# STATISTICAL CATEGORIZATION OF HUMAN HISTOLOGICAL IMAGES

*Dehua Zhao[1], Yixin Chen[1], and Hernan Correa[2]*

[1]Department of Computer Science, University of New Orleans, New Orleans
[2]Department of Pathology, Louisiana State University Health Sciences Center, New Orleans
[1]{dzhao,yixin}@cs.uno.edu, [2]hcorre@lsuhsc.edu

## ABSTRACT

Histology is the science of understanding the structure of animals and plants, and studying the functional implications of biological structures. In this paper, we propose a statistical modeling approach to human histological image categorization. Texture features of the images are characterized by localized Gabor filters. The probabilistic distribution of the texture patterns from each category is approximated by a finite Gaussian mixture model. Expectation maximization (EM) procedure and minimum message length (MML) principle are used to perform density estimation and model selection, respectively. Component-wise EM and weak component annihilation are applied to avoid the drawbacks of the standard EM. Experimental validation is provided based on images from different organs and parts of the body.

## 1. INTRODUCTION

Histology is the science of understanding the structure of animals and plants, and studying the functional implications of biological structures [8]. The knowledge of biological structures and their functions at the subcellular, cellular, tissue and organ levels is central to the understanding of mechanisms of disease and drug action. Therefore, histology provides a scientific foundation for clinical research, education, and practice. In order to make cell and tissue structure visible for microscopic examination, a specimen needs to go through several preparation steps and is mounted on a glass slide. Histology slides contain information of the body from the subcellular through organ levels. Interpreting histology slides requires expert knowledge and observational skills. The work is time consuming and human errors can never be completely avoided. Nevertheless, with proper training, a medical student can identify the structural features of specific cell types, tissues and organs from microscopic images, and understand structure-function relationships at different levels. Is it possible to train a computer program to interpret histology slides? This is the question we attempt to answer in this paper. In particular, we investigate statistical modeling based classification of human histological images.

Organizing images into semantically meaningful categories will not only help archive images but also be extremely useful for subsequent annotation and retrieval tasks [2]. There have been an abundance of prior work on general image categorization, especially within content-based image retrieval (CBIR). Readers are referred to [9] for a detailed review. Most existing CBIR methods may fail when directly extended to the medical image context; since "common CBIR-systems only have a rudimentary understanding of image content, and within such systems there is no distinction between important and unimportant features or between multiple objects in the image" [4]. Clinically useful information within medical images is in essence highly localized and varies drastically. Images obtained from different imaging modalities could be quite different, therefore, require different treatments. A review on medical image CBIR was given by Müller et al. [6]. Compared with other types of medical images, such as, computer tomography (CT), magnetic resonance image (MRI), X-rays, and ultrasound (US), the research on histological images is relatively rare. In I-Browse [10], Tang et al. extract features from image blocks of various size, and relate image blocks to a set of histological (semantic) labels. The semantic analyzer uses these labels to determine the category of the image based on domain knowledge. The performance of the system largely depends on domain knowledge and reasoning logic.

Histologists distinguish histological images through patterns that are defined by local structural elements. Figure 1 shows images from 10 categories corresponding to 10 human organs. We can see that histological images are essentially composed of texture patches. Patches from the same category follow certain patterns. However, these patterns usually vary from one image to another and even within the same image. In addition, there are some random optical noises due to intensity variance, focus of lens, and color variations. Therefore, a successful categorization system should be able to characterize these distinct patterns and, at the same time, overcome random variations and noises. In this paper, we propose a statistical modeling approach to histological image categorization. Multi-channel Gabor filters are used to extract texture features from subimages in a histological image. It is assumed that texture patterns from images of the same category follow the same distribution, which is defined by a Gaussian mixture probability density function. The expectation maximization (EM) procedure is used to estimate model parameters. Component-wise EM and weak component annihilation are used to avoid drawbacks of standard EM. Minimum message length (MML) model selection is used to choose the optimal model.

The remainder of the paper is organized as follows. Section 2 describes the feature extraction process. Section 3 presents Gaussian mixture model and component-wise EM parameter estimation. Section 4 discusses the model selection scheme. Experimental results and discussions are given in Section 5. Finally, we conclude in Section 6 along with a discussion of future work.

