ABSTRACT

In this paper, the development of a computer-aided system for the classification of neuroblastoma subtypes in terms of grade of neuroblastic differentiation is reported. The classification is carried out within a multi-resolution framework that follows a coarse-to-fine strategy. A novel segmentation approach using the Fisher-Rao criterion embedded in the generic Expectation-Maximization algorithm is presented. The results from multiple classifiers are aggregated using a classifier combining mechanism that involves voting and weighting priori classifier accuracies. The developed system, when tested on 19,551 image tiles, had the best overall accuracy of 96.89%. Additionally, multi-resolution combined with automated feature selection process resulted in a 34% savings in computation time when compared to a previously developed system. As a result, the performances of this system show promise for the computer-aided pathological assessment of the neuroblastic differentiation in practice.

Index Terms— Neuroblastoma, Multi-resolution, Image Segmentation, Pattern Classification, Classifier Combination

1. INTRODUCTION

Pathological analysis of tissue samples with computer vision and image analysis techniques has been an active research area for years, especially after the introduction of whole-slide digitizers. The focus of these efforts have been on quantitatively measuring and analyzing digital slides for breast cancer [1], cervical cancer [2], colonic mucosa [3] and prostate caner [4]. However, few successful efforts, to our best knowledge, have been made on the computerized classification of neuroblastoma (NB), a cancer mostly occurring in children.

In clinical practice, the prognosis of neuroblastoma is currently carried out by highly trained pathologists familiar with the International Neuroblastoma Classification System developed by Shimada et.al. [5]. In accordance with this system, grade of neuroblastic differentiation is one of the most prominent indicators of the class that each tumor sample should be categorized into. The categories of grades are undifferentiated, poorly differentiated and differentiating classes and these are defined in terms of their pathological characteristics. Typical tissue slide and example images associated with different differentiation grades are shown in Figure 1. Usually, the undifferentiated cases contain small to middle-sized NB cells with thin cytoplasm, none-to-few neurites and round to elongated nuclei. As for poorly differentiated cases, typical rosette patterns are often observed. Good indicators of differentiation class are large nuclei and cytoplasm, and the large ratio of diameter of cell to that of nucleus (typically > 2).

In our previous work [6-7], we have developed a unique segmentation and grade categorization algorithm. Although the classification accuracy of the developed system is relatively good for a small representative test set, the computational costs are prohibitive (4166 seconds for the slide shown in Figure 1(a) with classification accuracy of 96.46%). Additionally, the features for those studies were selected manually and the same features were used at each resolution. A small number of classifiers was trained and test, potentially excluding the contribution of other classifiers. To overcome these shortcomings, in this work, we employ a multi-resolution, multi-classifier approach with an automated feature selection process.

2. SYSTEM OVERVIEW

2.1. Image Acquisition
All images for this study were retrospectively collected from neuroblastoma patients according to an IRB approval. A digital scanner, ScanScope T2 digitizer (Aperio, San Diego, CA), is used to digitize the tissues at 40x magnification after they are stained by the haematoxylin and eosin (H&E)-staining process. Each slide is then compressed at approximately 1:40 compression ratio before they are fed into our grading system. The resulting slide size is typically around 1–2.5GB each.

### 2.2. Multi-resolution Framework

The size of a typical neuroblastoma image size is about 70k × 70k. Due to these relatively huge sizes, the input slide is divided into non-overlapping tiles of sizes 512 × 512. These tiles are then used as the inputs to our grading system rather than processing the whole tissue slide all at a time.

We employed a multi-resolution approach that decomposes every input image tile into multiple lower resolution-level representations. In our tests, a four layered multi-resolution hierarchy is built up with \((512 \times 512), (256 \times 256), (128 \times 128), (64 \times 64)\) as the set of tile sizes from the highest to the lowest resolution respectively. The underlying strategy of this classification system is to evaluate the classification results beginning with the lowest fine image resolution level and stopping at the resolution level where the performance of the classification mechanism satisfies a pre-determined criterion. At each resolution scale, such image analysis steps as image segmentation, feature construction, feature selection and classification are followed. In order to improve the overall classification performance, we used seven different combinations of feature extraction techniques and classifiers in practice. Additionally, we propose a novel two-step classifier combination method that yields the final classification result based on the decision information collected from the group of seven classifiers. The flowchart of the system is illustrated in Figure 2. Furthermore, for training purposes, 129 image tiles are cropped at random from each class of sample slides, hence a total of 387 image tiles for classifier training.

