# COMPUTER-AIDED GRADING OF NEUROBLASTIC DIFFERENTIATION: MULTI-RESOLUTION AND MULTI-CLASSIFIER APPROACH

*Jun Kong[1, 2], Olcay Sertel[1 ,2], Hiroyuki Shimada[3], Kim Boyer[1], Joel Saltz[2], Metin Gurcan[2]*

[1]Dept. of Electrical and Computer Engineering, The Ohio State University, Columbus OH
[2]Dept. of Biomedical Informatics, The Ohio State University, Columbus OH
[3]Dept. of Pathology and Laboratory Medicine, University of Southern California, Los Angeles, CA

## ABSTRACT

In this paper, the development of a computer-aided system for the classification of grade of neuroblastic differentiation is presented. This automated process is carried out within a multi-resolution framework that follows a coarse-to-fine strategy. Additionally, a novel segmentation approach using the Fisher-Rao criterion, embedded in the generic Expectation-Maximization algorithm, is employed. Multiple decisions from a classifier group are aggregated using a two-step classifier combiner that consists of a majority voting process and a weighted sum rule using priori classifier accuracies. The developed system, when tested on 14,616 image tiles, had the best overall accuracy of 96.89%. Furthermore, multi-resolution scheme combined with automated feature selection process resulted in 34% savings in computational costs on average when compared to a previously developed single-resolution system. Therefore, the performance of this system shows good promise for the computer-aided pathological assessment of the neuroblastic differentiation in clinical 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 has been on quantitatively measuring and analyzing digital slides for breast cancer [1], cervical cancer [2], colonic mucosa [3] and prostate cancer [4]. However, few methods, to our best knowledge, have been proposed for the computerized classification of neuroblastoma (NB), a cancer mostly occurring in children.

In clinical practice, the prognosis of neuroblastoma is 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 class-indicators. In terms of pathological characteristics established in this system, three grades are defined: undifferentiated (UD), poorly differentiated (PD) and differentiating (D). One typical tissue slide and example images associated with different differentiation grades are shown in Figure 1. In general, the undifferentiated cases contain small to middle-sized NB cells with thin cytoplasms, none-to-few neuroties 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).

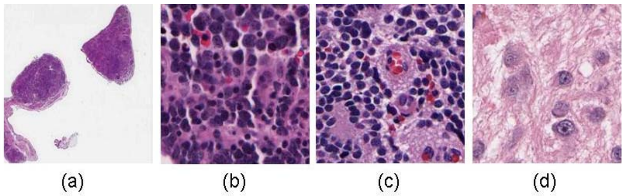

Figure 1: Typical tissue slide and the images associated with the three differentiation grades. (a) Typical tissue slide; Enlarged images for (b) undifferentiated (c) poorly-differentiated and (d) differentiated cases.

In our previous work [6-7], we have developed a novel 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 usually prohibitive. For instance, it takes 4166 seconds to process the slide shown in Figure 1(a) with classification accuracy of 96.46%. In these studies, the same set of manually selected features was used at each resolution. Furthermore, fewer classifiers were trained and tested, potentially excluding the contributions from other superior classifiers. To overcome these shortcomings, in this work, we employed a multi-resolution, multi-classifier approach with an automated feature selection process.