## 2. FEATURE EXTRACTION

Although color provides useful information for visual perception, it may not be desirable to directly use color as features for his-
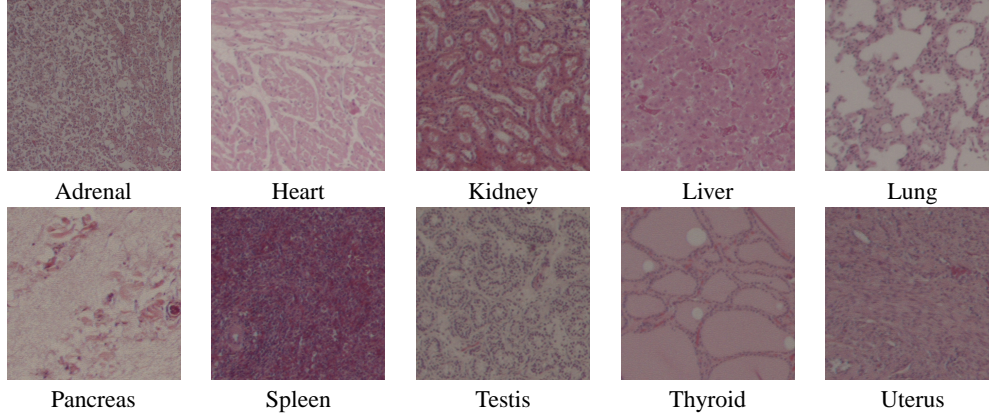
Figure 1: Histological images from 10 categories

tological image classification. This is because the color of histological image depends on the stained material. Since histologists can interpret gray scale images, we convert the image to LUV color space and use the L component for classification. Histogram equalization is then applied to the L component to eliminate luminance variances.

Histological images are composed of texture patches. There are a bunch of literatures devoted to texture analysis. Most successful and commonly used methods include: co-occurrence matrix, Laws filter masks, Gabor filters, Markov random fields, and fractals. Readers are referred to [7] for a comparative study of these methods. Gabor texture features become more favorable for several reasons. First, Gabor filters based on Gabor function allow one to choose arbitrary scale and orientation. Textural images are usually distinguishable with scale and orientation features. Second, biological evidence shows that the receptive field of organization of simple cells resembles the profile of Gabor element functions. Third, Gabor filters achieve minimum uncertainties in both spatial and frequency domain. Gabor filters have been successfully used in the texture image retrieval [5].

A 2D Gabor function consists of a modulating Gaussian shape envelop and a complex sinusoidal carrier:

$$g(x, y) = \left( \frac{1}{2\pi\sigma_x\sigma_y} \right) \exp^{-\frac{1}{2}\left( \frac{x^2}{\sigma_x^2} + \frac{y_2}{\sigma_y^2} \right)} \exp^{2\pi j(U+V)}$$

where $\sigma_x$ and $\sigma_y$ define the spreads of Gaussian envelop in x and y direction, respectively; $U$ and $V$ define the spatial frequency of the sinusoidal in Cartesian coordinates. Radial center frequency $F = \sqrt{U^2 + V^2}$ is usually adopted to measure the frequency in $cycles/pixel$.

Self-similar Gabor filters are generated through scaling ($s$) and orientation ($n$):

$$g_{sn}(x, y) = a^{-s} g(x^{'}, y^{'})$$

where $a = (F_h/F_l)^{\frac{1}{S-1}}$, $\theta = n\pi/O$, $x^{'} = a^{-s}(x\cos\theta + y\sin\theta)$, and $y^{'} = a^{-s}(-x\sin\theta + y\cos\theta)$. $F_h$ and $F_l$ are upper and lower center frequencies of interest, respectively. Parameters $S$ and $O$ are the total number of scales and total number orientations, respectively. Gabor filters comprising the Gabor filter bank are designed in a way that the half-magnitude supports of the filter responses in frequency domain touch each other [5].

Texture patches in a histological image are not homogenous. It's difficult to use global Gabor features to categorize histological images. Instead, a histological image is divided into sub-images or blocks. We assume that texture patterns within a block are homogenous, and features calculated from Gabor filter responses can characterize these patterns. For each block of size $M \times N$, mean $\mu_{sn}$ and standard deviation $\sigma_{sn}$ are calculated from Gabor responses corresponding to the filter with scale $s$ and orientation $n$. Therefore, a histological image is represented with finite number of feature vectors each corresponding to one block:

$$(\mu_{00}, \sigma_{00}, \mu_{01}, \sigma_{01}, \cdots, \mu_{S-1O-1}, \sigma_{S-1O-1}).$$

## 3. GAUSSIAN MIXTURE MODEL

A Gaussian mixture model is composed of several Gaussian components with different mixing weights or probabilities. A Gaussian mixture distribution with $k$ components can be written as

$$P(x|\Theta) = \sum_{k=1}^{K} p_k P(x|\mu_k, \Sigma_k) \tag{1}$$

with the constraint $\sum_{k=1}^{K} p_k = 1$. $x$ is the feature vector representing a realization of the random variables. $\Theta$ is the parameter set $(\mu_1, \Sigma_1, \mu_2, \Sigma_2, \cdots, \mu_K, \Sigma_K; p_1, p_2, \cdots, p_K)$. $p_k$ and $P(x|\mu_k, \Sigma_k)$ are the mixing probability and the probability density function of the $k$-th component, respectively.