### 3. IMAGE ANALYSIS

#### 3.1. Image Decomposition

In both training and testing phases, the whole procedure begins with the image decomposition in which each image is down-sampled in a way such that the lower resolution image can be used to perfectly reconstruct the bandwidth limited version of the next higher resolution image. Without the loss of generality, let us denote \(I^L\) as the input image tile at the full resolution, where \(L\) is the number of resolution hierarchies. Then, the next lower resolution image \(I^{L-1}\), following the down-sampling process demonstrated in [7], is a non-aliasing version in the spatial frequency domain of the next higher resolution tile.

#### 3.2. Image Segmentation

A new segmentation approach, EMLDA, presented in [6] was applied to each image tile. Briefly, this method uses the Linear Discriminant Analysis as the kernel of the generic EM algorithm and iteratively partitions the image in such a way that the Fisher-Rao criterion is maximized:

\[
J(V^* | \theta) = \max_{V, \theta} \frac{|V^T S_B(\theta)V|}{|V^T S_W(\theta)V|}
\]

where \(J(V | \theta)\) is the Fisher-Rao criterion to be maximized; \(S_B\) and \(S_W\) are the between- and within-class scatter matrices [8]. Furthermore, \(V\) is the projection matrix that maps data into a new feature space while \(\theta\) is the labeling configuration that represents the pattern in which the image is segmented. Both \(V\) and \(\theta\) are computed iteratively in the E- and M-step until \(J(V | \theta)\) converges to a local maximum.

#### 3.3. Feature Construction and Selection

All the features are extracted only from the segmented cytoplasm and neuropil regions since they bear the most discriminant information. Features including the entropy, mean and variance of the range of values within a local neighborhood, and the homogeneity degree of the co-occurrence matrix associated with each pixel in the R, G and B channels are extracted. Similarly, we convert each image from the RGB color space to the LAB* space where the same four features are extracted from each pixel either...
within the cytoplasm or neuropil region. Thereby, a total of 24 features are extracted from each pixel of interest.

Due to the “peaking phenomenon” [8], the choice of a subset of most discriminating features is conducive to improving the classification accuracy. Using a smaller number of features also reduces the overall computational complexity of the system. For feature selection, the Sequential Floating Forward Selection (SFFS) procedure is used [9]. Because of its inherent flexibility of dynamically adding and removing features, the SFFS method can generally converge to a good local maximum. In our tests, no particular preference on the choice of the classifier across all resolution levels is presented. However, each classifier has its own feature regions where it yields the best performance although the global performances of different classifiers may look similar.

3.4. Classification

In our experiments, a pool of seven classifiers or combinations of feature extraction techniques and classifiers, including K-Nearest Neighbor (KNN), Linear Discriminant Analysis (LDA)+KNN, LDA+Nearest Mean (NM), CORRLDA [10]+KNN, CORRLDA+NM, LDA+Bayesian and Support Vector Machine (SVM) with a linear kernel, are integrated into our system [8]. In both testing and training processes, each classifier classifies the data independently.

3.5. Weighted Voting and Resolution Transition Rule

The resulting labels assigned from the seven classifiers are combined by the combiner before the final classification decision is made. The combiner evaluates the outputs of all K classifiers (K=7 in this work) and comes up with a final decision \( \theta' \) that prefers to the decisions reached by the majority of the K classifiers:

\[
\theta' = \operatorname{ArgMax}_{i \in \{1, 2, \ldots, C\}} V(i)
\]

(2)

where \( V(i) \) is the number of votes for the \( i \)th class collected from the K classifiers and \( C \) is the number of classes (\( C=3 \) in this work).

Once the final label is determined by the combiner, we next evaluate whether or not the classification result at the current resolution level is good enough. The confidence degree on the classification result is defined as the sum of weights assigned to all classifiers that have the same decision as the combiner does. Furthermore, the weight associated with each classifier is computed by normalizing the classification accuracies of the K classifiers over the training data using the leave-one-out validation process. Therefore, the weight associated with each classifier indicates the prior classification accuracy and thus the degree of confidence we have on that particular classifier. The accumulated degree of confidence \( S \) is used as a measure of the degree of reliability associated with the classification result at a certain resolution level. In summary, the hypothesis test and the decision rule can be written as:

\[
H_0: \text{classification result is good enough;}
H_1: \text{go to the next higher resolution level for classification;}
\]

Decision Rule = \[ \begin{cases} H_0, & \text{if } S > \text{Threshold} \\ H_1, & \text{otherwise} \end{cases} \]

where:

\[
S = \sum_{\omega \in \{\theta_1, \theta_2, \ldots, \theta_M\}} w^i(\omega) = \sum_{\omega \in \{\theta_1, \theta_2, \ldots, \theta_M\}} \sum_{j=1}^{N} w^i(j)
\]

(3)

In Equation (3), \( w^i(j) \) is the recognition rate of the classifier \( i \) computed from the training data at resolution level \( j \) and \( M \) is the number of classifiers that also decide \( \theta' \), the label determined by the combiner, as the best class label. The set \( \{\theta_1, \theta_2, \ldots, \theta_M\} \) contains all the indices of the classifiers that coincide with the decision of the combiner.

