Published in final edited form as:
Proc IEEE Int Symp Biomed Imaging. 2011 March 30; : 2128–2131. doi:10.1109/ISBI.2011.5872833.

A COMPREHENSIVE FRAMEWORK FOR CLASSIFICATION OF NUCLEI IN DIGITAL MICROSCOPY IMAGING: AN APPLICATION TO DIFFUSE GLIOMAS

Jun Kong[★], **Lee Cooper**[★], **Fusheng Wang**[★], **Candace Chisolm**[†], **Carlos Moreno**[★,†], **Tahsin Kurc**[★], **Patrick Widener**[★], **Daniel Brat**[★,†], and **Joel Saltz**[★]

[★]Center for Comprehensive Informatics, Emory University, Atlanta, GA 30322

[†]Department of Pathology and Laboratory Medicine, Emory University, Atlanta, GA 30322

Abstract

In this paper, we present a comprehensive framework to support classification of nuclei in digital microscopy images of diffuse gliomas. This system integrates multiple modules designed for convenient human annotations, standard-based data management, efficient data query and analysis. In our study, 2770 nuclei of six types are annotated by neuropathologists from 29 whole-slide images of glioma biopsies. After machine-based nuclei segmentation for whole-slide images, a set of features describing nuclear shape, texture and cytoplasmic staining is calculated to describe each nucleus. These features along with nuclear boundaries are represented by a standardized data model and saved in the spatial relational database in our framework. Features derived from nuclei classified by neuropathologists are retrieved from the database through efficient spatial queries and used to train distinct classifiers. The best average classification accuracy is 87.43% for 100 independent five-fold cross validations. This suggests that the derived nuclear and cytoplasmic features can achieve promising classification results for six nuclear classes commonly presented in gliomas. Our framework is generic, and can be easily adapted for other related applications.

Index Terms

Nuclei classification; feature selection; microscopy image analysis; metadata model; diffuse glioma

1. INTRODUCTION

Diffuse gliomas are the most frequent primary brain cancers in the central nervous system and characterized by widespread invasiveness and a strong inclination to biologic progression [1]. Although some important knowledge about this disease has been accumulated over recent years, there still is a long way to go before a good understanding of these neoplasms is well established. Among the numerous biologic structures believed to be important in glioma studies, such as microvasculature, necrosis, and pseudopalidsades, neoplastic nuclei in gliomas stand out as a key element shedding much light on the nature of this disease. As the spatial densities and distributions of different classes of nuclei carry complementary information used for glioma diagnosis and grading, classification of nuclei represents a critical step in categorizing tumors for prognosis, treatment, and research. Additionally, the classification results may both suggest and be correlated with genetic alterations and, as a result, can facilitate better categorization of diffuse gliomas, understanding progression of disease, and prediction of disease prognosis and treatment response.

Although previous studies on nuclei classification involve various methods for different applications, such as breast cancer, and renal cell carcinoma [2, 4, 3], no standard framework regarding how to represent, manage and utilize large-scale analysis results to facilitate further processing on microscopy images, e.g. nuclei classification, is discussed. This becomes a more and more important issue, as the complexity and scale of data keeps increasing. This issue is particularly important for nuclei analysis because of the large volume of information to be synthesized. As a result, we develop a mechanism for standardized representation and management of analytical data results. With this system, data can be efficiently retrieved with comprehensive query supports. Additionally, we also establish a comprehensive framework that integrates all the essential components consisting of: 1) large-scale feature computation; 2) collection of human annotations for training data; 3) query for features of selected nuclei for further analysis; and 4) training and testing of nuclei classifiers. In this paper, we present such an integrative workflow that supports the end goal of classification of nuclei in microscopy images of glioma tissues.

2. NUCLEI CLASSIFICATION MECHANISM

Our workflow for nuclei classification involves multiple components: image partition, nuclei segmentation and nuclear feature computation supported by parallel computation infrastructure, imaging meta-data representation, microscopy database, human input interface, query mechanism, classifier training and testing process. Figure 1 illustrates a panoramic view of the complete working schema. Detailed discussions on each of these modules are given as follows.

