Details: Although the proposed approach to clustering histogram data is of general applicability, our primary interest is in the domain of image segmentation. In this contribution, we put a focus on segmentation according to color features. In this setting, the basic measurements are given by the three-dimensional color vectors. The objects $\mathbf{o}_i, i = 1,\ldots,n$ correspond to image sites located on a rectangular grid. In each dimension, the color values are discretized into 32 bins. For all sites, marginal feature histograms are computed in a local neighborhood. In order to determine initial values for the involved Gaussian distributions, a mixture model estimation step is performed prior to the PDC model optimization.

The Generative Model: One of the essential properties of our model is given by its generative nature. It is, therefore, reasonable to evaluate its quality by generating a new image from the learned statistics, i.e., we conducted experiments in which a learned model was used to re-generate its input by sampling. Two examples of this procedure are depicted in fig. 2. These results demonstrate, that the color content of the original image is well represented in its generated counterpart. However, the spatial relationships between the pixels, and thus the texture characteristics, are lost. This effect is due to the histogramming process which destroys these relations. Consequently, they cannot be taken into account by our generative model.

Evolution of Cluster-Assignments: a) The Multi-Scale Framework. In order to give some intuition about the dynamics of the multi-scale optimization, we produced a set of snapshots of the group assignments at various stages of the multi-scale pyramid (fig. 3). Cluster memberships are encoded by color/grey values. The series of images starts in the top left with a grouping at the coarsest stage, continuing to finer levels in a left-to-right and top-to-bottom fashion. For reference, the corresponding input image is also depicted. The interplay of the coarse to fine optimization and the splitting strategy is clearly visible. At



(a) (b)

Fig. 2. Sampling from the learned model : a) original image b) sampled image with four segments.

the coarsest stage, the grouping starts with two clusters. Then, groups are successively split as long as there is sufficient data for a reliable estimation. Upon convergence, results are mapped to the next finer level, at which the grouping cost function, i.e. the generalized free energy resulting from the complete data log-likelihood, is further minimized. These steps are repeated, until final convergence is reached.

Evolution of Cluster-Assignments: b) Phase Transitions in DA. Another interesting phenomenon in the development of assignments in the framework of deterministic annealing is the occurrence of phase transitions. To illustrate that point, we visualized a set of group assignments at various stages of the annealing process (fig. 4). For a better exposition of that particular point we dispensed with multi-scale optimization in these examples. Again, group memberships are visualized by different colors / grey levels. At high computational temperatures, the entropic term of the free energy dominates, leading to random cluster assignments. As the temperature parameter is gradually lowered, the most prominent structural properties of the data begin to emerge in the form of

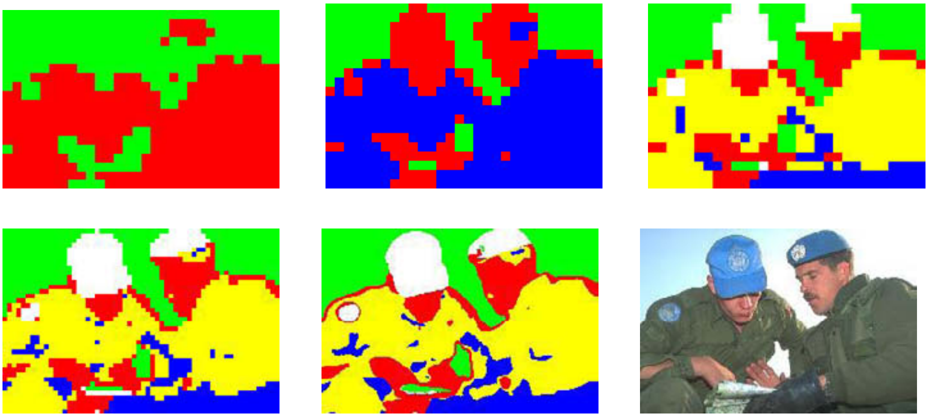


Fig. 3. Evolution of group assignments in the multi-scale framework.

