Simplifying Mixture Models Using the Unscented Transform

Jacob Goldberger, Hayit Greenspan, and Jeremie Dreyfuss

Abstract—Mixture of Gaussians (MoG) model is a useful tool in statistical learning. In many learning processes that are based on mixture models, computational requirements are very demanding due to the large number of components involved in the model. We propose a novel algorithm for learning a simplified representation of a Gaussian mixture that is based on the Unscented Transform, which was introduced for filtering nonlinear dynamical systems. The superiority of the proposed method is validated on both simulation experiments and categorization of a real image database. The proposed categorization methodology is based on modeling each image using a Gaussian mixture model. A category model is obtained by learning a simplified mixture model from all the images in the category.

Index Terms—Mixture of Gaussians, reduced model, Unscented Transform, weighted likelihood, clustering.

1 Introduction

THE Mixture of Gaussians (MoG) model is a flexible and powerful parametric framework for unsupervised data grouping. The mixture model has been widely used in clustering and density estimation where the model parameters can be estimated by the standard Expectation-Maximization (EM) algorithm. In many learning processes based on mixture models, the computational requirements are very demanding due to the large number of components involved in the model. Such a situation can be handled by simplifying the mixture model by reducing the number of components. One example of the need for model simplification is statistical inference in switching dynamic linear models, where performing exact inference with an MoG prior causes the number of Gaussian components representing the current belief to grow exponentially in time [1]. Another example for the need of model simplification can be found in mean-shift segmentation. The Parzen window estimator can be used to reveal the modes corresponding to dominant clusters in the data. This property is utilized in the mean-shift clustering algorithm [2]. The mean shift has been proven to be successful for color image segmentation. However, mean-shift segmentation can be quite expensive due to the large number of kernels involved in the density estimator. To reduce the model's complexity, we can apply the mean-shift procedure to a simplified mixture model [22]. Note that a fast version of mean shift using kd-trees only facilitates range searching but does not reduce the expensive computation associated with the large number of kernels. Additional applications that can benefit from simplifying the mixture model are nonparametric belief propagation [19], semiparametric particle filters [9], efficient decoding of lattice error correcting codes [14], and image clustering based on an MoG version of the information-bottleneck principle [6].

One may argue that we can find a simplified model by simply relearning a simpler MoG from the original data set using the standard EM algorithm. Apart from the computational complexity gain and avoiding the need for keeping the original data set, in many situations the mixture model is not obtained directly from clustering a given data set but as a result of a previous learning step (e.g., [9] and [14]). In these cases, there is no way to learn a simplified model using the EM algorithm. Instead, we only have an MoG representation of a multimodal distribution and we are interested in simplifying the model into a more compact model that can be efficiently used in the next step of the learning process or in the subsequent testing phase.

There is yet another aspect of Gaussian grouping, which can be beneficial. In many settings, input objects are naturally represented by multiple samples drawn from an underlying unimodal distribution. Hence, we can identify each Gaussian with a single object in a data set. Therefore, clustering such objects can be achieved through grouping the mixture components into a simplified model. Examples of such grouping tasks are image categorization [7] and efficient routing schemes for sensor networks [3].

A strong link exists between the mixture clustering problem, which is the focus of this study, and the problem of defining a meaningful and efficiently computed distance measure between two Gaussian mixtures. Any such distance measure induces a mixture clustering cost function. The mixture grouping algorithm searches for a simplified model that is most similar to the original multimixture model according to the specified distance measure. A common proximity criterion between two distributions is the Kullback-Leibler (KL) divergence. Given an n-component mixture model f, we can look for a simplified model $\hat{g} = \arg\min_g D(f\|g)$ where $D()$ is the KL divergence and the minimization is performed over all m-component MoGs. This criterion leads to an intractable optimization problem; there is not even a closed-form expression for the KL divergence between two MoGs let alone an analytic minimizer of its second argument.

Goldberger and Roweis [7] suggested a mixture clustering algorithm based on grouping of the mixture components. Soft versions of that clustering algorithm appear in [18] and [20]. The algorithm is derived from an approximation of the KL divergence between two MoGs f and g that is based on matching each component of f to the most similar component of g. This matching-based method approximates the KL divergence well if the Gaussian elements are far apart. However, if there is a significant overlap between the Gaussian elements, the assignment of a single component of $g(x)$ to each component of the original mixture $f(x)$ becomes less accurate. To handle overlapping situations, we propose a novel reduced representation of an MoG that is based on the Unscented Transform (UT). Goldberger et al. [5] showed that the UT can be used to obtain a good approximation for the KL divergence between two MoGs. In this study, we show that this approximation can be used to obtain an improved method for simplifying the MoG component structure. The UT and the KL approximation based on it are reviewed in Section 2. A novel algorithm for simplifying a mixture model that is based on the UT is presented in Section 3. Related work is discussed in Section 4, and experimental results are presented in Section 5. A preliminary version of this paper appeared in [4].

2 The Unscented Transform Approximation of KL

The Unscented Transformation is a method for calculating the statistics of a random variable, which undergoes a nonlinear

- *J. Goldberger is with the School of Engineering, Bar-Ilan University, Ramat-Gan 52900, Israel. E-mail: goldbej@eng.biu.ac.il.*
- *H. Greenspan and J. Dreyfuss are with the Department of Biomedical Engineering, Tel-Aviv University, Tel Aviv 69978, Israel. E-mail: hayit@eng.tau.ac.il, jeremied@lycos.com.*

Manuscript received 15 Oct. 2007; revised 2 Apr. 2008; accepted 7 Apr. 2008; published online 14 Apr. 2008.
Recommended for acceptance by J. Buhmann.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number TPAMI-2007-10-0707.
Digital Object Identifier no. 10.1109/TPAMI.2008.100.

