Relaxation Labeling for Cell Phase Identification

Dat T. Tran, Senior Member, IEEE, Tuan D. Pham, Senior Member, IEEE

Abstract—Gaussian mixture model (GMM) is used in cell phase identification to model the distribution of cell feature vectors. The model parameters, which are mean vectors, covariance matrices and mixture weights, are trained in an unsupervised learning method using the expectation maximization algorithm. Experiments have shown that the GMM is an effective method capable of achieving high identification rate. However, the GMM approach is not always effective because of ambiguity inherently existing in the cell phase data. To enhance the effectiveness of the GMM for solving this specific problem, the relaxation labeling (RL) is proposed to be used with the GMM. The RL algorithm is a parallel algorithm that updates the probabilities of cell phases by using correlation or mutual information between cell phases to reduce uncertainty among GMMs having overlapping properties.

I. INTRODUCTION

High content screening by automated fluorescence microscopy is becoming one of the most widely used research tools to assist scientists in understanding the complex process of cell division and mitosis [8]-[10]. Its power comes from the sensitivity and resolution of automated light microscopy with multi-well plates, combined with the availability of fluorescent probes that are attached to specific sub-cellular components, such as chromosomes and microtubules, for visualization of cell division or mitosis using standard epi-fluorescence microscopy techniques [30]. By employing a carefully selected reporter probes and filters, fluorescence microscopy allows specific imaging of phenotypes of essentially any cell component [17]. With these probes we can determine both the amount of a cell component, and most critically, its distribution within the cell relative to other components. Typically, 3-4 different components are localized in the same cell using probes that excite at different wavelengths. Any change in cell physiology would cause a redistribution of one or more cellular components, and this redistribution provides a certain cytological marker that allows for scoring of the physiological change.

An essential task for high content screening is to measure cell cycle progression (inter phase, prophase, metaphase, and telophase) in individual cells as a function of time. Cell cycle progress can be identified by measuring nuclear changes.

Automated time-lapse fluorescence microscopy imaging provides an important method for the observation and study of cellular nuclei in a dynamic fashion [12, 14]. Stages of an automated cellular imaging analysis consist of segmentation, feature extraction, identification, and tracking of individual cells in a dynamic cellular population; and the identification of cell phases is considered the most difficult task of such analysis [5].

In time-lapse microscopy images are usually captured in a time interval of more than 10 minutes. During this period dividing nuclei may move far away from each other and daughter cell nuclei may not overlap with their parents. Given the advanced fluorescent imaging technology, there still remain technical challenges in processing and analyzing large volumes of images generated by time-lapse microscopy.

The increasing quantity and complexity of image data from dynamic microscopy renders manual analysis unreasonably time-consuming. Therefore, automatic techniques for analysing cell-cycle progress are of considerable interest in the drug discovery process.

Being motivated by the desire to study drug effects on HeLa cells, an ovarian cancer cell line, we have applied several computational techniques for identifying individual cell phase changes during a period of time. To extract useful features for the cell-phase identification task, the image segmentation of large image sequences acquired by time-lapse microscopy is necessary. The extracted data can then be used to analyse cell phase changes under drug influence. Segmenting nuclei in time-lapse microscope can be performed by various methods such as thresholding, region growing, or edge detection [16]. Most of these algorithms take into account either the morphological information or the intensity information of the image. Problems may arise when trying to segment touching nuclei because it is very difficult to define the boundary of each individual nuclear. Watershed techniques can be used to segment touching objects [1]. To deal with the over-segmentation problem a post process is needed to merge the fragments. Umesh and Chaudhuri [29] used a connectivity based merging method to merge a tiny cell fragment with a nearby cell if it shares the maximum boundary with that cell. These authors applied their method on a set of 327 cells and a 98% correct segmentation result was reported. This method can only merge small fragments and fails if the size of cell fragments is above a preset value. The bigger fragments are considered as cell by this method. Bleau and Leon [1] used an iterative trial and test approach to merge small regions with their nearby larger regions based on a set of volume, depth, and surface criteria. These authors applied their method to segment the vesicles in live cells;
However no experimental results were reported.

To automate the process of classifying cellular phases using time-lapse fluorescence microscopic image sequences, we first apply a shape-and-size based method which merges the over-segmented nuclear fragments. Secondly we extract useful features to discriminate the shapes and intensities of different image cell phases. We then use these image features to train phase models to classify cancer cells at different phases. Since Gaussian mixture models (GMMs) are effective models with high identification accuracy and are widely used for pattern recognition, we therefore apply the GMM method to train phase models.

