An Optimal Reduced Representation of a MoG with Applications to Medical Image Database Classification

Jacob Goldberger
School of Engineering
Bar-Ilan University, Israel
goldbej@eng.biu.ac.il

Hayit Greenspan
The Engineering Faculty
Tel-Aviv University, Israel
hayit@eng.tau.ac.il

Jeremie Dreyfuss
The Engineering Faculty
Tel-Aviv University, Israel
jeremie_dreyfuss@walla.com

Abstract

This work focuses on a general framework for image categorization, classification and retrieval that may be appropriate for medical image archives. The proposed methodology is comprised of a continuous and probabilistic image representation scheme using Gaussian mixture modeling (MoG) along with information-theoretic image matching measures (KL). A category model is obtained by learning a reduced model from all the images in the category. We propose a novel algorithm for learning a reduced representation of a MoG, that is based on the Unscented-Transform. The superiority of the proposed method is validated on both simulation experiments and categorization of a real medical image database.

1. Introduction

The explosion in the last ten years of digital medical imaging techniques has led to a dramatic increase in the number of images that are acquired every day in any modern hospital. These images must be archived in the patients' personal file in a way that allows easy access when needed. That is why more and more hospitals purchase picture archiving and communication systems (PACS) to navigate through their rapidly growing databases. Lund university hospital (Sweden) for example, produces 15,000 new x-ray images per day and has them all available for access from any workstation in the hospital (StorageTek - case study)[1]. Content-based image retrieval (CBIR) is playing an increasing role in a wide range of clinical processes [10]. For the clinical decision-making it can be useful to refer to images of the same modality or the same anatomic region in order to identify certain pathologies. Moreover, the use of meta-data and alpha-numerical information coupled with CBIR can help the medical staff to decide on the most efficient course of action. The field of medical CBIR is still in its first steps. Most current systems propose solutions for images of specific modality or specific organ such as spine x-ray [1] or mammography [9]. A few systems can be found that focus on general medical image categorization and retrieval such as MedGIFT [11] and COBRA [3].

In content-based search, the goal is to retrieve the most-similar images to a query image introduced to the system. The images belonging to the query-image *category* are the ones we wish to retrieve. Every CBIR system is based on three phases: feature extraction, effective indexing, and a retrieval system. In the following we will focus exclusively on the problem of image indexing. This phase is crucial in order to ensure an efficient retrieval, which is as important requirement as the precision of the retrieval. In Zhang et al. [14], SVM are trained for different categories of images using a training set with known class labels. The SVMs are next used to categorize new images as they enter into the database. Zhang et al. used SVM to model low-level features (texture and shape descriptors) of the training data. The SVM they used is modified to fit a K-nearest-neighbor (KNN) classification scheme. This method has the advantage of being more efficient than regular SVMs. Another approach for efficient classification is through dimensionality reduction. One such approach is the Karhunen-Loeve Transform (KLT) that can be calculated based on the approximation proposed by Chandrasekaran et al. [2] using low-rank singular value decomposition (SVD).

Our approach is based on the probabilistic and continuous framework for supervised image category modeling and matching that was proposed by Greenspan and Pinhas [6]. Each image or image-set (category) is represented as a MoG distribution. Images (categories) are compared and matched via a probabilistic measure of similarity between distributions known as the Kullback-Leibler (KL) distance. Representing a large image set with a mixture of Gaussians may lead to a very complex model (with a large number of mixture parameters). In this paper we propose a more compact model for the image set representation (a reduced MoG). We introduce an efficient way for learning this model that

[1] http://www.sun.com/storagetek/success-stories/

1-4244-1180-7/07/$25.00 ©2007 IEEE

can be incorporated into various classification algorithms. The rest of the paper is organized as follows. The problem of clustering the MoG components and previous attempts to solve it, are reviewed in Section 2. Sections 3 and 4 present a novel clustering based on the Unscented-Transform. Section 5 shows a simulation demonstrating that the proposed method outperforms previously suggested methods. Finally, an application of the proposed clustering algorithm to efficient categorization of a real medical image database is presented in section 6.