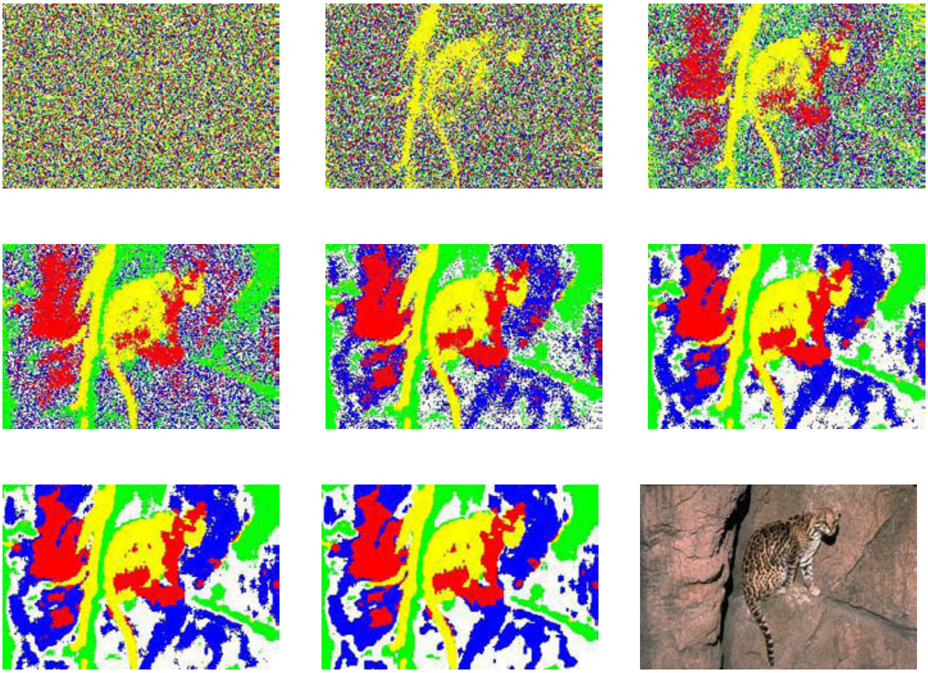
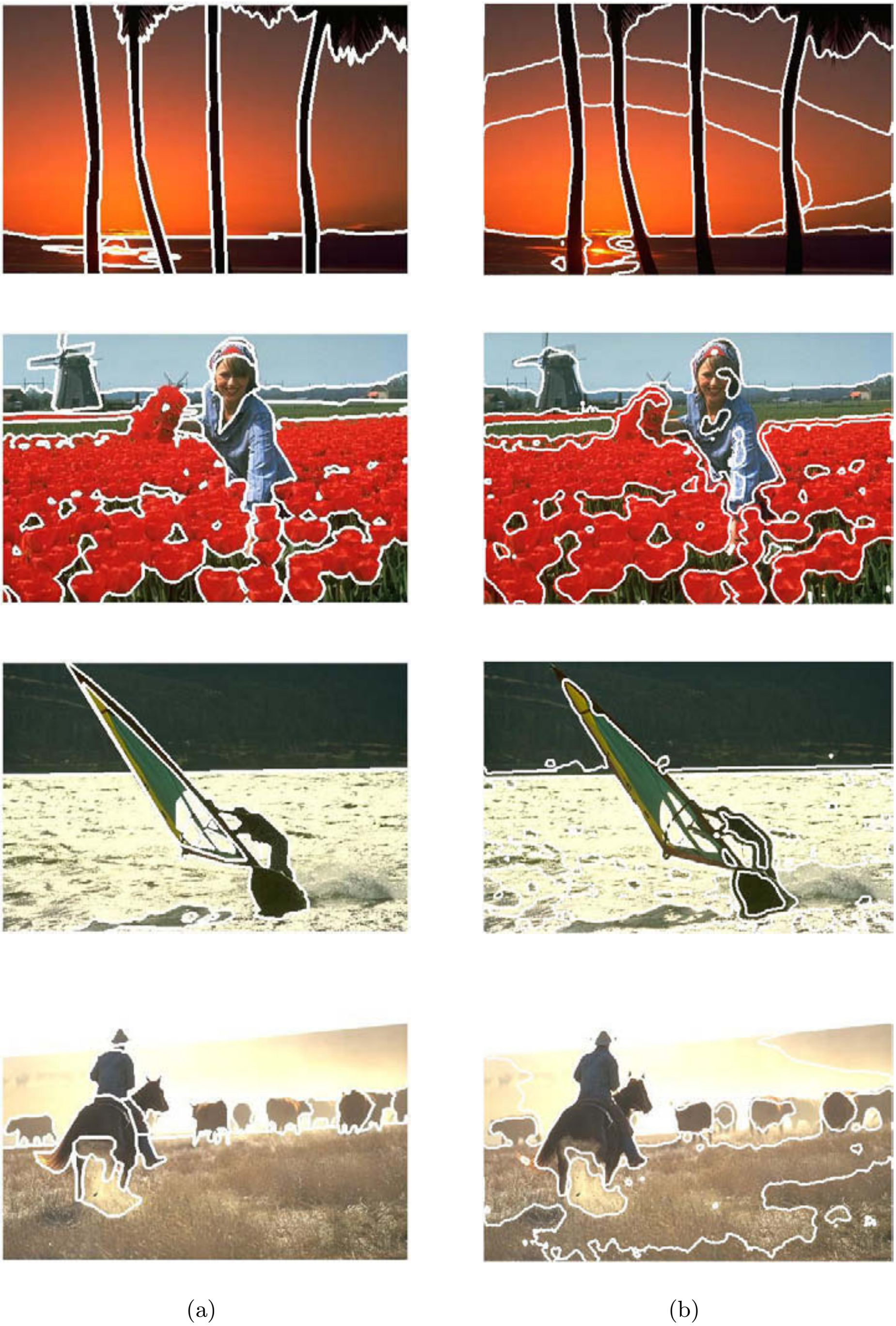