However the GMM approach is not always effective because the ambiguity inherently existing in the labeling of cell phase feature vectors is treated in an inflexible way by its deterministic rules. Basing our motivation on this reason, we propose an improved algorithm over the cell phase-based GMM approach using the relaxation labeling (RL) in which the deterministic identification of the GMM-based approach is only an initial process of the probabilistic labeling. Results from this initial labeling will then be updated using the RL technique until a convergence is reached. The RL technique was first introduced by Rosenfeld et al. [26] to solve problems in pattern recognition. It is a parallel algorithm that updates the probabilities of labels or classes by using interactive information between unknown objects with respect to the reference labels to reduce uncertainty among labels having interchanging properties. Since being first developed to tackle problems in image analysis, its flexible framework has attracted many researchers in the broad field of pattern recognition. In speaker recognition, the RL approach has been originally applied for the vector quantization (VQ) method by Pham et al. [23, 24]. The results using several codebook sizes obtained from the RL approach are more favorable than those from the VQ method.

In this paper we apply the framework of relaxation labeling to dealing with the GMM-based cell phase identification. The rest of this paper is organized as follows. In Section 2, we briefly review the GMM method. We then describe the procedures of RL in Section 3. The combination of the GMM and the RL for cell phase identification is presented in Section 4. Experiments are given in Section 5 to compare the performance of the proposed approach with the conventional GMM method.

II. GAUSSIAN MIXTURE MODEL

This section presents the GMM method for cell phase identification. Parameter estimation equations for training phase models are presented first. The GMM method for phase identification is then described as a maximum likelihood classifier.

A. The GMM algorithm

Let \( X = \{x_1, x_2, \ldots, x_T\} \) be a set of \( T \) vectors, each of which is a \( d \)-dimensional feature vector extracted by digital signal processing. Since the distribution of these vectors is unknown, it is approximately modeled by a mixture of Gaussian densities as the weighted sum of \( c \) component densities, given by the equation

\[
p(x_t | \lambda) = \sum_{i=1}^{c} w_i N(x_t, \mu_i, \Sigma_i)
\]

where \( \lambda \) denotes a prototype consisting of a set of model parameters \( \lambda = \{w_i, \mu_i, \Sigma_i\} \), \( w_i \), \( i = 1, \ldots, c \), are mixture weights, and \( N(x_t, \mu_i, \Sigma_i) \), \( i = 1, \ldots, c \), are the \( d \)-variate Gaussian component densities with mean vectors \( \mu_i \) and covariance matrices \( \Sigma_i \)

\[
N(x_t, \mu_i, \Sigma_i) = \frac{1}{(2\pi)^{d/2} |\Sigma_i|^{1/2}} \exp\left(-\frac{1}{2}(x_t - \mu_i)^T \Sigma_i^{-1} (x_t - \mu_i)\right)
\]

To train the GMM, these parameters are estimated such that they best match the distribution of the training vectors. The most widely used training method is the maximum likelihood (ML) estimation. For a set of training vectors \( X \), the likelihood of the GMM is

\[
p(X | \lambda) = \prod_{t=1}^{T} p(x_t | \lambda)
\]

The aim of ML estimation is to find a new parameter model \( \lambda \) such that \( p(X | \lambda) \geq p(X | \bar{\lambda}) \). Since the expression in (3) is a nonlinear function of parameters in \( \lambda \), its direct maximization is not possible. However, parameters can be obtained iteratively using the expectation-maximization (EM) algorithm [6]. An auxiliary function \( Q \) is used

\[
Q(\lambda, \lambda) = \sum_{i=1}^{c} \sum_{t=1}^{T} \log[p(i | x_t, \lambda)] = \sum_{i=1}^{c} \sum_{t=1}^{T} \log[w_i N(x_t, \mu_i, \Sigma_i)]
\]

where \( p(i | x_t, \lambda) \) is the a posteriori probability for acoustic class \( i \), \( i = 1, \ldots, c \) and satisfies

\[
p(i | x_t, \lambda) = \frac{w_i N(x_t, \mu_i, \Sigma_i)}{\sum_{k=1}^{c} w_k N(x_t, \mu_k, \Sigma_k)}
\]

and

\[
\sum_{i=1}^{c} p(i | x_t, \lambda) = 1
\]