2. The Clustering Task

Assume that we are given a mixture density f composed of n d-dimensional Gaussian components:

$$f(y) = \sum_{i=1}^{n} \alpha_i N(y; \mu_i, \Sigma_i) = \sum_{i=1}^{n} \alpha_i f_i(y) \qquad (1)$$

We want to cluster the components of 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 MoG(m), one way to formalize the goal of clustering is to say that we wish to find the element g of MoG(m) "closest" to f under some distance measure.

A common proximity criterion is the cross-entropy from f to g, i.e. $\hat{g} = \arg\min_g KL(f||g) = \arg\max_g \int f \log g$, where $KL()$ is the Kullback-Leibler divergence and the minimization is performed over all g in MoG(m). 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.

Zhang and Kwok [15] proposed a mixture clustering algorithm based on the l_2 norm between the mixture models. Goldberger and Roweis [5] suggested a mixture clustering algorithm based on grouping the mixture components. Soft versions of that clustering algorithm appear in [12, 13]. This matching based method approximates well the KL-divergence if the Gaussian elements are far apart. However, if there is a significant overlap between the Gaussian elements, then the assignment of a single component of $g(x)$ to each component of $f(x)$ becomes less accurate. To handle overlapping situations we propose a novel reduced representation of a MoG based on the Unscented transform. Goldberger et al. [4] showed that we can utilize the Unscented transform mechanism to obtain a good approximation for the KL-divergence between two MoGs. In this study we show that this approximation can be used to cluster the MoG components. The Unscented transform and the KL-approximation that is based on it, are reviewed in the next section.

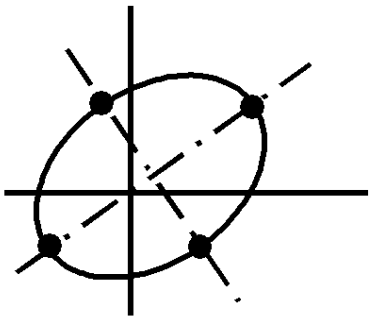

Figure 1. The sigma points of the Unscented transform

3. The Unscented Transform

The Unscented transformation is a method for calculating the statistics of a random variable which undergoes a non-linear transformation [7]. It is successfully used for nonlinear filtering. The Unscented Kalman filter (UKF) [8] 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 non-linear function, the UKF uses the true nonlinear function and approximates the distribution of the function output. Following [4] we show how we can utilize the Unscented transform mechanism to obtain an approximation for the KL-divergence between two MoGs.

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

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

such that $(\sqrt{\Sigma})_k$ is the k-th column of the matrix square root of Σ. Let $UDU^\top$ be the singular value decomposition of Σ, such that $U = \{U_1, ..., U_d\}$ and $D = \text{diag}\{\lambda_1, ..., \lambda_d\}$ then $(\sqrt{\Sigma})_k = \sqrt{\lambda_k} U_k$. These sample points completely capture the true mean and variance of the normal distribution $f(x)$ (see Figure 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:

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

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 then 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.

The Unscented transform can be used to approximate the KL-divergence between the following two MoGs:

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

Since $KL(f||g) = \int f \log f - \int f \log g$, 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 non-linear function $\log g(x)$ which can be approximated using the Unscented transform. Hence:

$$\int f \log g \approx \frac{1}{2d} \sum_{i=1}^{n} \alpha_i \sum_{k=1}^{2d} \log g(x_{i,k}) \tag{3}$$

such that:

$$\begin{aligned} x_{i,k} &= \mu_i + (\sqrt{d\Sigma_i})_k \qquad k = 1, ..., d, \\ x_{i,d+k} &= \mu_i - (\sqrt{d\Sigma_i})_k \qquad k = 1, ..., d. \end{aligned} \tag{4}$$

To simplify notations we denote the Unscented-Transform-Approximation (3) by:

$$UTA(f, g) = \frac{1}{2d} \sum_{i=1}^{n} \alpha_i \sum_{k=1}^{2d} \log g(x_{i,k}) \tag{5}$$