4. RESULTS

In our system, all experiments are carried out on a Linux cluster consisting of 64 nodes, each of which has dual 2.4 GHz Opteron 250 processors, 8 GB of RAM and two 250 GB SATA drives installed. In our experiments, 32 out of the 64 nodes are used. By dividing the image into non-overlapping tiles, we took advantage of parallel computing in this cluster.

Classification results of a typical undifferentiated case, shown in Figure 1(a), with different threshold settings (i.e. different confidence level configurations) are shown in Table 1, where D, scaled by 0.1, in the first column represent the thresholds when resolution level is escalated from 1 to 2, 2 to 3 and 3 to 4. The entries in the last column, in addition, represent the numbers of “background” tiles, ones containing no structure of interest, detected using the technique reported in [7].

Table 1. Classification results of a typical undifferentiated case with different threshold configurations, where L represents the resolution level; Acc is the classification accuracy given the ground truth; T is the time cost.

<table>
<thead>
<tr>
<th>D (x 0.1)</th>
<th>T (sec)</th>
<th>Acc (%)</th>
<th>#L1</th>
<th>#L2</th>
<th>#L3</th>
<th>#L4</th>
<th>#Bg</th>
</tr>
</thead>
<tbody>
<tr>
<td>7 8 9</td>
<td>1738</td>
<td>95.01</td>
<td>3797</td>
<td>134</td>
<td>36</td>
<td>80</td>
<td>10569</td>
</tr>
<tr>
<td>9 8 7</td>
<td>1749</td>
<td>95.85</td>
<td>2761</td>
<td>761</td>
<td>429</td>
<td>93</td>
<td>10572</td>
</tr>
<tr>
<td>8 8 8</td>
<td>1705</td>
<td>95.38</td>
<td>3405</td>
<td>368</td>
<td>167</td>
<td>105</td>
<td>10571</td>
</tr>
<tr>
<td>9 9 9</td>
<td>2591</td>
<td>96.66</td>
<td>2772</td>
<td>407</td>
<td>255</td>
<td>610</td>
<td>10572</td>
</tr>
<tr>
<td>10 10 10</td>
<td>2760</td>
<td>96.89</td>
<td>0</td>
<td>2182</td>
<td>797</td>
<td>1066</td>
<td>10571</td>
</tr>
</tbody>
</table>

Of the all threshold sets, it can be concluded that, in general, the classification performances associated with the higher resolution levels are better than those of the lower resolution ones. However, better classification accuracies are obtained at the cost of longer computational time expenditures. In addition, the multi-resolution classification system demonstrates a good robustness since the
classification accuracies corresponding to different threshold configurations are always maintained at a satisfying level, comparable to the ones (97.67%, 98.45%, 98.45% and 98.97% from the lowest to the highest resolution levels) over the training data. Furthermore, the time costs are also reduced by 34% as compared to those of the previous system. The visualized version of the classification results presented in Table 1 is shown in Figure 3, where the classification maps as well as the level maps, i.e. images indicating the resolution level at which the final decision on each image tile is made, are demonstrated.

Each pixel in images shown in Figure 3 represents the outcome from a $512 \times 512$ image tile extracted from the original neuroblastoma slide as large as $75607 \times 68481$. Thus, 19,551 tiles are processed in this case. Color blue, cyan and yellow assigned to each pixel in the images on the left column represent undifferentiated, poorly-differentiated and differentiating classes respectively while the background pixel is shown in white. Within each image in the right column, the brightest intensity represents either level 1 (i.e. the lowest resolution) or the background while the darkest regions indicate those areas where the classification decisions are made on the level 4 (i.e. the full resolution).

5. CONCLUSIONS

This study demonstrates an automated system that classifies neuroblastoma according to the grade of differentiation within a multi-resolution framework. An automated feature selection algorithm (SFFS) considerably reduces the computational cost (66% of that of a previously developed system) while maintaining accuracy levels. The multiple classifiers combined with a voting mechanism and the resolution level transition rule further help achieve a classification system with good classification accuracies (the best accuracy of 96.89%). The overall performance of the developed system shows the promising feasibility in helping pathologists classify the neuroblastoma images.

6. REFERENCES