A block in a histological image is assumed to be generated by one of these Gaussian components with the corresponding mixing probability. Blocks from the same category are assumed to be independently generated by a Gaussian mixture model, so the marginal probability of an image is the product of probabilities of all blocks in the image. For a testing image, the likelihood of individual internal block generated by the model of one category is calculated using formula (1). The likelihood of the image generated by the model is just the product of the likelihoods of all blocks in the image. The image is assigned to the category with the maximal likelihood.

A Gaussian mixture model treats data samples as realizations of continuous random variables, thus avoids quantization process as in discrete statistical models. The training process of one category is independent to the data of other categories. Therefore our design has very good scalability.

## 4. PARAMETER ESTIMATION AND MODEL SELECTION

The training process is to find the model that most probably generates data. For Gaussian mixtures, the estimation is usually solved with iterative Expectation Maximization (EM) procedure. The number of components is determined with minimum message length (MML [3]) principle.

The iterative EM algorithm first initializes the model parameters, and then breaks down each iteration into two steps: expectation and maximization. Membership probability of each data sample is calculated in E step given model parameters, and model parameters are updated in M-step. The expectation formula is defined as follows:

$$P(z_m^{(i)} = 1|x^{(i)}, \Theta_{old}) =$$

$$\frac{p_{m\_old} P(x^{(i)}|\mu_{m\_old}, \Sigma_{m\_old})}{\sum_{k=1}^{K} p_{k\_old} P(x^{(i)}|\mu_{k\_old}, \Sigma_{k\_old})} \triangleq \omega_m^{(i)} \quad (2)$$

where $z^{(i)}$ is the missing label corresponding to sample $x^{(i)}$. $z^{(i)}$ has K components taking values of 0 and 1, only one component can be 1. It tells which component generates $x^{(i)}$. $\Theta_{old}$ is the initial parameter set or the parameter set updated in previous iteration. The formula is nothing but Bayesian formula calculating posterior probability of the sample data $x^{(i)}$ being generated by component $m$ given the model parameters. Formulae to update model parameters are listed below, where $n$ is total number of instances in the training data set $x$.

$$p_{m\_new} = \frac{\sum_{i=1}^{n} \omega_m^{(i)}}{\sum_{k=1}^{K} \sum_{i=1}^{n} \omega_k^{(i)}}, \quad (3)$$

$$\mu_{m\_new} = \left(\sum_{i=1}^{n} \omega_m^{(i)}\right)^{-1} \sum_{i=1}^{n} x^{(i)} \omega_m^{(i)}, \quad (4)$$

$$\Sigma_{m\_new} = \left(\sum_{i=1}^{n} \omega_m^{(i)}\right)^{-1} \sum_{i=1}^{n} (x^{(i)} - \mu_{m\_new})(x^{(i)} - \mu_{m\_new})^T \omega_m^{(i)}. \quad (5)$$

The major drawback with the standard EM algorithm is its convergence to local maximum. Component-wise EM [1] can to certain extent avoid this drawback. Component-wise EM [1] algorithm differs from the standard EM in that only one component rather than all components gets updated at one iteration while the likelihood being kept in the ascending trend, so some local maxima can be avoided. Another problem with the standard EM algorithm as discussed in [3] is that the learned covariance matrices of some components tend to be very close to singular matrices especially when the number of components is larger than the underlying true value. This makes the learned model meaningless. Eliminating these so called *weak components* can avoid this problem. The weak component is the one with almost 0 mixing probability, meaning that it is not supported by the data, thus being annihilated. With the combination of component-wise EM algorithm and weak component annihilation, drawbacks of the standard EM algorithm can, to certain extent, be avoided.