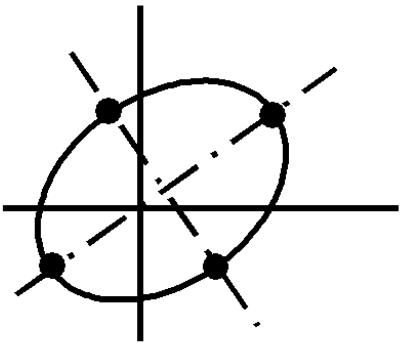

Fig. 1. The sigma points of the UT along the eigenvectors of the covariance matrix of the normal distribution.

transformation [13]. It is successfully used for nonlinear filtering. The Unscented Kalman filter (UKF) [13], [12] is more accurate, more stable, and far easier to implement than the extended Kalman filter (EKF). In cases where the process noise is Gaussian, it is also better than the particle filter, which is based on Monte Carlo simulations. Unlike the EKF, which uses the first-order term of the Taylor expansion of the nonlinear function, the UKF uses the true nonlinear function and approximates the distribution of the function output. Following [5], we show how the UT can be utilized to obtain an approximation for the KL divergence between two MoGs.

We shall first review the UT. Let x be a d-dimensional normal random variable $x \sim f(x) = N(\mu, \Sigma)$ and let $h(x) : R^d \to R$ be an arbitrary nonlinear function. We want to approximate the expectation of $h(x)$, which is $\int f(x)h(x)dx$. With the UT approach, a set of $2d$ "sigma points" are chosen as follows:

$$\begin{aligned} x_k &= \mu + (\sqrt{d\Sigma})_k \qquad k = 1, \ldots, d, \\ x_{d+k} &= \mu - (\sqrt{d\Sigma})_k \qquad k = 1, \ldots, d, \end{aligned}$$

such that $(\sqrt{\Sigma})_k$ is the kth column of the matrix square root of Σ. Let $UDU^\top$ be the singular value decomposition of Σ, such that $U = \{U_1, \ldots, U_d\}$ and $D = \mathrm{diag}\{\lambda_1, \ldots, \lambda_d\}$ then $(\sqrt{\Sigma})_k = \sqrt{\lambda_k}U_k$. In other words, U_k is an eigenvector of Σ corresponding to the eigenvalue λ_k. The sigma points completely capture the true mean and variance of the normal distribution $f(x)$ (see Fig. 1). The uniform distribution over the sigma points $\{x_k\}_{k=1}^{2d}$ has mean μ and covariance matrix Σ. Given the sigma points, we define the following approximation:

$$E(h(x)) = \int f(x)h(x)dx \approx \frac{1}{2d}\sum_{k=1}^{2d} h(x_k). \tag{1}$$

Similarly, we can obtain an estimation for the variance of $h(x)$. Although this approximation algorithm resembles a Monte Carlo method, no random sampling is used, thus only a small number of points are required. It can be verified that if $h(x)$ is a linear or even a quadratic function of x, the approximation is precise. The basic Unscented method can be generalized. The mean of the Gaussian distribution μ can be also included in the set of sigma points [11].

The UT can be used to approximate the KL divergence between the following two MoGs. Let f and g be two d-dimensional MoGs such that

$$f = \sum_{i=1}^{n} \alpha_i f_i = \sum_{i=1}^{n} \alpha_i N(\mu_i, \Sigma_i) \quad \text{and} \quad g = \sum_{j=1}^{m} \beta_j g_j.$$

Since $D(f\|g) = \int f \log f - \int f \log g$, it is sufficient to show how we can approximate $\int f \log g$. The linearity of the construction of f from its components yields

$$\int f \log g = \sum_{i=1}^{n} \alpha_i \int f_i \log g = \sum_{i=1}^{n} \alpha_i E_{f_i}(\log g).$$

Assume that x is a Gaussian random variable $x \sim f_i$, then $E_{f_i}(x) = \mu_i$ and $E_{f_i}(\log g(x))$ is the mean of the random variable $\log g(x)$, which is a nonlinear function of x. This mean can be approximated using the UT:

$$\int f \log g \approx \frac{1}{2d}\sum_{i=1}^{n} \alpha_i \sum_{k=1}^{2d} \log g(x_{i,k}), \tag{2}$$

such that

$$\begin{aligned} x_{i,k} &= \mu_i + \left(\sqrt{d\Sigma_i}\right)_k \qquad k = 1, \ldots, d, \\ x_{i,d+k} &= \mu_i - \left(\sqrt{d\Sigma_i}\right)_k \qquad k = 1, \ldots, d. \end{aligned} \tag{3}$$

To simplify notations, we denote the UT approximation (UTA) (2) by

$$\mathrm{UTA}(f, g) = \frac{1}{2d}\sum_{i=1}^{n} \alpha_i \sum_{k=1}^{2d} \log g(x_{i,k}). \tag{4}$$

If the covariance matrices of the two MoGs are restricted to be diagonal, the computational complexity of the UTA is significantly reduced. Assuming the covariance matrices of the components of f have the following form:

$$\Sigma_i = diag\left(\sigma_{i,1}^2, \ldots, \sigma_{i,d}^2\right) \qquad i = 1, \ldots, n,$$

then the sigma points are simply

$$\mu_i \pm \sqrt{d}\,\sigma_{i,k} \qquad k = 1, \ldots, d.$$

The diagonal structure of the covariance matrices of the components of g can be utilized to further reduce the complexity of computing the value $\log(g)$ on the $2dn$ sigma points. In fact, if f has a huge number of components, and g has a much smaller number, the most useful case is probably where f is diagonal and g uses full covariances, so as to give best representation with small number of components. The UTA is summarized below.