If the covariance matrices of the two MoG are restricted to be diagonal the computational complexity of the Unscented approximation is significantly reduced. Assume the covariance matrices of the components of f have the following form:

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

then the sigma points are simply:

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

An alternative approximation for the KL distance between two MoGs f and g is based on matching a Gaussian from g to each component of f. The formula of the approximation, which we dub Gaussian-Match-Approximation (GMA), is:

$$GMA(f, g) = \sum_{i=1}^{n} \alpha_i \max_j \int f_i \log g_j \tag{6}$$

the GMA proximity measure can be used to derive a reduced representation of a given MoG f as follows [5]. The MoG $\hat{g}$ is an optimal reduced approximation for f if:

$$\hat{g} = \arg\max \text{GMA}(f, g) \tag{7}$$

such that the maximization is performed over all $g \in$ MoG(m). The UTA is known to better approximate the KL distance between two MoGs than the GMA [5, 12]. Utilizing the GMA proximity measure as a criterion for obtaining a reduced representation of a given MoG, motivates using the UTA as a cost function in order to obtain a better reduced approximation. The algorithm derived from this reasoning is presented in the next section.

4. The Unscented Transform based Clustering

Given a 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 a m-component MoG g that maximizes the expression:

$$UTA(f, g) = \frac{1}{2d} \sum_{i=1}^{n} \alpha_i \sum_{k=1}^{2d} \log g(x_{i,k}) \tag{8}$$

where $x_{i,k}$ are the sigma-points of f. To optimize this equation we can consider the sigma-points as a deterministic weighted set of samples from f. The free-energy function in this case is:

$$-FE(q) = \sum_{i} \alpha_i H(q_{ik}) + \sum_{ijk} \alpha_i q_{ik}(j) \log(\beta_j g_j(x_{i,k})) \tag{9}$$

where q_{ik} is a discrete distribution on the m components of g and H is the entropy function. Minimizing the free-energy yields an iteration of the EM algorithm for learning the reduced model. The E-step is:

$$w_{ikj} = \frac{\beta_j g_j(x_{ik})}{g(x_{ik})} \tag{10}$$

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

$$\begin{aligned} \beta_j &= \frac{1}{2d} \sum_{ik} \alpha_i w_{ikj} \\ \mu'_j &= \frac{\sum_{ik} \alpha_i w_{ikj} x_{ik}}{\sum_{ik} \alpha_i w_{ikj}} \\ \Sigma'_j &= \frac{\sum_{ik} \alpha_i w_{ikj} (x_{ik} - \mu_j)(x_{ik} - \mu_j)^\top}{\sum_{ik} \alpha_i w_{ikj}} \end{aligned} \tag{11}$$

where μ'_j and Σ'_j are the updated parameters of the j-th component of the reduced MoG g. From the general EM-algorithm theory, it can be verified that the expression $UTA(f,g)$ is monotonically increasing during the EM iterations. Hence at the convergence point we obtain a reduced model that is (locally) optimal according of the UTA proximity measure.

The Unscented Clustering Algorithm:

input: $f = \sum_{i=1}^{n} \alpha_i N(\mu_i, \Sigma_i)$ and m

output: $g = \sum_{j=1}^{m} \beta_j N(\mu'_i, \Sigma'_i)$

such that UTA(f,g) is maximal.

E-step: $w_{ikj} = \frac{\beta_j g_j(x_{ik})}{g(x_{ik})}$ s.t. $x_{i,k} = \mu_i \pm (\sqrt{d\Sigma_i})_k$

M-step:

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

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

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

For comparison, in the algorithms presented in [5, 12], the E-step equation (10) is:

$$w_{ij} = \frac{\beta_j e^{-\lambda KL(f_i||g_j)}}{\sum_l \beta_l e^{-\lambda KL(f_i||g_l)}} \tag{12}$$

such that in [12] λ is set to 1 and in [5] λ is set to ∞. In the M-step of [5, 12], 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.