2.1. Parallel Processing of Images

Imaging datasets in our glioma study contain several hundreds of microscopy images and present a significant computational challenge for image analysis. Due to large sizes, data structures and intermediate results computed during whole slide image analysis may exceed available main memory on a machine. Moreover, processing a large image slide may take very long. For these reasons, we tile whole slide images into non-overlapping 4096 × 4096-pixel regions to permit parallel analysis. To scale up the analysis, we process images with the help of a large-scale, high-performance computation infrastructure where a cluster of computer nodes is used for executing jobs simultaneously. This infrastructure configuration currently consists of seven Dell 1950 1U rack mount units. Each unit is configured with Dual Xeon E5420 CPUs running with four cores at 2.5Ghz for a total of eight cores per node. Each node has one job slot per core, with 56 job slots in total.

2.2. Nuclear Feature Computation

Abundant information on cellular and tissue morphology can be derived from digitized pathology images. In diffuse gliomas, tumor cell nuclei are of significant interest to the scientific community [5]. However, capturing and analyzing this information in large-scale microscopy images presents a serious challenge with the subjective human reviewing process. Computerized image analysis, as an alternative, provides an opportunity for quantitative measurement of complex micro-anatomical features for biology entities in pathologic slides [6]. We have developed a suite of image analysis tools for segmenting and characterizing nuclei [7]. To identify nuclei in a reliable way, we apply mathematical morphology operations to images for normalizing background regions degraded by artifacts arising from tissue preparation and the scanning process. This operation substantially separates the foreground from the normalized background and allows recognition of nuclei by the straightforward thresholding method. Overlapped nuclei are subsequently separated using the watershed technique. A co mplementary set of features is derived from each identified nucleus. These features can be grouped into four broad categories: nuclear morphometry, region texture, intensity, and gradient statistics [7]. Since nuclear morphology is believed to be informative for distinguishing types of gliomas, morphometric features such as the degree of elongation, size, and regularity are included. Nuclear texture information is also captured with multiple descriptors, as there is a significant variation in texture across nuclei of distinct categories due to the clumping of chromatin. Features relevant to cytological intensity and intensity gradient are included in the feature set as well. Additionally, we apply the same set of texture and gradient features to “cytoplasm” regions surrounding nuclei and combine these features with nuclear features for better representation of nuclei of distinct types. Since the true cellular borders of glioma cells cannot be resolved on Hematoxylin and Eosin (H&E) stained images, “cytoplasm” refers to a fixed-distance radius surrounding a nucleus. Figure 2 presents glioma nuclei and “cytoplasm” regions from which features are derived. In aggregate, 74 features are computed to represent each nucleus.

2.3. Nuclei Classification

In our study, we use Quadratic Discriminant Analysis (QDA), because of its good performances, as a way to differentiate nuclei of distinct classes [10]. As we have collected a large population of nuclei for each type, we fit a normal distribution $\mathcal{N}(\mu_i, \Sigma_i)$ to each class of nuclei by the following step:

$$\begin{aligned}\sum\nolimits_i = \sum_{j=1}^{N_i} (x_{ij} - \mu_i)(x_{ij} - \mu_i)^T = (X_i - \widetilde{\mu_i})(X_i - \widetilde{\mu_i})^T \\ = R_i^T Q_i^T Q_i R_i = R_i^T R_i\end{aligned} \qquad (1)$$

where Σ_i, μ_i, and N_i are the estimated covariance matrix, mean and the number of samples of class i; $x_{ij} \in \Re^n$ is the j^{th} sample in class i; $X_i = [x_{i1}, \cdots, x_{iN_i}]$ and $\widetilde{\mu_i} = [\mu_i, \cdots, \mu_i]$; Additionally, by QR-decomposition, we have: $(X_i - \widetilde{\mu_i})^T = Q_i R_i$, where Q_i and R_i are unitary and upper triangular matrix, respectively. In our experiments, as we have $N_i > n$, $\forall i$, we truncate Q and R by taking the first n columns of Q and the first n rows of R.