The basis of the EM algorithm is that if \( Q(\lambda, \lambda) \geq Q(\bar{\lambda}, \bar{\lambda}) \) then \( p(X | \lambda) \geq p(X | \bar{\lambda}) \) [13]. Setting derivatives of the \( Q \)
function with respect to $\lambda$ to zero, the following re-estimation formulas are found

\[
\begin{align*}
\overline{w}_i &= \frac{1}{T} \sum_{t=1}^{T} p(i | x_t, \lambda) \\
\mu_i &= \frac{1}{\sum_{t=1}^{T} p(i | x_t, \lambda)} \sum_{t=1}^{T} p(i | x_t, \lambda)x_t \\
\Sigma_i &= \frac{1}{\sum_{t=1}^{T} p(i | x_t, \lambda)} \sum_{t=1}^{T} p(i | x_t, \lambda)(x_t - \mu_i)(x_t - \mu_i)' 
\end{align*}
\]  

(7) (8) (9)

The algorithm for training the GMM is described as follows

1. Generate the \textit{a posteriori} probability $p(i | x_t, \lambda)$ at random satisfying (6)
2. Compute the mixture weight, the mean vector, and the covariance matrix following (7), (8) and (9)
3. Update the \textit{a posteriori} probability $p(i | x_t, \lambda)$ according to (5) and compute the $Q$ function using (4)
4. Stop if the increase in the value of the $Q$ function at the current iteration relative to the value of the $Q$ function at the previous iteration is below a chosen threshold, otherwise go to step 2.

\section{The GMM in cell phase identification}

Let $\lambda_k$, $k = 1, \ldots, N$, denote cell phase models of $N$ phases. Given a feature vector $x$, a classifier is designed to classify $x$ into $N$ cell phase models by using $N$ discriminant functions $g_k(x)$, computing the similarities between the unknown $x$ and each phase model $\lambda_k$ and selecting the model $\lambda_k^*$ if

\[
k^* = \arg \max_{1 \leq k \leq N} g_k(x)
\]

(10)

In the minimum-error-rate classifier, the discriminant function is the \textit{a posteriori} probability $g_k(x) = p(\lambda_k | x)$

(11)

Using the Bayes’ rule

\[
p(\lambda_k | x) = \frac{p(\lambda_k)p(x | \lambda_k)}{p(x)}
\]

(12)

and assuming equal likelihood of all phases, i.e., $p(\lambda_k) = 1/N$. Since $p(x)$ is the same for all phase models, the discriminant function in (11) is equivalent to the following [24]

\[
g_k(x) = p(x | \lambda_k)
\]

(13)

Finally, the decision rule used for cell phase identification is as follows

\[
\text{Select phase } k^* \text{ if } k^* = \arg \max_{1 \leq k \leq N} p(x | \lambda_k)
\]

(14)

where $p(x | \lambda_k)$ is given in (1).

\section{III. RELAXATION LABELING}

The RL algorithms consist of four models when first introduced by Rosenfeld et al. [26] - discrete, fuzzy, linear probabilistic, and nonlinear probabilistic models. However, the last model offers the best performance for identification problems and due to this reason, only the nonlinear probabilistic relaxation model is studied here for speaker identification. Let a set of objects $A = \{a_1, a_2, \ldots, a_T\}$ and a set of labels $A = \{\lambda_1, \lambda_2, \ldots, \lambda_N\}$. An initial probability is given to each object $a_i$ having each label $\lambda_k$, which is denoted as $p_i(\lambda_k)$. These probabilities satisfy the following condition

\[
\sum_{k=1}^{N} p_i(\lambda_k) = 1, \forall a_i \in A, \quad 0 \leq p_i(\lambda_k) \leq 1
\]

(15)

The RL updates the probabilities $p_i(\lambda_k)$ in (15) using a set of compatibility coefficients $r_{it}(\lambda_k, \lambda_l)$, where $r_{it}(\lambda_k, \lambda_l) : A \times A \mapsto [-1, 1]$, whose magnitude denotes the strength of compatibility. The meaning of these compatibility coefficients can be interpreted as follows

\[
\begin{array}{ll}
< 0: \lambda_k, \lambda_l \text{ are incompatible for } a_i \text{ and } a_{i'} \\
= 0: \lambda_k, \lambda_l \text{ are independent for } a_i \text{ and } a_{i'} \\
> 0: \lambda_k, \lambda_l \text{ are compatible for } a_i \text{ and } a_{i'}
\end{array}
\]

(16)