5. Simulation Results

In order to compare the quality of the proposed Unscented-based approximation as well as its processing efficiency we conducted the following simulation experiment. In each session we sample a random mixture of many Gaussians f and we search for a reduced representation of f using a mixture of a small number of components. The original mixture models were randomly sampled according to the following rules. The number of Gaussians in the original 2-dimensional MoG f was 20. We search for an optimal reduced model that is composed of 5 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.

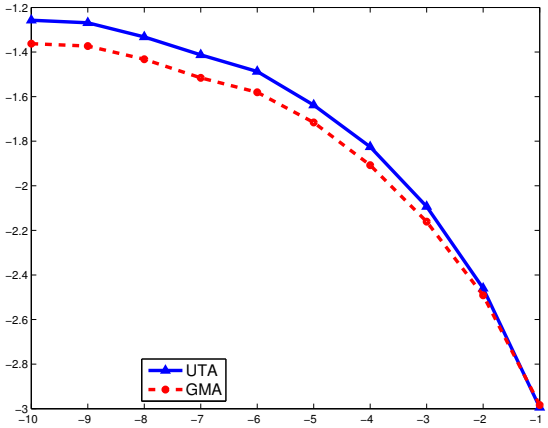

Figure 2. A comparison between two reduced-model algorithms on simulation data. The first reduced model is the Unscented-Based (UTA) and the second is matched based (GMA). The graph shows the (Monte-Carlo) cross-entropy of the original model and the reduced model, as a function of ϵ on a logarithmic scale.

In addition to the method proposed in this paper, we have also implemented the matching-based learning method suggest by Goldberger and Roweis [5] (see equation (6)). Another important issue is how to asses the quality of the approximation obtained from the learning methods. It was validated several times [4, 12] that the distance measure based on the Unscented-Transform (see Section 2) is the best method to measure the distance between two MoGs in terms of accuracy and computational complexity. It is obvious that the reduced model based on the Unscented-Transform is best approximating the original model using the Unscented-Transform distance since it was chosen exactly to optimize this criterion. Instead we measure the approximation quality based on a Monte-carlo simulation (based on 10000 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 1000 times for each ϵ. Figure 2 shows the simulation results. It shows the KL score as a function of $\log_2(\epsilon)$. As can be seen, better approximation results were obtained using the Unscented-Transform based reduction algorithm.

6. Experimental Results

To evaluate the two different reduction methods, we have used MoGs computed from a set of 1502 medical images which were pre-labeled by an expert as belonging to 21 different categories, two of which being MRI images, two other CT images and the rest digital radiographies. 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 [6], a five-dimensional feature space is used to represent the images, including intensity, texture (contrast, scale) and position (x,y). An unsupervised clustering done by an Expectation-Maximization (EM) algorithm is used to compute the MoGs of each image, thus giving a compact representation of homogenous regions in the feature space. This representation can be easily rendered by replacing each pixel in the image by the mean intensity value of the Gaussian it has been clustered to. One can see from Figure 3 where several examples of images and their MoGs are shown that the present modeling of images still maintains their visual content.

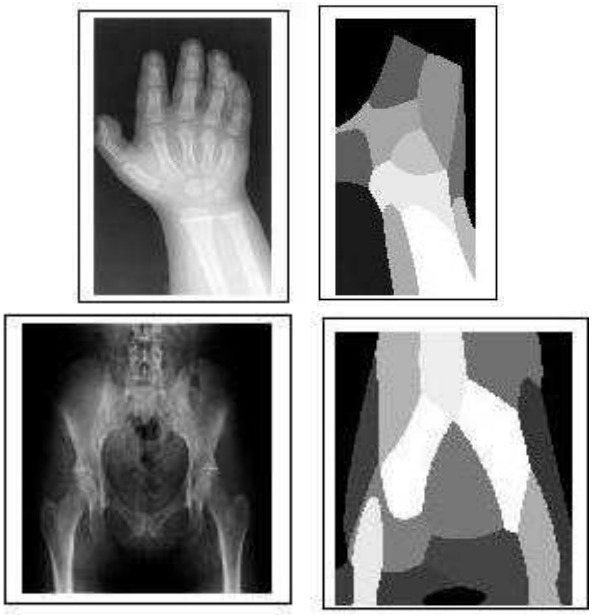

Figure 3. Examples of images (left) and their MoG modeling (right). The model is shown via segmentation of the image pixels based on the MoG.