When assuming the normal distribution for each class, the discriminant function for class i becomes:

$$\begin{aligned}D_i(x) = -\tfrac{1}{2}(x - \mu_i)^T \sum\nolimits_i^{-1} (x - \mu_i) - \tfrac{1}{2}\log\left|\sum\nolimits_i\right| + \log\pi_i \\ = -\tfrac{1}{2}(x - \mu_i)^T R_i^{-1} (R_i^T)^{-1} (x - \mu_i) - \tfrac{1}{2}\log|R_i|^2 + \log\pi_i \\ = -\tfrac{1}{2}\widetilde{x_i}^T \widetilde{x_i} - \log\prod_j \lambda_i^j + \log\pi_i \\ = -\tfrac{1}{2}\widetilde{x_i}^T \widetilde{x_i} - \sum_j \log\lambda_i^j + \log\pi_i\end{aligned} \qquad (2)$$

where π_i is the prior probability for class i; λ_i^j is the j^{th} eigen-value of R_i; and $\widetilde{x_i} = (R_i^T)^{-1}(x - \mu_i)$.

Recall that R_i is an upper triangular matrix, as a result, $\prod_j \lambda_i^j = \prod_j r_i^j$, where r_i^j is the j^{th} diagonal entry of R_i. Therefore, the discriminant function in (2) can be further simplified:

$$
\begin{aligned}
D_i(x) &= -\tfrac{1}{2}\widetilde{x_i}^T\widetilde{x_i} - \sum_j \log\lambda_i^j + \log\pi_i \\
&= -\tfrac{1}{2}\widetilde{x_i}^T\widetilde{x_i} - \sum_j \log r_i^j + \log\pi_i
\end{aligned}
\tag{3}
$$

Following the *Maximum A Posteriori* (MAP) rule [10], the classification mechanism becomes:

$$
\theta^*(x) = \arg\max_{i\in\{1,2,\dots,C\}} D_i(x) \tag{4}
$$

where C is the number of classes.

2.4. Data Representation and Management

Results generated from nuclei analysis contain rich information, including nuclear features, boundary coordinates, human classification, image metadata, patient information, and analytical provenance information. To provide a generic representation of data objects, data types, and their relationships, we develop *Pathology Analytical and Imaging Standards* (PAIS) [8, 9] for modeling and managing pathology imaging analysis results and tissue characterizations in our framework. The logical PAIS model consists of 62 classes and is realized in our framework as an XML schema. Analysis results are stored in XML documents (PAIS XML documents) conforming to this schema before they are transferred over the network to a PAIS database for storage and management. PAIS database provides relational tables and spatial tables to manage PAIS data. Our implementation is based on IBM Infosphere Warehouse Enterprise Edition 9.7.2 with DB2 and DB2 Spatial Extender as the database engine, running on a Dell PowerEdge T410 with CentOS 5.5. The spatial database engine manages markup boundaries as spatial data types and provides efficient spatial query support such as *contains* or *intersects* predicates. The database implementation has significant advantages. 1) Efficiency. Although the scale of data is large (about half million nuclei markups each whole-slide image on average), the relational database is very efficient on querying the data with its indexing methods and query optimization techniques. For example, it takes less than one second to count the nuclei for a whole-slide image; 2) Comprehensive query support, such as metadata queries and spatial queries, and correlation queries across different data; 3) Highly expressive power of queries. SQL query language with the spatial extended capabilities makes it most easy for humans to express queries in a declarative way. For instance, we use the spatial query shown in Figure 3 to identify nuclei boundaries based on human marked points (one internal point marked per nucleus) and retrieve corresponding nuclear features of 2770 training nuclei selected by human experts from the database. This query takes only about 10 seconds.