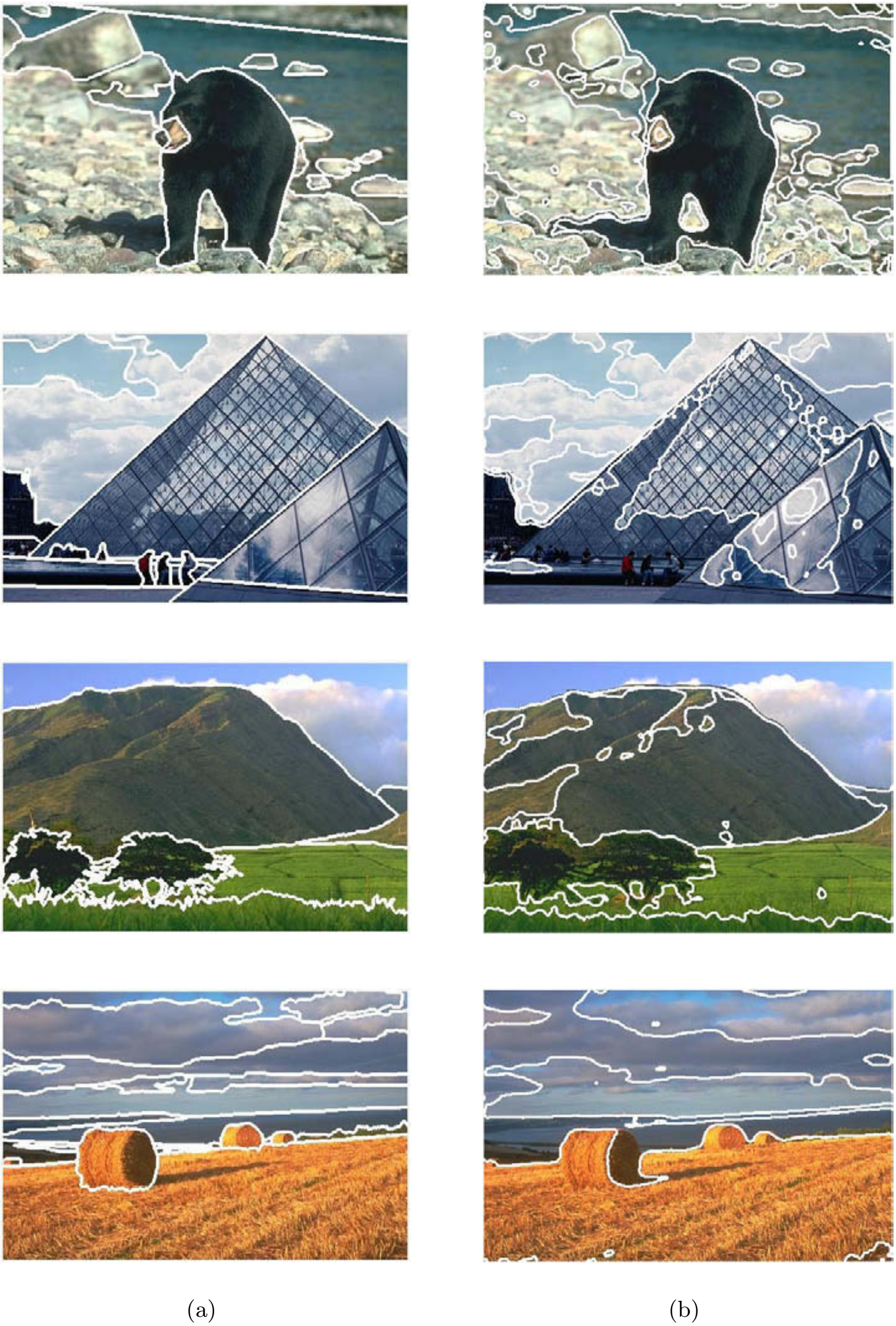
Fig. 4. Evolution of group assignments in deterministic annealing.