1-4244-1437-7/07/$20.00 ©2007 IEEE

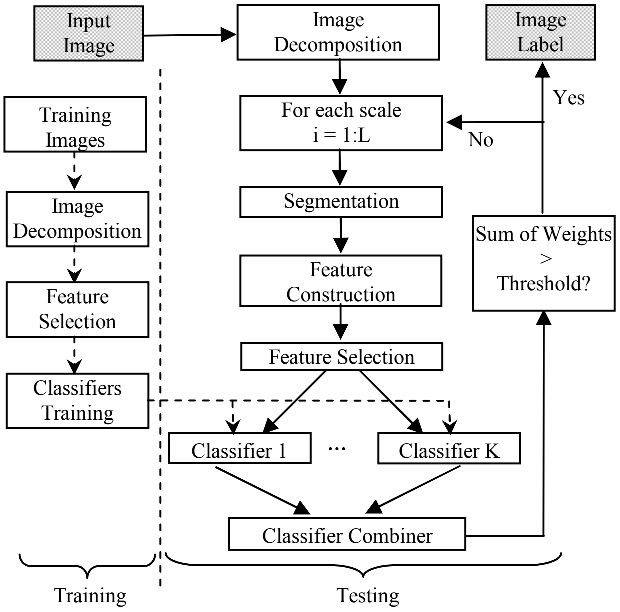

Figure 2: Flowchart of the developed classification system

## 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). 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

Due to the overwhelmingly large image size, each tumor slide is split into non-overlapping tiles of size $512 \times 512$ before they are processed one at a time by our grading system. Aiming at increasing the classification efficiency as much as possible, we employed a multi-resolution approach that decomposes every input image tile into multiple resolution 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 resolution images and stopping at the resolution level where the classification performance satisfies a pre-determined criterion.

As shown in Figure 2, multiple image analysis steps consisting of image segmentation, feature construction, feature selection and classification are followed at each resolution scale. As a way of improving the overall classification performance, seven combinations of feature extractors and classifiers are used, followed by a two-step classifier combiner that makes the final decision after aggregating information from the group of classifiers.

## 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 $I^{L-1}$ is created following the down-sampling process stated in [7].

### 3.2. Image Segmentation

In the computerized system, a new segmentation approach, EMLDA presented in [6], was applied to each image tile. In summary, 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:

$$J(V^* \mid \theta^*) = \underset{V,\theta}{Max} \frac{|V^T S_B(\theta) V|}{|V^T S_W(\theta) V|} \qquad (1)$$

where $J(V \mid \theta)$ 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 feature subspace, 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 $J(V \mid \theta)$ 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 discriminative 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 the L, $A^*$, and $B^*$ image channels are extracted. As a result, the constructed feature vector consists of 24 elements.

Due to the "peaking phenomenon" [8], the choice of a subset of most discriminating features is conducive to improving the classification accuracy. Additionally, using fewer features also contributes to the decrease in the computational complexity. Therefore, in practice, we employed the Sequential Floating Forward Selection (SFFS) procedure [9], as it yields a strong flexibility of adding and removing features in a dynamical way.

### 3.4. Classification

In the experiments, a pool of seven 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 [8], are integrated into our system, with each one working independently.

In our experiments, no particular preference on the choice of the classifier across all resolution levels is observed. Each classifier contributes to the improvement of the classification accuracy in an approximately equal manner. Removing any one of the seven classifiers will result in an inferior recognition rate at least at one resolution level. This is due to the fact that each classifier, in our system, has its own feature regions where it yields the best performance, although their global performances look similar. Integrating multiple rather than a single classifier into our system, therefore, improves the resulting system performance.

### 3.5. Classifier Combiner

The resulting labels assigned by the multiple classifiers are combined using a two-step integration strategy before the final classification decision is made.

***Step 1.*** The combiner evaluates the outputs of all K classifiers (K=7 in this work) and produces a final decision $\theta^*$ that prefers to the decisions supported by the majority of the K classifiers:

$$\theta^* = Arg \underset{i \in \{1,2,\dots C\}}{Max} V(i) \qquad (2)$$

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

***Step 2.*** After the label is voted, we next evaluate whether or not the classification result at the current resolution level is sufficiently good. The confidence degree on the classification result is defined as the sum of weights assigned to classifiers that agree with the combiner, since the combination scheme using the sum rule usually outperforms the others [11]. Each classifier weight is obtained by normalizing the classification accuracies of the K classifiers over the training data using the leave-one-out validation process, thus, indicating the degree of confidence on each classifier. As a result, the hypothesis test and the resulting decision rule can be written as:

*$H_0$: classification result is good enough; quit the process;*
*$H_1$: go to the next higher resolution level for classification;*

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

$$where: \qquad S = \sum_{i=1}^{K} \tilde{w}^l(i)\delta_{i\theta^*} = \sum_{i=1}^{K} \frac{w^l(i)}{\sum_{j=1}^{K} w^l(j)} \delta_{i\theta^*} \qquad (3)$$

In (3), $w^l(i)$ is the priori recognition rate of the classifier $i$ at resolution level $l$ and $\delta_{ij}$ is the Kronecker delta function.