Input: Two d-dimensional Gaussian mixtures f and g such that $f = \sum_{i=1}^{n} \alpha_i N(\mu_i, \Sigma_i)$

Output: UTA approximation of $\int f \log g$:

$$\mathrm{UTA}(f, g) = \tfrac{1}{2d}\textstyle\sum_{i=1}^{n} \alpha_i \sum_{k=1}^{2d} \log g(x_{ik})$$
$$\text{s.t. } x_{ik} = \mu_i \pm (\sqrt{d\Sigma_i})_k.$$

3 The Unscented Clustering Algorithm

The general mixture simplification task can be formalized as follows: Assume that we are given a mixture density f composed of n d-dimensional Gaussian components:

$$f = \sum_{i=1}^{n} \alpha_i f_i = \sum_{i=1}^{n} \alpha_i N(\mu_i, \Sigma_i). \tag{5}$$

We want to simplify the MoG f into a reduced mixture of $m < n$ components. If we denote the set of all (d-dimensional) Gaussian mixture models with at most m components by $\mathrm{MoG}(m)$, one way to formalize the goal of clustering is to say that we want to find the element g of $\mathrm{MoG}(m)$ "closest" to f under some distance measure.

Utilizing the UTA distance measure, the clustering task can be stated as follows: Given an MoG $f = \sum_{i=1}^{n} \alpha_i f_i$, we want to find a reduced MoG representation $g = \sum_{j=1}^{m} \beta_j g_j$ that best approximates f based on UTA measure. More formally, we want to find an m-component MoG g that maximizes the following expression:

$$\mathrm{UTA}(f,g) = \frac{1}{2d}\sum_{i=1}^{n}\alpha_i \sum_{k=1}^{2d} \log g(x_{ik}), \tag{6}$$

where x_{ik} are the sigma points of f. To optimize this equation, we can consider the sigma points as a deterministic set of weighted samples from f. The UTA expression can be viewed as a weighted log-likelihood of the distribution $g()$ based on the sigma points. Hence, the (weighted) free-energy function [17] in this case is

$$FE(q) = \sum_{ik}\alpha_i H(q_{ik}) + \sum_{ijk}\alpha_i q_{ik}(j)\log\big(\beta_j g_j(x_{ik})\big), \tag{7}$$

where q_{ik} is a discrete distribution on the m components of g, and H is the entropy function. Note that if the hidden random variable of the MoG f is not uniformly distributed (i.e., not all the coefficients α_i are $1/n$) the expression $\mathrm{UTA}(f,g)$, viewed as a function of g, has no strict probabilistic meaning. It can be viewed as a weighted log-likelihood of the sigma points. Similarly the weighted free energy (7) has no strict probabilistic meaning. In Appendix A, however, we prove that the EM theory can be extended to this case of weighed samples. Minimizing the (negative) free energy yields an iteration of the EM algorithm for learning the reduced model. The E-step is

$$w_{ikj} = \frac{\beta_j g_j(x_{ik})}{g(x_{ik})}, \tag{8}$$

$$i = 1,\ldots,n, \qquad k = 1,\ldots,2d, \qquad j = 1,\ldots,m.$$

The probabilistic interpretation of w_{ikj} is the posterior probability that the sigma point x_{ik} was generated using the jth component of g. The M-step is

$$\begin{aligned}
\beta_j &= \frac{1}{2d}\sum_{ik}\alpha_i w_{ikj},\\
\mu'_j &= \frac{\sum_{ik}\alpha_i w_{ikj} x_{ik}}{\sum_{ik}\alpha_i w_{ikj}},\\
\Sigma'_j &= \frac{\sum_{ik}\alpha_i w_{ikj}\Big(x_{ik}-\mu'_j\Big)\Big(x_{ik}-\mu'_j\Big)^{\top}}{\sum_{ik}\alpha_i w_{ikj}},
\end{aligned} \tag{9}$$

where μ'_j and Σ'_j are the updated parameters of the jth component of the reduced MoG g. The general EM-algorithm theory guarantees that the expression $\mathrm{UTA}(f,g)$ monotonically increases during the EM iterations. Hence, at the convergence point, we obtain a reduced model $g(x)$ that is (locally) optimal according to the UTA proximity measure. The UTA Clustering (UTAC) algorithm is summarized below.

The UTAC algorithm:
Input: A d-dimensional mixture $f = \sum_{i=1}^{n}\alpha_i N(\mu_i,\Sigma_i)$ and the desired reduction parameter m
Output: $g = \sum_{j=1}^{m}\beta_j N(\mu'_i,\Sigma'_i)$ such that $\mathrm{UTA}(f,g)$ is maximal.
E-step: $w_{ikj} = \frac{\beta_j g_j(x_{ik})}{g(x_{ik})}$ s.t. $x_{ik} = \mu_i \pm (\sqrt{d\Sigma_i})_k$
M-step:

$$\beta_j = \tfrac{1}{2d}\textstyle\sum_{ik}\alpha_i w_{ikj}$$

$$\mu'_j = \frac{\sum_{ik}\alpha_i w_{ikj} x_{ik}}{\sum_{ik}\alpha_i w_{ikj}}$$

$$\Sigma'_j = \frac{\sum_{ik}\alpha_i w_{ikj}(x_{ik}-\mu'_j)(x_{ik}-\mu'_j)^{\top}}{\sum_{ik}\alpha_i w_{ikj}}.$$

4 Related Work

An alternative approximation of the KL distance between two MoGs $f = \sum_{i=1}^{n}\alpha_i f_i$ and $g = \sum_{j=1}^{m}\beta_j g_j$ is based on matching each component of f to one of the Gaussians of g. The match approximation is [7]

$$D(f\|g) \approx \sum_{i=1}^{n}\alpha_i \min_j D(f_i\|g_j). \tag{10}$$

Concentrating on the similar task of approximating $\int f\log g$, the formula of the match-based approximation, which we dub Gaussian Match Approximation (GMA), is

$$\mathrm{GMA}(f,g) = \sum_{i=1}^{n}\alpha_i \max_j \int f_i \log g_j \approx \int f\log g. \tag{11}$$

We can also consider a variant of (11) based on a soft component assignment [7], [18]:

$$\int f\log g \approx \frac{1}{\lambda}\sum_{i=1}^{n}\alpha_i \log\left(\sum_j \beta_j \exp\left(\lambda\int f_i\log g_j\right)\right), \tag{12}$$

such that in [18] λ is set to 1 and in [7] λ is set to ∞. Every distance measure $d(,)$ between two MoGs naturally induces a mixture simplification criterion. Given an MoG f, we look for a simplified mixture model g such that $d(f,g)$ is minimal. The GMA proximity measure can be used to derive a reduced representation of a given MoG f as follows [7]: The MoG $\hat{g}$ is an optimal reduced approximation for f if

$$\hat{g} = \arg\max_g \mathrm{GMA}(f,g) \tag{13}$$

such that the maximization is performed over all $g \in \mathrm{MoG}(m)$.