stable group memberships. This process continues with the lesser pronounced characteristics of the data manifesting themselves, until a grouping solution with the predefined number of five clusters is reached at low temperatures. A theo-



(a) (b)

Fig. 5. Segmentation results: a) human segmentation, b) PDC segmentation.



(a)

(b)

Fig. 6. Segmentation results: a) human segmentation, b) PDC segmentation.

retical investigation of the critical temperatures for phase transitions in the case of K-Means clustering is given by Rose et al. in [15]. It is shown, that the first split is determined by the variance along the first principal axis of the data. The critical temperature for further phase transitions is more difficult to compute due to inter cluster influences. Because of the structural analogies of our method to K-Means, comparable results are expected to hold for PDC.

Comparative Evaluation: Judging the quality of a given segmentation is difficult due to the fact that ground truth is unavailable in most cases. Furthermore the segmentation is often only one item in a large context of processing steps. In those cases it is only natural, as Borra and Sakar point out [1], to judge the segmentation quality with respect to the overall task. In contrast to this view, Malik et al. examined human image segmentation [8] experimentally. Their results indicate a remarkable consistency in the segmentation of given images among different human observers. This finding motivates their current effort to construct a database of human segmented images from the Corel collection for evaluation purposes, which is publicly available. This set of images has been chosen as our testbed, making a direct comparison between our novel segmentation model and human performance possible. Figures 5 and 6 depict the best (w.r.t. PDC) human segmentation in comparison to the segmentation results achieved by parametric distributional clustering for four segments. Segment boundaries for both human and machine segmentation are given by thick white lines. It is obvious that segmentations which require high-level sematic knowledge like shadows cannot be reproduced by our method but segmentations based on low level color information are reliably inferred.