## 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. Dividing the image into non-overlapping image tiles, we can take full advantage of parallel computing with this cluster.

For training purposes, 129 image tiles are cropped at random from tumor slides of each grading class, hence a total of 387 image tiles for classifier training. The testing database used in our experiments comprises 10 studies with 3, 2 and 5 whole-slide tumor samples from UD, PD and D grades, individually. The resulting best overall classification rate is 92.75%±7.42%.

Classification results of a typical UD case, shown in Figure 1(a), with different threshold settings ( i.e. different confidence level configurations) are shown in Table 1, where, $\xi$ scaled by 0.1, in the first column represents the set of 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 identified 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.

| $\xi$ ( x 0.1 ) | T (sec) | Acc (%) | #L1 | #L2 | #L3 | #L4 | #Bg. |
|---|---|---|---|---|---|---|---|
| 7 8 9 | 1738 | 95.01 | 3797 | 134 | 36 | 80 | 10569 |
| 9 8 7 | 1749 | 95.85 | 2761 | 761 | 429 | 93 | 10572 |
| 8 8 8 | 1705 | 95.38 | 3405 | 368 | 167 | 105 | 10571 |
| 9 9 9 | 2591 | 96.66 | 2772 | 407 | 255 | 610 | 10572 |
| 10 10 10 | 2760 | 96.89 | 0 | 2182 | 797 | 1066 | 10571 |

Of all the 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 more computational 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%, on average, as compared to those of the previous system [7]. In Figure 3, the classification results presented in Table 1 are visualized using classification maps and level maps, i.e. images indicating the class label and the resolution level at which the final decision on each image tile is made. Each pixel shown in Figure 3 represents a 512× 512 image sub-region cropped from the original tumor slide consisting of 14,616 tiles. Color blue, cyan and yellow assigned to each pixel in images on the left column represent UD, PD and D classes respectively, while the background pixel is shown in white color. In each image on 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 full resolution level.

## 5. CONCLUSIONS

This study demonstrates an automated system that classifies neuroblastoma according to the grade of differentiation within a multi-resolution framework. Combined with this multi-resolution paradigm, an automated feature selection algorithm (SFFS) considerably reduces the computational cost (66% of that of a previously developed single-resolution system) while maintaining accuracy levels. A two-step strategy combining multiple classifiers further helps the system yield good classification accuracies (the best accuracy of 96.89%). As a result, the developed system shows a great promise in assisting pathologists to classify neuroblastoma images.