The EM algorithm can be utilized to find the best-reduced mixture g based on the GMA criterion. In the EM algorithms presented in [7] and [18], the E-step (8) is

$$w_{ij} = \frac{\beta_j e^{-\lambda D(f_i\|g_j)}}{\sum_l \beta_l e^{-\lambda D(f_i\|g_l)}}. \tag{14}$$

In the case $\lambda = \infty$, the E-step is simply

$$w_{ij} = \begin{cases} 1 & j = \arg\min_k D(f_i\|g_k),\\ 0 & \text{else.}\end{cases} \tag{15}$$

In the M-step, the updated Gaussian g_j is obtained by collapsing $\frac{\sum_i w_{ij}\alpha_i f_i}{\sum_i w_{ij}\alpha_i}$ into a single Gaussian. The term collapsing means shifting from a given distribution to a normal distribution that has the same first and second moments. The M-step updating of the mixture coefficient β_j is $\beta_j = \sum_i \alpha_i w_{ij}$. To simplify notation, we dub this GMA Clustering algorithm as GMAC. The algorithm is summarized below.

The GMAC algorithm:
Input: A d-dimensional mixture $f = \sum_{i=1}^{n}\alpha_i N(\mu_i,\Sigma_i)$, the desired reduction parameter m, and a softness parameter λ.
Output: $g = \sum_{j=1}^{m}\beta_j N(\mu'_i,\Sigma'_i)$ such that $\mathrm{GMA}(f,g)$ is maximal.

E-step: $w_{ij} = \frac{\beta_j e^{-\lambda D(f_i\|g_j)}}{\sum_l \beta_l e^{-\lambda D(f_i\|g_l)}}$

M-step: $g_j = N(\mu'_j,\Sigma'_j)$ is obtained by collapsing $\frac{\sum_i w_{ij}\alpha_i f_i}{\sum_i w_{ij}\alpha_i}$ into a single Gaussian:

$\mu'_j = \sum_i w_{ij}\alpha_i\mu_i / \sum_i w_{ij}\alpha_i$
$\Sigma'_j = \sum_i w_{ij}\alpha_i(\Sigma_i + (\mu_i-\mu'_j)(\mu_i-\mu'_j)^{\top}) / \sum_i w_{ij}\alpha_i$

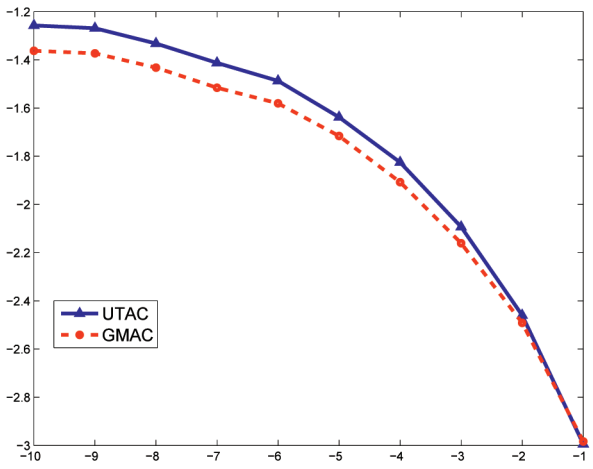

Fig. 2. Comparison between two reduced models UTAC and GMAC on simulation data. The graph shows the (Monte Carlo) cross-entropy of the original model and the reduced model, as a function of ϵ on a logarithmic scale.

The UTA was found to be a better approximation of the KL distance between two MoGs than the GMA for various applications such as image categorization [5], [7]. Utilizing the GMA proximity measure as a criterion for obtaining a reduced representation of a given MoG motivates the approach of this study of using the UTA as a cost function to obtain a better simplified model. In Appendix B, we show that if the proposed UTAC algorithm is modified such that all the sigma points of the same Gaussian component are constrained to be assigned to the same Gaussian component, the modified UTAC algorithm is exactly the EM algorithm based on the GMA approximation (14) such that λ is set to be $2d$.

5 Experimental Results

5.1 Simulation Results

To assess the quality of the proposed Unscented-based approximation, we conducted the following simulation experiment. In each session, we sample a random mixture f of many Gaussian components and we search for a reduced representation g using a mixture with a small number of components. The original mixture models were randomly sampled according to the following rules. The number of Gaussians in the original 2D MoG f was 20. We search for an optimal reduced model comprising five components. For each Gaussian $N(\mu, \Sigma)$, μ was sampled from $N(0, I)$ and Σ was sampled from the Wishart distribution as follows: The entries of a matrix $A_{2\times 2}$ were independently sampled from $N(0, 1)$ and we set $\Sigma = \epsilon AA^T$. The parameter ϵ controls the size of the covariance matrices. As we decrease ϵ, the Gaussians that compose the MoG are further apart.

As a comparison to the proposed method, the matching-based learning method proposed by Goldberger and Roweis [7] [see (11)] was also implemented. Another important issue is how to assess the quality of the approximation obtained from the learning methods. It was validated several times [5], [18] that the distance measure based on the UT (see Section 2) is a good method to measure the distance between two MoGs in terms of accuracy and computational complexity. However, we can obviously expect that the reduced model based on the UT best approximates the original model using the UT distance since it was chosen exactly to optimize this criterion. Hence, it is meaningless to use the UT-based distance measure to compare the quality of different model reduction methods. Instead, we measure the approximation quality based on a Monte Carlo simulation (based on 10,000 samples) of the KL distance between the original and the reduced MoGs. In other words, the score we utilized is the (Monte Carlo approximation of the) asymptotic log-likelihood of data sampled from the original model f based on the reduced model. The experiment was repeated 1,000 times for each ϵ. Fig. 2 shows the simulation results. The KL score is shown as a function of $\log_2(\epsilon)$. As can be seen, better approximation results were obtained for all values of ϵ using the UT-based reduction algorithm.