Another important issue is to choose an appropriate number of Gaussian components. With too many components, the model is likely to overfit the training data. On the other hand, a model with very few components may not be able to accurately approximate the underling density function. In our work, Minimum Message Length (MML) principle is adopted to select the number of mixture components. MML captures both data and model complexity. The message length $L$ is expressed as

$$L(\Theta, \mathcal{D}) = \frac{N}{2} \sum_{m=1}^{K} \log(\frac{np_m}{12}) + \frac{K}{2} \log \frac{n}{12} +$$

$$\frac{K(N+1)}{2} - \log P(\mathcal{D}|\Theta) \quad (6)$$

where $\mathcal{D}$ is the training data set containing $n$ training samples, and each sample is a vector of dimension $d$. $K$ is the number of components, and $N$ is the number of parameters specifying each component. In our case, $N = d + d * (d+1)/2$. $p_m$ is the mixing probability of the $m^{th}$ component. The term $\log P(\mathcal{D}|\Theta)$ is the pure data driven part representing the likelihood that the model generates $\mathcal{D}$. Other terms describe the model complexity.

The following algorithm describes the training process. The number of components takes from $K_{min}$ to $K_{max}$. k-means clustering algorithm is applied to set the initial model parameters. The difference of model message lengths of neighboring iterations is used as the termination criterion. The model with the minimum message length is selected.

**Algorithm 1** *Training Algorithm*

**Input:** Data Vector $X_{n \times d}$
$\quad\quad K_{min}, K_{max}, L_{min} = +\infty, \delta_p, \delta_L$
**Output:** $\Theta_{opt}, K_{opt}$
1 **FOR** $K = K_{min}, \cdots, K_{max}$
2 $\quad$ set $\Theta = Kmeans(X, K)$
3 $\quad$ set $L_2 = +\infty$
4 $\quad$ set $K_1 = K$
5 $\quad$ **REPEAT**
6 $\quad\quad$ set $L_1 = L_2$
7 $\quad\quad$ **FOR** $l = 1, \cdots, K_1$
8 $\quad\quad\quad$ calculate $\omega_m^{(i)}$ where (2) where
$\quad\quad\quad i = 1, \cdots, n$ and $m = 1, \cdots, K_1$
9 $\quad\quad\quad$ update $p_k$ with (3) where $k = 1, \cdots, K_1$
10 $\quad\quad\quad$ **IF** $p_l > \delta_p$
11 $\quad\quad\quad\quad$ update $\mu_l$ and $\Sigma_l$ with (4) and (5)
12 $\quad\quad\quad$ **ELSE**
13 $\quad\quad\quad\quad$ set $p_l = 0$
14 $\quad\quad\quad\quad$ swap $p_l$ and $p_{K_1}$
15 $\quad\quad\quad\quad$ swap $\mu_l$ and $\mu_{K_1}$
16 $\quad\quad\quad\quad$ swap $\Sigma_l$ and $\Sigma_{K_1}$
17 $\quad\quad\quad\quad$ set $K_1 = K_1 - 1$
18 $\quad\quad\quad\quad$ **IF** $K_1 < K_{min}$
19 $\quad\quad\quad\quad\quad$ break repeat loop
20 $\quad\quad\quad\quad$ **END IF**
21 $\quad\quad\quad$ **END IF**
22 $\quad\quad$ **END FOR**
23 $\quad\quad$ calculate $L_2$ with (6)
24 $\quad$ **UNTIL** $|L_1 - L_2| < \delta_L$
25 $\quad$ **IF** $L_2 < L_{min}$
26 $\quad\quad$ set $L_{min} = L_2$
27 $\quad\quad$ set $K_{opt} = K_1$
28 $\quad$ **END IF**
29 $\quad$ set $\Theta_{opt} =$ nonzero components in $\Theta$
30 **END FOR**

| Histological Images | | | Classification Accuracy | |
|---|---|---|---|---|
| ID | Cate. Name | Num. of Images | Training | Testing |
| C1 | adrenal | 25 | 0.800 | 0.480 |
| C2 | heart | 114 | 0.844 | 0.808 |
| C3 | kidney | 20 | 0.825 | 0.650 |
| C4 | liver | 107 | 0.848 | 0.813 |
| C5 | lung | 288 | 0.902 | 0.872 |
| C6 | pancreas | 120 | 0.656 | 0.608 |
| C7 | spleen | 18 | 0.930 | 0.933 |
| C8 | testis | 25 | 0.750 | 0.440 |
| C9 | thyroid | 39 | 0.952 | 0.900 |
| C10 | uterus | 22 | 0.920 | 0.730 |

Table 1: The category labels of the images and 5-fold crossvalidation results.