For computing the compatibility coefficients, two possible methods employ the concepts of statistical correlation and mutual information. The two methods are based on those developed by Peleg and Rosenfeld [20]. The correlation-based estimate of the compatibility coefficients is defined as

\[
r_{it}(\lambda_k, \lambda_l) = \frac{\sum_{t=1}^{T} [p_t(\lambda_k) - \overline{p}(\lambda_k)] [p_t(\lambda_l) - \overline{p}(\lambda_l)]}{\sigma(\lambda_k)\sigma(\lambda_l)}
\]

(17)

where $p_t(\lambda_l)$ is the probability of $a_{i'}$ having label $\lambda_l$ and $a_{i'}$ are the neighbors of $a_i$, $\overline{p}(\lambda_l)$ is the mean of $p_t(\lambda_l)$ for
all $a_t$, and $\sigma(\lambda_t)$ is the standard deviation of $p_t(\lambda_t)$. To alleviate the effect of dominance among labels, the modified coefficients are

$$r^*_t(\lambda_t, \lambda_t) = [1 - p_t(\lambda_t)][1 - p_t(\lambda_t)]r^*_t(\lambda_t, \lambda_t)$$  \hspace{1cm} (18)

and the mutual-information based estimate of the compatibility coefficients is

$$r^*_t(\lambda_t, \lambda_t) = \log\frac{\sum_{t=1}^T p_t(\lambda_k) p_t(\lambda_l)}{\sum_{t=1}^T p_t(\lambda_k) \sum_{l=1}^T p_t(\lambda_l)}$$  \hspace{1cm} (19)

The compatibility coefficients in (19) must be scaled in order to take values in the range $[-1, 1]$.

The updating factor for the estimate $p_t(\lambda_k)$ at the $m$-th iteration is

$$q_t^{(m)}(\lambda_k) = \sum_{l=1}^T d^{(m)} \left[ \sum_{k=1}^N r_t(\lambda_k, \lambda_l) p_t^{(m)}(\lambda_l) \right]$$  \hspace{1cm} (20)

where $d^{(m)}$ are the parameters that weight the contributions to $a_t$ coming from its neighbors $a_t$, and subject to

$$\sum_{t=1}^T d^{(m)} = 1$$  \hspace{1cm} (21)

The updated probability $p_t^{(m+1)}(\lambda_k)$ for object $a_t$ is given by

$$p_t^{(m+1)}(\lambda_k) = \frac{\sum_{k=1}^N p_t^{(m)}(\lambda_k)[1 + q_t^{(m)}(\lambda_k)]}{\sum_{k=1}^N p_t^{(m)}(\lambda_k)[1 + q_t^{(m)}(\lambda_k)]}$$  \hspace{1cm} (22)

The RL algorithm can be outlined as follows

1. Estimate the initial probabilities for each object satisfying (15)
2. Compute the compatibility coefficients using (18) or (19)
3. Calculate the updating factor defined in (20)
4. Update the probabilities for each object using the updating rule in (22)

Repeat steps 3 and 4 until the change in the probability is less than a chosen threshold or equal to a chosen number of iterations

IV. GMM-RL FOR CELL PHASE IDENTIFICATION

It now becomes clear that for a successful performance of the relaxation process, the initial label probabilities and the compatibility coefficients need to be well determined. Wrong estimates of these parameters will lead to algorithmic instabilities. In the GMM-RL-based cell phase identification, the initial probabilities in the RL are defined as the a posteriori probabilities in (12). Objects are now feature vectors considered in the GMM and labels are phases. The GMM-RL algorithm for speaker identification is stated as follows

1. Estimate the initial probabilities for each phase $\lambda_k$ using the a posteriori probabilities in (12)

$$p_t(\lambda_k) = p(\lambda_k | x_t) = \frac{p(x_t | \lambda_k) p(\lambda_k)}{\sum_{k=1}^N p(x_t | \lambda_k) p(\lambda_k)}$$  \hspace{1cm} (23)

where $p(\lambda_k) = 1/N$ and $p(x_t | \lambda_k)$ is computed as in (1)

2. Compute the compatibility coefficients using (18) or (19)
3. Calculate the updating factor defined in (20), where $d^{(m)} = 1/T$ for simplicity.
4. Update the probabilities for each phase using the updating rule in (22)
5. Repeat steps 3 and 4 until the change in the probability is less than a chosen threshold or equal to a chosen number of iterations