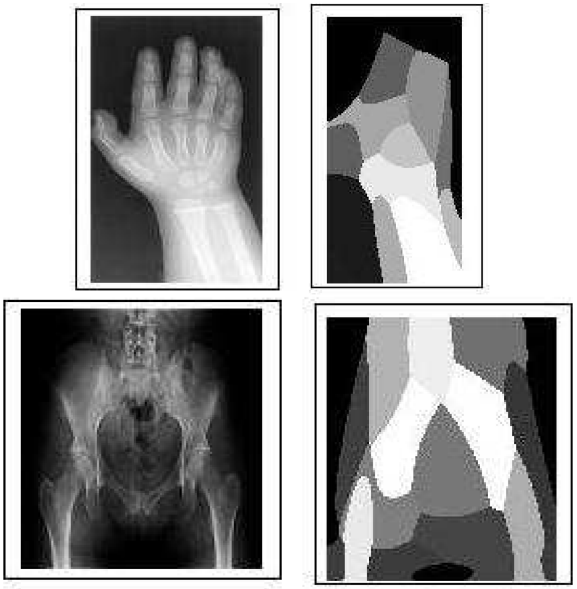

Fig. 3. (Left) Examples of images and (right) their MoG modeling. The model is shown via segmentation of the image pixels based on the MoG.

5.2 Medical Data Categorization

To evaluate the two different reduction methods, we used MoGs computed from a set of 1,680 medical images, which were prelabeled by an expert as belonging to 21 different categories. The categories include MRI images of the knee and spine, CT images of the abdomen and skull, as well as 17 classes of digital radiographs. The X-ray images are a subset of the IRMA project [13], which is being collected and labeled by experts in an extensive ongoing research effort toward medical content retrieval in picture archiving and communication systems (PACSs). The data consists of medical radiographs taken from clinical routine at the Department of Diagnostic Radiology, Aachen University Hospital, Aachen, Germany. The images are taken secondary digital, i.e., scanned from conventional film-based radiographs at a high resolution (typical 2,000 $\times$ 3,000 pixels) and are downscaled to a typical resolution of 300 $\times$ 500 pixels (8-bit) [15]. Images are classified by medical experts according to the imaging modality, the examined region, the image orientation with respect to the body, and the biological system under evaluation.

The images in the database show not only poor contrast but also great intensity variability thus presenting an interesting challenge for a modeling and classification task. As proposed by Greenspan and Pinhas [8], a 5D feature space is used to represent the images, including intensity, texture (contrast, scale), and (x, y) coordinates of the pixel. Unsupervised clustering, based on the EM algorithm, is used to compute the MoGs of each image, thus giving a compact representation of homogenous regions in the feature space. Utilizing the (x, y) coordinates as additional features is a standard way to impose spatial consistency into the MoG image representation. Each image is represented using a mixture of 16 full-covariance Gaussian components. This representation can be easily rendered by replacing each pixel in the image by the mean intensity value of the Gaussian to which it has been clustered. Fig. 3 provides several examples of images and their MoGs

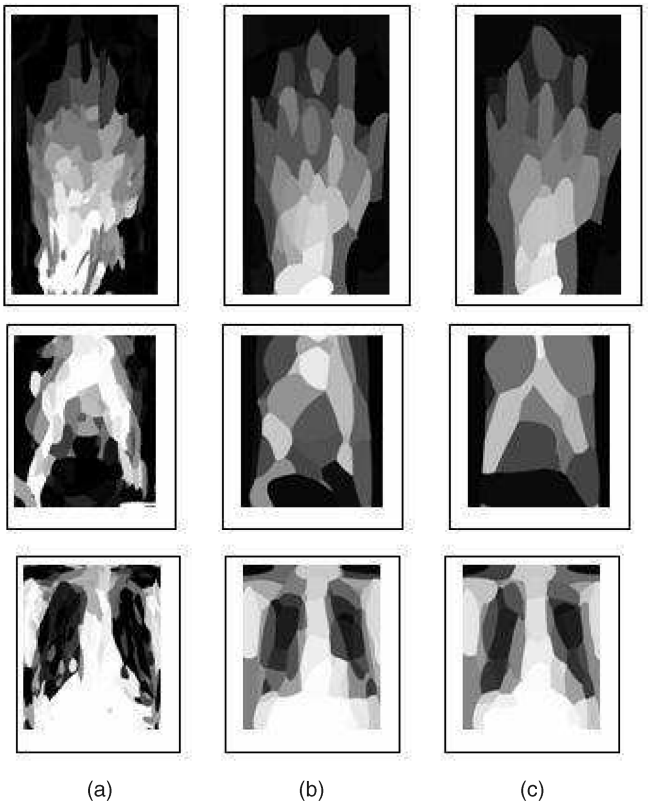

Fig. 4. (Left to right) Full Model and Reduced Models with 10 percent and 5 percent of the Full Model's size for (top to bottom) the "Arm," "Pelvis," and "Chest posterior anterior" categories. The category model is shown by finding the most likely Gaussian component at each pixel (x, y) and plotting the Gaussian mean intensity at that pixel.

showing that the modeling of images still preserve a very low resolution of their visual content.

To test our reduction methods, the images were divided such that part of the images in each category serve as a training set and the remaining serve as a testing set. For each category in the training set, an exhaustive model, which we will refer to as the *Full Model*, is obtained by merging all the image MoGs from the same category together into a mixture of MoGs. The merging is done simply by summing all the MoGs in a category and normalizing each Gaussian's weight in the mixture by the number of images in the category. Depending on the category, the number of components in the full model varies between several hundreds and several thousands. Afterward, both the matching-based (GMAC) and the Unscented-based (UTAC) clustering algorithms are applied to the Full Model to create reduced MoGs with fewer Gaussians. The data-driven motivation behind the reduction of the Full Model is that all the MoGs of images from the same category can be seen as the same original MoG, which has undergone the addition of noise to its parameters. The reduction process can then be seen as aiming to find the true original values of these parameters. Fig. 4 shows the Full Model and the Unscented Reduced Models of different orders for several image categories.