In order to test our reduction methods, the images were divided such that 70% of the images in each category serve as a training set and the remaining 30% 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 calculated by merging together all the MoGs from the same category. The merging is done simply by summing together all the MoGs in a category and normalizing each Gaussian's weight in the mixture by the number of images in the category. Afterwards, both the Match Reduction and Unscented Reduction algorithms are applied to the Full-Model in order to create reduced MoGs with less Gaussians. The 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 finding the true original values of these parameters. One of the most difficult problems being of course that there is no way of knowing what the order of the original MoG was. This is the motivation for selecting the order of the reduced models as a certain percentage of the Full-Model's order. Figure 4 shows the Full-Model and the Unscented Reduced Models of different orders for several image categories.

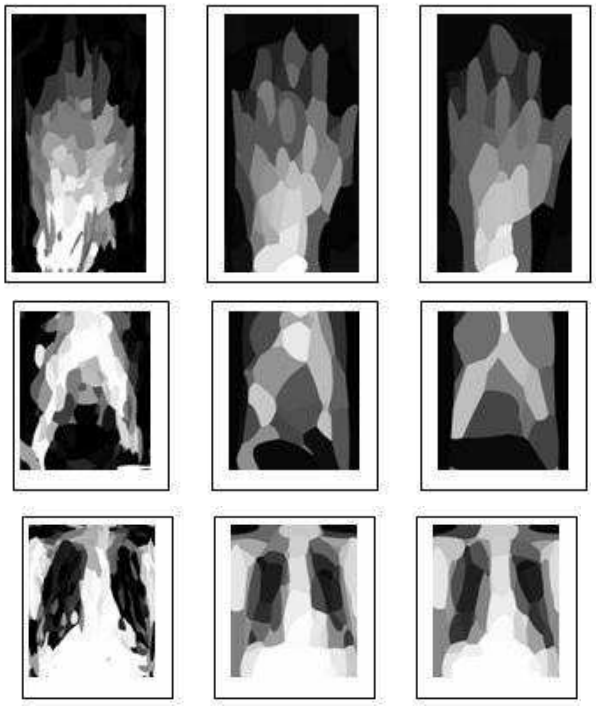

Figure 4. Full-Model and Reduced Models with 10%, 5% of the Full-Model's size (left to right) for the "Arm", "Pelvis" and "Chest posterior anterior" categories (top to bottom).

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 to the category for which this distance is minimal. For the reasons that were explained in Section 3, the most appropriate method to use, in order to compare the MoGs, in the approximation of the KL distance based on the Unscented Transform [4]. The 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 Figure 5. We compare category modeling based on the Full-Model and on reducing the number of Gaussians to 3%, 10% ,20% and 25% of the original number of Gaussians in each category. From Figure 5 we see clearly that the best classification results are obtained when using the UTA. Another impor-

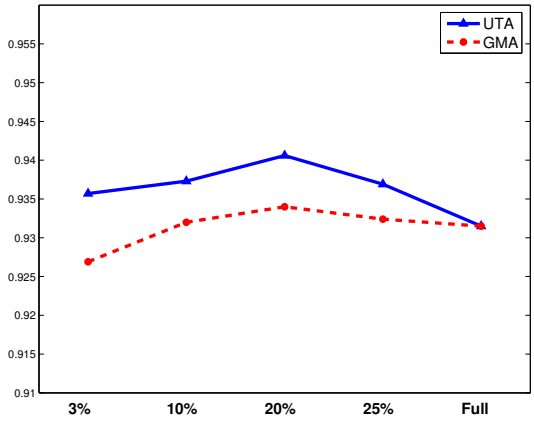

Figure 5. Comparison between two reduced models (real medical data). The first reduced model is the Unscented-Based (UTA) and the second is matched based (MGA). The graph shows the classification results of the two reduced models using the Unscented transform distance measure.