3. EXPERIMENTS AND RESULTS

To test the classification mechanism, we derive our training and testing data from 29 whole-slide microscopy images of samples covering a wide spectrum of glioma variations. A panel of neuropathologists selects 2770 nuclei of six classes via a third-party microscopy image viewer [11] by making one dot within the region of each nucleus of interest. Table 1 summarizes the six nuclei classes and the number of samples for each class included in our experiments, where c_1 =‘Neoplastic Astrocyte’ (NA), c_2 =‘Neoplastic Oligodendrocyte’ (NO), c_3 =‘Reactive Endothelial’ (RE), c_4 =‘Mitotic Figure’ (MF), c_5 =‘Reactive Astrocyte’ (NA), c_6 =‘Debris’ (DB). The coordinates of the dots and class labels from each slide are exported into an XML file from which a text file is generated and imported to the table *validation.nucleipoint* in the PAIS database. From PAIS database, we can submit a joint query combining multiple tables as demonstrated in Figure 3 and retrieve class labels and features of all nuclei selected by human experts.

As it is well known that features used for sample representation may not contribute to the end classification performance uniformly, it is helpful to identify the subset of features capturing the best discriminating information for recognition improvement [10]. As a result, we use the Sequential Floating Forward Selection (SFFS) procedure to narrow down the “discriminating” features prior to QDA. In Table 2, we present the mean confusion matrices without and with SFFS when QDA+MAP classification method is repeatedly applied to 100 independent five-fold cross validation tests. From Table 2, it is noticed that recognition accuracies for all classes but neoplastic astrocyte (c_1) are improved when SFFS is used. For class of neoplastic astrocyte, the performances with and without SFFS are comparable.

In order to investigate the best possible classification accuracy with this nuclei set, we also apply Gaussian Mixture Model [12] to our dataset and compare performance across different classification methods with nuclei, “cytoplasm”, and nuclei in conjunction with “cytoplasm” features. These results are shown in Table 3. The addition of features from “cytoplasm” regions significantly boosts recognition performance for both GMM and QDA classifier. Additionally, when combining SFFS, the performance is even more improved. When SFFS +QDA is used, the numbers of features with the use of nuclear, cytoplasmic, and combined features are 20, 22, and 34, respectively. The best overall accuracy reaches 87.43%. As a result, this system can be potentially used not only to facilitate clinical researchers in various cancer-related research, but also to inform pathologists of such critical information as spatial distributions and percentages of numerous types of nuclei for better diagnosis.

4. CONCLUSIONS

In this paper, we propose a comprehensive framework for classification of nuclei in whole-slide digital microscopy images, with a specific application to gliomas. We develop a generic and standardized approach for representing image analytical results, and an efficient database implementation with powerful query support. The framework provides an integrated workflow including imaging feature extraction with high performance computing infrastructure, convenient human annotation collection method, analysis result representation, spatial database query mechanisms and the final classification pipelines. When repeating the five-fold cross validation experiment for 100 times on 2770 nuclei of six types, we achieve an average accuracy of 87.43% with the SFFS+QDA+MAP classifier. These results demonstrate the efficacy of the integrative framework for analysis of nuclei in large-scale microscopy images.

Acknowledgments

This research is supported in part by Grant Number R01LM009239 from the NLM, by NCI Contract No. N01-CO-12400 and 94995NBS23 and HHSN261200800001E, and by NSF CNS 0615155, and 79077CBS10.

References

1. Brat DJ, Prayson RA, Ryken TC, Olson JJ. Diagnosis of malignant glioma: role of neuropathology. J Neurooncolology. 2008; 89(3):287–311.
2. Schnorrenberg F, Pattichis CS, Kyriacou K, Vassiliou M, Schizas CN. Computer-aided classification of breast cancer nuclei. Technol Health Care. 1996; 4(2):147–161. [PubMed: 8885093]
3. Schuefler, PJ.; Fuchs, TJ.; Ong, CS.; Roth, V.; Buhmann, JM. Computational tma analysis and cell nucleus classification of renal cell carcinoma. 32nd DAGM; 2010.