Note that if we apply EM algorithm directly on the pixels of all the images in a given category to learn a simplified category model, all the pixels in the same (x, y) coordinates are considered as spatially similar and there is no notion of pixels that are all belonging to the same image. The hierarchical grouping, which is based on MoG simplification, avoids this problem by first separately clustering all the pixels in the same image. This is apart from the computational complexity gain.

We compare category modeling based on the Full Model and on reducing the number of Gaussians in each category model to 20, 15, 10, and 5. Following the model-generation step, a classification experiment is performed on the testing set. The distance between each image MoG and each category model is computed, and the images are classified in the category for which this distance is minimal. We use the two KL-approximation methods GMA and UTA to classify a test image based on the distance between the image MoG and a category MoG. Each classification experiment is repeated 10 times, each time training and testing with different images. The results presented show the mean of those 10 runs. The classification results are presented in Fig. 5. In the top row (Figs. 5a and 5b), 10 percent of the data is used for training, and in the bottom row (Figs. 5c and 5d), 20 percent is used. In the left column (Figs. 5a and 5c), the UTA distance measure is used for classification, and in the right column (Figs. 5b and 5d), the GMA distance measure is used for classification.

Fig. 5 clearly indicates that the best classification results are obtained when using the UTA both for model learning and for classification. Another important observation from Fig. 5 is that the Full Model does not provide the best results. The reduced model, therefore, is not just a technical step used to overcome the computational complexity of large models. It is also a learning step applied to the models of the category images to obtain an improved category model. The results of the classification experiments also underline the importance of the reduced model's size. One can see that the classification score is improved by all of the considered reductions, but this improvement is limited and in our experiment (see Fig. 5a) the limit was reached for 10 components. Further reduction could not improve the classification score. Another interesting observation is that the worst combination is simplifying the mixture model using the UTAC algorithm and classifying using the GMA distance metric. In this case, a strict mismatch exists between training and the test models where in the test models, unlike training models, sigma points from the same Gaussian are forced to be in the same group.

6 Conclusion

In this study, we have proposed an efficient probabilistic categorization method. We specifically address the problem of intractability of an MoG representation based on huge number of components. We introduce an algorithmic technique for learning an optimal reduced version of the original MoG. We have also demonstrated the applicability of the clustering algorithm to efficient representation and retrieval of medical image categories. This technique can be used for situations other than image categorization, such as robot path planning, nonlinear dynamical systems, and speech analysis. Several problems are left for future research. There are several ways to select the sigma points that are used in the UT (e.g., [11]), and given an MoG simplification problem, one can search for a variant that yields the best approximation. Another research problem (that exists in every clustering algorithm) is finding the optimal number of components in the reduced model.

Appendix A

In this appendix, we extend the EM algorithm for training MoG models to the case of weighted likelihood functions. Let $x_1, \ldots, x_n$ be n IID samples from the density function $f(x; \theta)$ where θ is an unknown parameter set of a mixture model (and the discrete mixture random variable is hidden). Denoting the hidden variable by y, we obtain $f(x) = \sum_y f(y, x)$. In the standard parameter-estimation case, we look for a maximum-likelihood estimation of θ, namely we want to find