tant observation from Figure 5 is that the Full-Model does not provide the best results. The reduced model, therefore, is not just a technical step to overcome the computational-complexity of large models. It is also a learning step that is applied on the models of the category-images to obtain an improved category model. The results of the classification experiments also underline the importance in the choice of the reduced model's size. One can see that the classification score is improved by any of the considered reductions, but this improvement is limited and in our experiment the limit was reached for a reduction to 20% of the original size. After that, the more drastic was the reduction the less it improved the classification score. If the reduction level was chosen independently for every category it's fair to assume that a higher classification score could be reached.

7. Conclusion

In this work we presented a clustering algorithm for efficient representation of medical image categories. We specifically addresses the problem of intractability of a MoG representation based on huge number of components. We introduce an algorithmic technique for learning an optimal collapsed version of the original MoG. This technique can be used for other situations than image categorization, such as robot path planning, non-linear dynamical systems and speech analysis.

References

[1] S. Antani, J. Cheng, J. Long, L. Rodney Long, and G. Thoma. Medical validation and cbir of spine x-ray images over the Internet. *Proc. SPIE Int. Soc. Opt. Eng.*, 2006. 1

[2] S. Chandrasekaran, B. Manjunath, Y. Wang, J. Winkeler, and H. Zhang. An eigenspace update algorithm for image analysis. *Graphical Models and Image Processing: GMIP*, 1997. 1

[3] E. A. El-Kwae, H. Xu, and M. R. Kabuka. Content-based retrieval in picture archiving and communication systems. *Journal of Digital Imaging*, 2000. 1

[4] J. Goldberger, H. Greenspan, and S. Gordon. An efficient similarity measure based on approximations of KL-divergence between two Gaussian mixtures. *International Conference on Computer Vision (ICCV)*, 2003. 2, 4, 5

[5] J. Goldberger and S. Roweis. Hierarchical clustering of mixture model. *In Neural Information Processing Systems, (NIPS)*, 2004. 2, 3, 4

[6] H. Greenspan and A. Pinhas. Medical image categorization and retrieval for pacs using the GMM-KL framework. *IEEE Trans. on Info. Technology in BioMedicine*, 2006. 1, 5

[7] S. Julier and J. K. Uhlmann. A general method for approximating nonlinear transfromations of probability distributions. *Technical report, RRG, Dept. of Engineering Science, University of Oxford*, 1996. 2

[8] S. Julier and J. K. Uhlmann. A new extension of the Kalman filter to non-linear systems. *Proc of AeroSense: The 11th International Symposium on Aerospace/Defence Sensing, Simulation and Control*, 1997. 2

[9] P. Korn, N. Sidiropoulos, C. Faloutsos, E. Siegel, and Z. Protopapas. Fast and effective retrieval of medical tumor shapes. *IEEE Trans. on Knowledge and Data Engineering*, 1998. 1

[10] H. Muller, N. Michoux, D. Bandon, and A. Geissbuhler. A review of content-based image retrieval systems in medical application - clinical benefits and future directions. *International Journal of Medical Informatics*, 2004. 1

[11] H. Muller, A. Rosset, J. Vallee, and A. Geissbuhler. Comparing feature sets for content-based medical information retrieval. *SPIE Medical Imaging*, 2004. 1

[12] N. Petrovic, A. Ivanovic, N. Jojic, S. Basu, and T. Huang. Recursive estimation of generative models of video. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2006. 2, 3, 4

[13] N. Vasconcelos. On the complexity of probabilistic image retrieval. *International Conference on Computer Vision (ICCV)*, 2001. 2

[14] H. Zhang, A. Berg, M. Maire, and J. Malik. SVM-KNN: Discriminative nearest neighbor classification for visual category recognition. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2006. 1

[15] K. Zhang and J. T. Kwok. Simplifying mixture models through function approximation. *In Neural Information Processing Systems, (NIPS)*, 2006. 2