5 Conclusion

In this contribution, a novel model for unsupervised image segmentation has been proposed. It is based on robust measurements of local image characteristics given by feature histograms. As one of the main contributions, it contains a continuous model for the group-specific distributions. In contrast to existing approaches, our method thus explicitly models the noise-induced errors in the histogramming of image content. Being based on the theoretically sound maximum likelihood framework, our approach makes all modeling assumptions explicit in the cost function of the corresponding generative model. Moreover, there exists an informative connection to information theoretic concepts. The Information Bottleneck model offers an alternative interpretation of our method as a way to construct a simplified representation of a given image while preserving as much of its relevant information as possible. Finally, the results demonstrate the good performance of our model, often yielding close to human segmentation quality on the testbed.

Acknowledgement. The authors are grateful to N. Tishby for various discussions on the bottleneck framework for learning. We also thank J. Puzicha for valuable contributions at the beginning of this project. This work has been supported by the German Research Foundation (DFG) under grant #BU 914/3–1 and by cooperations with INFINEON and InfoTerra GmbH.

References

1. S. Borra and S. Sakar. A framework for performance characterization of intermediate–level grouping modules. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(11):1306–1312, 1997.
2. Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory.* John Wiley & Sons, 1991.
3. I. Csizàr and G. Tusnady. Information geometry and alternating minimization procedures. In E. J. Dudewicz et al, editor, *Recent Results in Estimation Theory and Related Topics*, Statistics and Decisions, Supplement Issue No. 1. Oldenbourg, 1984.
4. A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society B*, 39:1–38, 1977.
5. R. J. Hathaway. Another interpretation of the EM algorithm for mixture distributions. *Statistics and Probability Letters*, 4:53–56, 1986.
6. F. Heitz, P. Perez, and P. Bouthemy. Multiscale minimization of global energy functions in some visual recovery problems. *CVGIP: Image Understanding*, 59(1):125–134, 1994.
7. A. Jain and R. Dubes. *Algorithms for Clustering Data.* Prentice Hall, Englewood Cliffs, NJ 07632, 1988.
8. D. Martin, C. Fowlkes, D. Tal, and J. Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proc. ICCV'01*, 2001.
9. R. M Neal and G. E. Hinton. A view of the EM algorithm that justifies incremental, sparse, and other variants. In M. I. Jordan, editor, *Learning in Graphical Models.* MIT Press, 1999.
10. F. Pereira, N. Tishby, and L. Lee. Distributional clustering of english words. In *30th International Meeting of the Association of Computational Linguistics*, pages 183–190, Columbus, Ohio, 1993.
11. J. Puzicha and J. M. Buhmann. Multiscale annealing for unsupervised image segmentation. *Computer Vision and Image Understanding*, 76(3):213–230, 1999.
12. J. Puzicha, T. Hofmann, and J. M. Buhmann. Histogram clustering for unsupervised segmentation and image retrieval. *Pattern Recognition Letters*, 20:899–909, 1999.
13. J. Puzicha, T. Hofmann, and J. M. Buhmann. A theory of proximity based clustering: Structure detection by optimization. *Pattern Recognition*, 2000.
14. K. Rose, E. Gurewitz, and G. Fox. A deterministic annealing approach to clustering. *Pattern Recognition Letters*, 11:589–594, 1990.
15. K. Rose, E. Gurewitz, and G. Fox. Statistical mechanics and phse transitions in clustering. *Physical Review Letters*, 65(8):945–948, 1990.
16. Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905, 2000.
17. N. Tishby, F. Pereira, and W. Bialek. The information bottleneck method. In *Proc. of the 37th annual Allerton Conference on Communication, Control, and Computing*, 1999.