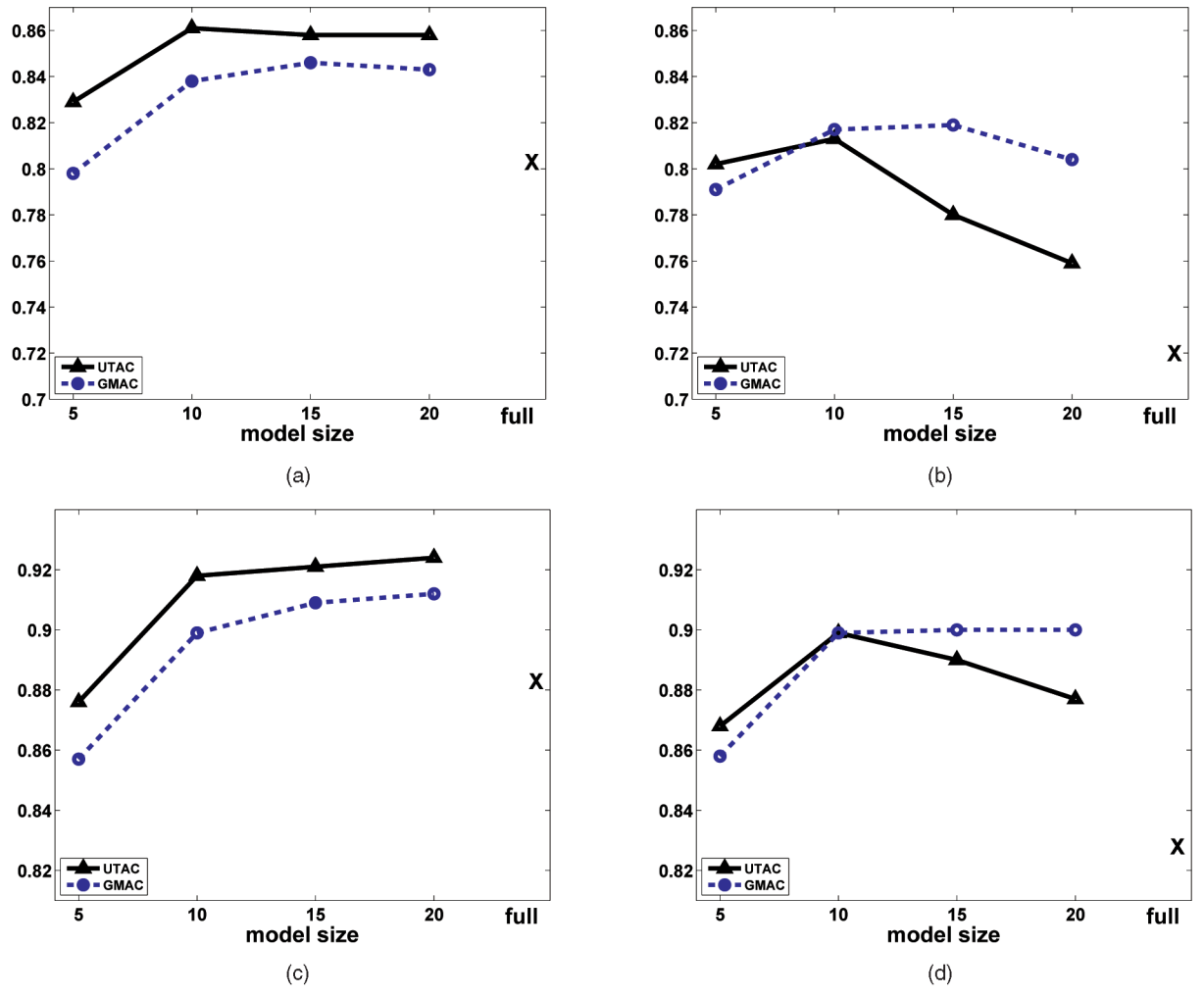

Fig. 5. Comparison of the two reduced models UTAC and GMAC on a real medical image. The graphs show the classification results of the two reduced models as a function of the size of the reduced model. In (a) and (b), 10 percent of the data is used for training, and in (c) and (d), it is 20 percent. In (a) and (c), the UTA is used for classification, and in (b) and (d), the GMA is used for classification.

$$\theta_{ML} = \arg\max_{\theta} \sum_{i=1}^{n} \log f(x_i;\theta).$$

Assume that we are also given n positive numbers $w_1,\ldots,w_n$ and the target is to find θ that maximizes the following expression:

$$\sum_{i=1}^{n} w_i \log f(x_i;\theta). \tag{16}$$

Expression (16) has no strict probabilistic meaning. It can be viewed as a weighted log-likelihood function, where w_i signifies the importance or the weight of the data point x_i. The weighted likelihood concept has been previously developed for a variety of purposes, and it offers a simple alternative to Bayesian statistics. The approach combines all relevant prior information through a weighted version of the likelihood function. The most popular application is weighted least square regression [10], [21].

In this appendix, we show that the EM theory can be extended to the case of weighted likelihood functions. We follow the Neal and Hinton interpretation of the EM algorithm as a coordinate ascent of the free-energy function [17].

For every set of probability distributions $q_i(y)$, $i=1,\ldots,n$ defined on the ith hidden mixture random variable, we define the weighted free energy as follows:

$$FE(q,\theta) = \sum_i w_i \sum_y q_i(y) \log f(y,x_i;\theta) + \sum_i w_i H(q_i), \tag{17}$$

where H is the entropy function. The Jensen inequality implies that for every $i=1,\ldots,n$:

$$\log f(x_i) \geq \sum_y q_i(y) \log f(y,x_i) + H(q_i). \tag{18}$$

Multiplying each side of (18) by the *positive* number w_i and summing over all the indices i we obtain

$$\sum_{i=1}^{n} w_i \log f(x_i;\theta) \geq FE(q,\theta). \tag{19}$$

It can be easily verified that

$$\sum_{i=1}^{n} w_i \log f(x_i;\theta) = FE(q,\theta) + \sum_i w_i D(q_i(y)\|f(y|x_i)), \tag{20}$$

where D is the KL divergence. Hence, taking $q_i(y) = f(y|x_i)$, (19) turns into an equality. Therefore,

$$\arg\max_{\theta} \sum_{i=1}^{n} w_i \log f(x_i;\theta) = \arg\max_{\theta} \max_{q} FE(q,\theta).$$

The EM algorithm approximates the double maximization problem via alternating maximization. We have, therefore, shown that

there is a monotone improvement of the weighted likelihood function during the EM iterations.

APPENDIX B

In this appendix, we derive an explicit relation between the mixture simplification algorithm presented in this paper, namely the UTAC and a previously suggested algorithm called the GMAC algorithm [7]. Let $f = \sum_{i=1}^{n} \alpha_i f_i$ be a d-dimensional mixture model such that $f_i = N(\mu_i, \Sigma_i)$. Let $x_{i1}, \ldots, x_{i2d}$ be the sigma points of f_i. The UTAC algorithm is essentially a standard EM algorithm applied on the $2dn$ sigma points to find the maximum weighted-likelihood simplified MoG.

Next, we derive the equations of a modified version of the UTAC algorithm, where all the sigma points of the same Gaussian component are constrained to be assigned to the same simplified Gaussian component. The weighted free energy of a simplified mixture $g = \sum_{j=1}^{m} \beta_j g_j$ according to the constrained model is

$$FE(q) = \sum_i \alpha_i H(q_i) + \sum_{ij} \alpha_i q_i(j) \left(\log \beta_j + \sum_k \log g_j(x_{ik}) \right). \quad (21)$$

The EM algorithm derived from the weighted free-energy function (21) is the following. The E-step is

$$w_{ij} = \frac{\beta_j \prod_{k=1}^{2d} g_j(x_{ik})}{\sum_l \beta_l \prod_{k=1}^{2d} g_l(x_{ik})}. \quad (22)$$

w_{ij} is the weight of the soft assignment of f_i to g_j. Note that since the sigma points precisely capture the first two moments of the Gaussian component f_i:

$$D(f_i \| g_j) = \frac{1}{2d} \sum_{k=1}^{2d} \log \frac{f_i(x_{ik})}{g_j(x_{ik})}. \quad (23)$$