4. Boucheron LE, Manjunath BS, Harvey NR. Use of imperfectly segmented nuclei in the classification of histopathology images of breast cancer. IEEE ICASSP. 2010:666–669.

5. Gupta M, Djalilvand A, Brat DJ. Clarifying the difuse gliomas: an update on the morphologic features and markers that discriminate oligoden-droglioma from astrocytoma. Am J Clin Pathol. 2005; 124:755–768. [PubMed: 16203285]

6. Kong J, Cooper L, Sharma A, Kurc T, Brat D, Saltz J. Texture based image recognition in microscopy images of diffuse gliomas with multi-class gentle boosting mechanism. ICASSP. March.2010 :457–460.

7. Cooper L, Kong J, et al. An integrative approach for in silico glioma research. IEEE Transactions on Biomedical Engineering. October; 2010 57(10):2617–2621. [PubMed: 20656651]

8. Wang F, Kurc T, Widener P, Pan T, Kong J, Cooper L, et al. High-performance systems for in silico microscopy imaging studies. Data Integration in the Life Sciences. Lecture Notes in Computer Science. 2010; 6254/2010:3–18.

9. PAIS wiki. https://web.cci.emory.edu/confluence/display/PAIS

10. Duda, RO. Pattern classification and scene analysis. Wiley; New York, USA: 1973.

11. Aperio Inc. “http://www.aperio.com/”

12. McLachlan, G.; Peel, D. Finite mixture models. Hoboken, NJ: John Wiley & Sons, Inc; 2000.

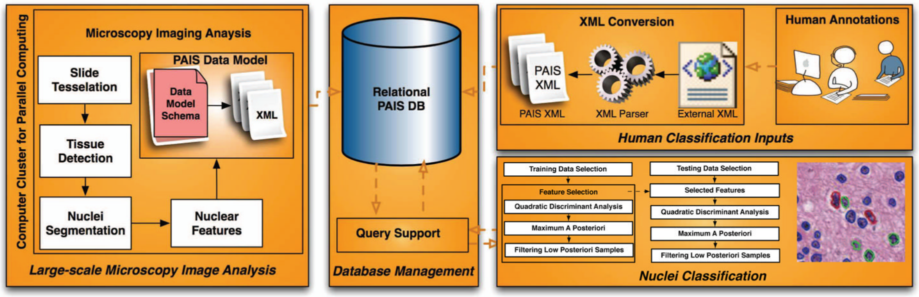


Fig. 1.

The overall framework to support classification of glioma nuclei from microscopy images

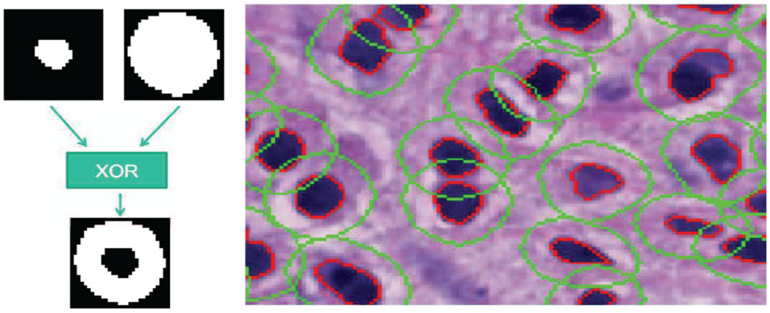
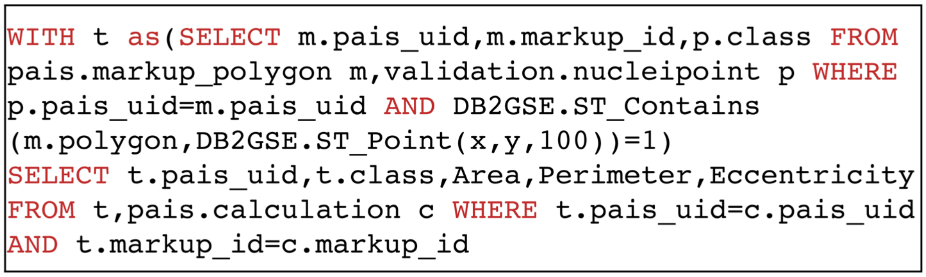


Fig. 2.