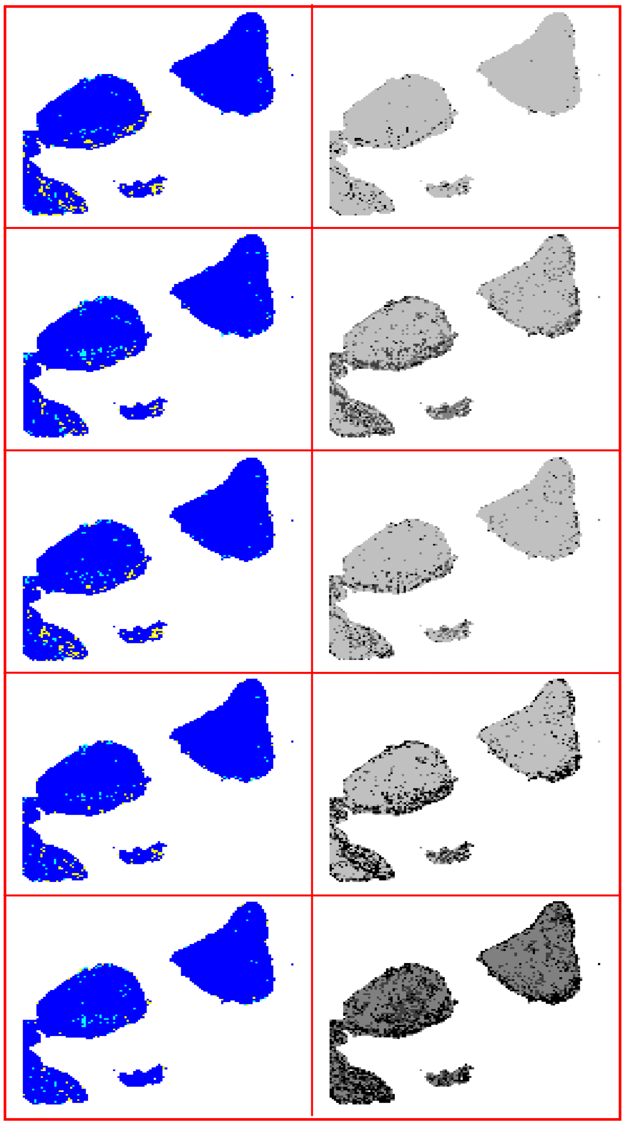

Figure 3: From top to bottom, images on the left column are the classification results using the multi-resolution approach when $\xi$ is configured as Table1; Images on the right columns are the corresponding resolution level maps.

## 6. REFERENCES

[1] H. Kobatake, Y. Yoshinaga, and M. Murakami, “Automatic detection of malignant tumors on mammogram,” *Proc. IEEE ICIP*, vol. 1, pp. 407-410, Austin, 1994.

[2] F. Hallouche, A.E. Adams, O.R. Hinton, G. Relf, M.S.Lakshmi, and G.V. sherbet, “Image processing for cell cycle analysis and discrimination in metastatic variant cell lines of B16 murine melanoma,” *Pathology*, vol.60, pp. 76-81, 1992.

[3] A.N. Esgiar, R.N.G. Naguib, B.S. Sharif, M.K. Bennett and A. Murray, “Microscopic image analysis for quantitative measurement and feature identification of normal of cancerous colonic mucosa,” *IEEE Trans. on Information Technology in Biomedicine*, vol. 2, no.3, pp. 197-203, 1998.

[4] S. Doyle, C. Rodriguez, A. Madabhushi, J. Tomaszeweski, and M. Feldman, “Detecting prostatic adenocarcinoma from digitized histology using a multi-scale hierarchical classification approach,” *Proc. IEEE, EMBS*, pp.4759-4762, New York City, 2006.

[5] H. Shimada et al., “The International Neuroblastoma Pathology Classification (the Shimada System),” *Cancer*, vol. 86, no. 2, pp.364-372, 1999.

[6] J. Kong, H. Shimada, K. Boyer, J. Saltz and M. Gurcan., “Image analysis for automated assessment of grade of neuroblastic differentiation,” *IEEE, International Symposium on Biomedical Imaging: Macro to Nano*, pp.61-64, Washington D.C., 2007.

[7] J. Kong, H. Shimada, K. Boyer, J. Saltz and M. Gurcan., “A new multi-resolution analysis framework for classifying grade of neuroblastic differentiation,” *Technical Report,* BMI-OSU, # *OSUBMI_TR_2007_n02,* 2007.

[8] K. Fukunaga, “Introduction to Statistical Pattern Recognition,” 2nd ed., New York, *Academic press*, 1990.

[9] P. Pudil, F.J. Ferri, J. Novovicova, and J. Kittler, “Floating search methods for feature selection with nonmonotonic criterion functions,” *CVIP*, vol.2, pp.279-283, Jerusalem, Israel, 1994.

[10] M. Zhu, A.M. Martinez, “Selecting principal components in a two-stage LDA algorithm,” *CVPR*, vol. 1, pp.132-137, NY, 2006.

[11] J. Kittler, M. Hatef, R P.W. Duin and J. Matas, “On combining classifiers,” *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol.20, No.3, pp.226-239, 1998.