Equation (23) can be rewritten in the following form:

$$\log \left(\prod_{k=1}^{2d} g_j(x_{ik}) \right) = -2dD(f_i \| g_j) + \mathrm{c} \quad (24)$$

such that the c is a function of f_i and is the same for all $j = 1, \ldots, m$. Hence, substituting (24) in (22) we obtain

$$w_{ij} = \frac{\beta_j e^{-2dD(f_i \| g_j)}}{\sum_l \beta_l e^{-2dD(f_i \| g_l)}}, \quad (25)$$

which is exactly the E-step of the soft version of the GMAC algorithm [18] where λ is set to be $2d$. The M-step consists of reestimating the parameters of the simplified model g. Let μ'_j and Σ'_j be the parameters of the jth component of j. Then, the M-step is

$$\mu'_j = \frac{\sum_i \alpha_i w_{ij} \sum_k x_{ik}}{2d \sum_i \alpha_i w_{ij}} = \frac{\sum_i \alpha_i w_{ij} \mu_i}{\sum_i \alpha_i w_{ij}}. \quad (26)$$

Using a similar method, we can obtain an updated expression for Σ'_j. The M-step therefore reestimates the Gaussian g_j as a collapsed version of $\frac{\sum_i w_{ij} \alpha_i f_i}{\sum_i w_{ij} \alpha_i}$ into a single Gaussian. This again coincides with the M-step of the Gaussian matching method [7], [18].

REFERENCES

[1] Y. Bar-Shalom and X. Li, *Estimation and Tracking: Principles, Techniques and Software.* Artech House, 1993.

[2] D. Comaniciu and P. Meer, "Mean Shift: A Robust Approach toward Feature Space Analysis," *IEEE Trans. Pattern Analysis and Machine Intelligence*, pp. 603-619, 2002.

[3] J. Davis and I. Dhillon, "Differential Entropic Clustering of Multivariate Gaussians," *Proc. 20th Ann. Conf. Neural Information Processing Systems (NIPS)*, 2006.

[4] J. Goldberger, H. Greenspan, and J. Dreyfuss, "An Optimal Reduced Representation of a MoG with Applications to Medical Image Database Classification," *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2007.

[5] J. Goldberger, H. Greenspan, and S. Gordon, "An Efficient Similarity Measure Based on Approximations of KL-Divergence between Two Gaussian Mixtures," *Proc. Ninth IEEE Int'l Conf. Computer Vision (ICCV)*, 2003.

[6] J. Goldberger, H. Greenspan, and S. Gordon, "Unsupervised Image-Set Clustering Using an Information Theoretic Framework," *IEEE Trans. Pattern Analysis and Machine Intelligence*, pp. 449-458, 2006.

[7] J. Goldberger and S. Roweis, "Hierarchical Clustering of Mixture Model," *Proc. 18th Ann. Conf. Neural Information Processing Systems (NIPS)*, 2004.

[8] H. Greenspan and A. Pinhas, "Medical Image Categorization and Retrieval for PACS Using the GMM-KL Framework," *IEEE Trans. Information Technology in Biomedicine*, pp. 190-202, 2007.

[9] B. Han, D. Comaniciu, Y. Zhu, and L. Davis, "Incremental Density Approximation and Kernel-Based Bayesian Filtering for Object Tracking," *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2004.

[10] F. Hu, "The Asymptotic Properties of Relevance Weighted Likelihood Estimations," *Canadian J. Statistics*, pp. 45-60, 1997.

[11] S. Julier, "The Scaled Unscented Transformation," *Proc. Am. Control Conf. (ACC '02)*, pp. 4555-4559, 2002.

[12] S. Julier and J.K. Uhlmann, "Unscented Filtering and Nonlinear Estimation," *Proc. IEEE*, pp. 401-422, 2004.

[13] S. Julier, J.K. Uhlmann, and H.F. Durrant-Whyte, "New Method for the Nonlinear Transformation of Means and Covariances in Filters and Estimators," *IEEE Trans. Automatic Control*, pp. 477-482, 2000.

[14] B. Kurkoski and J. Dauwels, "Message-Passing Decoding of Lattices Using Gaussian Mixtures," *Proc. 30th Symp. Information Theory and Its Applications (SITA '07)*, pp. 877-882, 2007.

[15] T.M. Lehmann, M. Guld, T. Deselaers, D. Keysers, H. Schubert, K. Spitzer, H. Ney, and B.B. Wein, "Automatic Categorization of Medical Images for Content-Based Retrieval and Data Mining," *Computerized Medical Imaging and Graphics*, pp. 143-155, 2005.

[16] T.M. Lehmann, M. Guld, O. Thies, B. Fisher, K. Spitzer, D. Keysers, H. Ney, M. Kohnen, H. Schubert, and B.B. Wein, "Content-Based Image Retrieval in Medical Applications," *Methods of Information in Medicine*, pp. 354-361, 2004.

[17] R. Neal and G. Hinton, "A View of the EM Algorithm That Justifies Incremental, Sparse, and Other Variants," *Learning in Graphical Models*, In M.I. Jordan, MIT Press, pp. 355-368, 1999.

[18] N. Petrovic, A. Ivanovic, N. Jojic, S. Basu, and T. Huang, "Recursive Estimation of Generative Models of Video," *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2006.

[19] E.B. Sudderth, A. Torralbs, W.T. Freeman, and A.S. Willsky, "Describing Visual Scenes Using Transformed Dirichlet Processes," *Proc. 20th Ann. Conf. Neural Information Processing Systems (NIPS)*, 2006.

[20] N. Vasconcelos, "On the Complexity of Probabilistic Image Retrieval," *Proc. Eighth Int'l Conf. Computer Vision (ICCV)*, 2001.

[21] X. Wang, C. van Eeden, and J. Zidek, "Asymptotic Properties of Maximum Weighted Likelihood Estimators," *J. Statistical Planning and Inference*, pp. 37-54, 2004.

[22] K. Zhang and J.T. Kwok, "Simplifying Mixture Models through Function Approximation," *Proc. 20th Ann. Conf. Neural Information Processing Systems (NIPS)*, 2006.