## 5. EXPERIMENTAL RESULTS AND DISCUSSIONS

To validate the methods we have described, we implemented the histological image categorization system and tested with 778 histological images from 10 categories as shown in table 1. These images are of $40\times$ magnification (4 objective lens $\times$ 10 ocular lens) stored in JPEG format with size $3072 \times 3840$. To reduce computational cost, each image is downsampled to $1536 \times 1920$. The block size is $64 \times 64$. The Gabor filter bank consists of filters of 3 scales and 6 orientations. Five fold crossvalidation is conducted on these histological images. The average train classification accuracy is 84.3%, and the average testing accuracy is 72.3%. The results are also shown in table 1.

Several interesting observations are worthy mentioning. First, even though the number of Gaussian components are chosen from 3 to 9 during the training process, all categories take value of 3, 4 or 5; and most of them take 4 as optimal. As we know that all organs, no matter how complicated its functionalities are, are composed of 4 basic tissue types (connective, epithelial, nerve, and muscle). The observation suggests that Gaussian mixture model can to certain extent reflect underlying biological compositions and that the learning results during categorization stage can be extremely useful in subsequent annotation tasks. Second, many images of C1 (adrenal) are misclassified as C6 (pancreas), and vice versa. This observation suggests that the designed classifier doesn't classify these categories. The interesting thing is that these two categories are glands. This suggests that the performance may be improved by arranging classifiers in hierarchy. For example, an organ can be first classified as *gland* versus *not gland*, and further divided into detailed categories. The last observation is that the misclassification rates of the boundary images are much higher than that of the non-boundary images. The boundary images are usually taken from the boundary of the slide or the boundary of the organ. Most of the misclassified boundary images contain regions from other organs.

## 6. SUMMARY AND FUTURE WORK

In this paper, a statistical modeling approach to histological image categorization is proposed. Texture patterns of the images are characterized by localized Gabor features. Finite Gaussian mixture model is used to approximate the distribution of texture patterns from the same category. The well-known expectation maximiza-

tion (EM) procedure and MML principle are used to perform parameter estimation and model selection, respectively. Component-wise EM and weak component annihilation are used to avoid drawback of standard EM. Experiment shows promising results.

The current system has several limitations. First, we assume that blocks in a histological image are identically and independently distributed. This assumption is somewhat restrictive. The model that captures spatial correlation of subimages may increase the categorization accuracy. Second, current study focus on histological images of single magnification. Test on histological images over a range of magnifications is needed to show the robustness of the method. We leave these to future work. We will also add more images and images from new categories to test the performance and scalability.

## 7. REFERENCES

[1] G. Celeux, S. Chretien, F. Forbes, and A. Mkhadri, "A Component-Wise EM Algorithm for Mixtures," *Technical Report 3764, INRIA, Rhone Alpes, France*, 1999.

[2] Y. Chen and J. Z. Wang, "Image categorization by learning and reasoning with regions," *Journal of Machine Learning Research*, vol. 5, pp. 913–939, 2004.

[3] M. A. T. Figueiredo and A. K. Jain, "Unsupervised Learning of Finite Mixture Models," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 24, no. 3 pp. 381–396, 2002.

[4] T. Lehmann, B. Wein, J. Bredno, F. Vogelsang, and M. Kohnen, "Content-based Retrieval in Medical Applications: A Novel Multi-step Approach," *Proceddings SPIE*, 3972, pp. 312-320, 2000.

[5] B. S. Manjunth and W. Y. Ma, "Texture Features for Browsing and Retrieval of Image Data," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 18, no. 8, pp. 837–842, 1996.

[6] H. Müller, N. Michoux, D. Bandon and A. Geissbuhler, "A Review of Content-Based Image Retrieval Systems in Medical Applications - Clinical Benefits and Future Directions," *Computer Vision and Image Understanding*, vol. 77, pp. 111–132, 1999.

[7] T. Randen and J. H. Husoy, "Filtering for texture classification: a comparative study," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 21, no. 4, pp. 291–310, 1999.

[8] M. H. Ross, G. I. Kaye, and W. Pawlina, *Histology: A Text and Atlas*, 4th edition, Lippincott Williams & Wilkins, 2002.

[9] A. W.M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-based image retrieval at the end of the early years," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 22, no. 12, pp. 1348–1381, 2000.

[10] L. H. Tang, R. Hanka, and H. H.S. Ip, "Histological image retrieval based on semantic content analysis," *IEEE trans. Information Technology in Biomedicine*, vol. 7, no. 1, pp. 26–36, 2003.