Features characterizing nuclei are derived from nuclei (red) and "cytoplasm" regions (green).



```
WITH t as(SELECT m.pais_uid,m.markup_id,p.class FROM
pais.markup_polygon m,validation.nucleipoint p WHERE
p.pais_uid=m.pais_uid AND DB2GSE.ST_Contains
(m.polygon,DB2GSE.ST_Point(x,y,100))=1)
SELECT t.pais_uid,t.class,Area,Perimeter,Eccentricity
FROM t,pais.calculation c WHERE t.pais_uid=c.pais_uid
AND t.markup_id=c.markup_id
```

Fig. 3.

An example SQL query to retrieve features of nuclei based on human marked points

Table 1

The number of nuclei samples for each of the six classes is shown.

Nuclei Class	**c_1 (NA)**	**c_2 (NO)**	**c_3 (RE)**	**c_4 (MF)**	**c_5 (NA)**	**c_6 (DB)**
Sample #	638	479	519	425	401	308

Table 2

The mean confusion matrices without and with SFFS (in %) are presented when QDA+MAP classification method is applied to 100 five-fold cross validation runs.

	c_1 (NA)	**c_2 (NO)**	**c_3 (RE)**	**c_4 (MF)**	**c_5 (RA)**	**c_6 (DB)**
c_1(NA)	**89.84, 89.72**	2.06, 1.75	5.78, 5.15	0.50, 0.79	0.59, 0.72	1.23, 1.87
c_2(NO)	4.02, 2.18	**90.42, 92.39**	3.12, 1.99	0.94, 1.34	0.15, 0.44	1.35, 1.66
c_3(RE)	14.33, 8.08	0.91, 0.71	**80.26, 85.26**	0.23, 0.19	0.41, 2.07	3.86, 3.69
c_4(MF)	4.53, 2.14	1.05, 0.79	2.27, 1.34	**86.42, 92.93**	2.62, 1.25	3.11, 1.55
c_5(RA)	11.13, 7.97	2.77, 2.89	19.95, 8.73	1.48, 1.78	**63.24, 77.43**	1.42, 1.20
c_6(DB)	10.00, 3.76	1.40, 1.35	11.65, 7.33	3.18, 2.24	0.12, 1.24	**73.65, 84.08**

Table 3

Performances (in %) associated with different features and classification mechanisms are compared.

	GMM		**QDA (“Nuc.”, “Cyto.”, “Nuc.+Cyto.”)**	
	“Nuc.”	**“Nuc.+Cyto.”**	**No SFFS**	**SFFS (20, 22, 34)**
c_1 (NA)	72.90	89.62	74.47, 88.46, **89.84**	73.82, 85.46, 89.72
c_2 (NO)	85.02	90.53	84.94, 77.23, 90.42	86.14, 79.48, **92.39**
c_3 (RE)	64.19	80.36	64.64, 66.57, 80.26	66.13, 72.84, **85.26**
c_4 (MF)	79.07	86.45	78.44, 77.07, 86.42	78.77, 84.63, **92.93**
c_5 (RA)	52.90	63.68	51.17, 54.40, 63.25	51.87, 66.24, **77.43**
c_6 (DB)	64.99	74.14	63.31, 60.53, 73.65	64.42, 66.07, **84.08**
Overall	70.53	82.08	70.44, 72.63, 81.97	71.04, 77.00, **87.43**