It should be noted that $p_t(\lambda_k)$ obtained from (23) after $m$ iterations is the a posteriori probability used in (11). Therefore, the decision rule for cell phase identification is as follows

$$\text{Select phase } k^* \text{ if}$$

$$k^* = \arg\max_{1 \leq k \leq N} p_t(\lambda_k)$$  \hspace{1cm} (24)

V. EXPERIMENTAL RESULTS

A. Data Set

The data set contains 375841 cells in 892 nuclear sequences provided by the Department of Cell Biology at the Harvard Medical School to evaluate the presented modeling methods for the cell phase identification system. The average number of cells per sequence is 421. Imaging was performed by time-lapse fluorescence microscopy with a time interval of 15 minutes. Two types of sequences were used denoting drug treated and untreated. Cell cycle progress was affected by drug and some or all of the cells in the treated sequences were arrested in metaphase. Cell cycle progress in the untreated sequences was not affected. Cells without drug treatment will usually undergo one division during this period of time.
B. Feature Extraction

After the nuclear segmentation has been performed, it is necessary to perform a morphological closing process on the resulting binary images in order to smooth the nuclear boundaries and fill holes inside the nuclei. These binary images are then used as a mask on applied the original image to arrive at the final segmentation. From this resulting image, features can be extracted. The ultimate goal for feature selection is to assign correct phase to cells via the training of some identification technique. In this work, a set of cell-nuclear features are extracted based on the experience of biologists. To identify the shape and intensity differences between different cell phases, a set of 7 features are extracted. These features include maximum intensity, mean, stand deviation, major axis, minor axis, perimeter, and compactness [5]. Because the feature values have different ranges, the scaling of features was therefore necessary by calculating the z-scores [16]

\[ z_{ij} = \frac{x_{ij} - m_j}{s_j} \]  

(25)

where \( x_{ij} \) is the \( j \)-th feature of the \( i \)-th sequences, \( m_j \) the mean value of all \( T \) cells for feature \( j \); and \( s_j \) the mean absolute deviation, that is

\[ s_j = \frac{1}{T} \sum_{i=1}^{T} |x_{ij} - m_j| \]  

(26)

We then divided the data set into 5 subsets for training 5 GMMs and a subset for identification. Each of the 5 training sets for 5 phases contained 5000 cells, which were extracted from the cell sequences labeled from 590 to 892. The identification set contained sequences labeled from 1 to 589. There were 249,547 cells in this identification set.

C. Initializations and Constraints

The GMMs were initialized as follows. Mixture weights, mean vectors, and covariance matrices were initialized with essentially random choices. Covariance matrices were diagonal, i.e. \( \sigma_k \) were variances. A variance limiting constraint was applied to all GMMs using diagonal covariance matrices [25]. This constraint placed a minimum variance value \( \sigma_{min}^2 = 0.1 \) on elements of all variance vectors in the GMMs in our experiments.

The sequences labeled from 1 to 589 in the identification set were used to calculate the updated probabilities in the RL scheme. The length \( T \) in (19), (20) and (21) is the length of the sequence where the feature vector \( x \) was being identified.

D. Experimental Results

The identification error rates in percentage obtained from the GMM and GMM-RL methods are presented in Figure 1. It can be seen that the GMM-RL yields the better performance in all model sizes which are 4, 8, 16, 32, 64 and 128 Gaussians comparing with the GMM. The lowest identification error rate was 11.74% obtained from the 128-mixture GMMs. The total cells for identification are 249,547 extracted from the cell sequences labeled from 1 to 589.

![Fig. 1. Cell phase identification error rates (in %) for the Gaussian mixture model-Relaxation labeling (GMM-RL) and Gaussian mixture model (GMM) methods.](image)

VI. CONCLUSION

We have presented the combined Gaussian mixture modeling and Relaxation labeling method for cell phase identification. Cell features were used to train the five phase models, which were interphase, prophase, metaphase, anaphase, and arrested metaphase. Phase information in cell sequences was used to update probabilities in the Relaxation labeling scheme. The combined Gaussian mixture and Relaxation labeling method was achieved better identification results comparing with the Gaussian mixture modeling method.

ACKNOWLEDGMENTS

The data set was provided by our collaborator - HCNR Center for Bioinformatics, Harvard Medical School. The second author acknowledges the financial support received from the ARC-DP grant DP0665598.

REFERENCES


[27] D. Tran and T. Pham, "Fuzzy and Markov models for written language verification", WSEAS Conference, Lisbon, Portugal, June 16-18 2005


