Image Classification Optimization of High Resolution Tissue Images

M. Kozlovszky*, K. Hegedűs, G. Windisch,
L. Kovács, and G. Pintér

Óbuda University, John von Neumann Faculty of Informatics, Budapest, Hungary,
kozlovszky.miklos@nik.uni-obuda.hu,
http://nik.uni-obuda.hu

Abstract. Generic image classification methods are not performing well on tissue images. Such software solutions are producing high number of false negative and positive results, which prevents their clinical usage. We have created the MorphCeck high resolution tissue image processing framework, which enables us to collect morphological and morphometrical parameter values of the examined tissues. Size of such tissue images can easily reach the order of 100 MB - 1 GB. Therefore, the image processing speed and effectiveness has an important factor. Our main goal is to accurately evaluate high resolution H-E (hematoxilin-eozin) stained colon tissue sample images, and based on the parameters classify the images into differentiated sets according to the structure and the surface manifestation of the tissues. We have interfaced our MorphCheck tissue image measurement software framework with the WND-CHARM general purpose image classifier and tried to classify high resolution tissue images with this combined software solution. The classification is by default initiated with a large training set and three main classes (healthy, adenoma, carcinoma), however the new image classification/process wall-clock time was intolerable high on single core PC. The processing time is depending on the size/resolution of the image and the size of the training set. Due to the tissue specific image parameters the classification effectiveness was promising. So we have started a development process to decrease the processing time and further increase the accuracy of the classification. We have developed a workflow based parallel version of the MorphCheck and WND-CHARM classifier software. In collaboration with the MTA SZTAKI Application Porting Centre the WND-CHARM has been ported to some distributed computing infrastructure (DCI). The paper introduces the steps that were taken to optimize WND-CHARM applications running faster using DCIs and some performance results of the tissue image classification process.

Keywords: Application porting, medical image processing workflow, HP-SEE, gUSE, scalability, MorphCheck, WND-CHARM

* Authors would like to thank Semmelweis University and Major & Co. to provide us annotated tissue samples for processing and classification. Authors would like to thank for the technical support of the WND-CHARM developer team as well.

1 Introduction

High resolution tissue image analysis and classification is a hot research topic nowadays. Since the most of the cases it means processing high-resolution images against large image databases/training sets that are really data-, and compute-intensive challenge, parallelization of this application is highly needed to decrease the computational time. This paper shows how we have created an image classification service using the Grid and cloud User Support Environment (gUSE [1]) to evaluate images processed by our Morphcheck (MC) software and evaluated it by "Weighted Neighbor Distance Using Compound Hierarchy of Algorithms Representing Morphology" (WND-CHARM [2]) software.

1.1 MorphCheck

MorphCheck (MC) is a high resolution tissue image analyzer framework, which processes high resolution digital tissue images. MorphCheck software framework is capable to effectively recognize -with its extendable algorithm repository- large number of differentiated tissue structures (such as surface epithelium, gland structures, lamina muscularis, submucosa etc.), and measure their morphological and morphometrical properties. The software supports both some vendor specific tissue scanner image formats and regular image standards (such as Tagged Image File Format /tiff/, Joint Photographic Experts Group /jpg/). It supports various colorization schemes like: HE (Hematoxilin-Eozin), DAB (3,3'-Diaminobenzidin), multi-color FISH. It contains various texture based algorithms, furthermore intensity and structure based algorithms (such as K-means, region growth, etc.)[4].

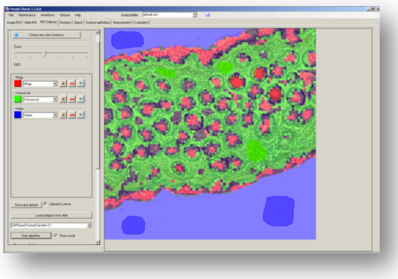

Fig. 1. MorphCheck tissue image analyzer framework GUI

1.2 WND-CHARM

WND-CHARM is an acclaimed open source image classifier application developed at National Institute of Aging (NIA, NIH)[5] that supports generic image

analysis methods. WND-CHARM extracts a large set of (normal: 1025, or extended: 2873) image features including polynomial decompositions, high contrast features, pixel statistics, and textures. These features are computed on: the raw image; transforms of the image; and second transforms of the image. The feature values are then used to classify test images into a set of pre-defined image classes. WND-CHARM is using WND-5 as image classifier. We are using recently the v1.30.227 as the baseline application for our MC-WND image classification service and for performance measurements as well. We have modified the original WND-CHARM software to include our MC tissue parameters. We have modified both the training and the classification part of the software.

1.3 gUSE

gUSE is basically a virtualization environment providing large set of high-level DCI (Distributed Computing Infrastructure, such as supercomputers, grids or clouds) services by which interoperation among classical service and desktop grids, clouds and clusters, unique web services and user communities can be achieved in a scalable way. gUSE has a graphical user interface, which is called WS-PGRADE. All part of gUSE is implemented as a set of Web services. gUSE supports various DCIs, and its execution concept is heavily based on workflows. The definition (graph, etc.) of workflows and their jobs are stored in a local storage. Job executions on DCIs requires user level authentication, and this can be managed transparently via the WS-PGRADE.

HP-SEE's Life Science Portal (Bioinformatics eScience Gateway) The Bioinformatics eScience Gateway is based on gUSE and operates within the Life Science VO of the HP-SEE[6] infrastructure. We have used this facility to implement our workflow and create the MC-WND tissue image classification service.

Fig. 2. HP-SEE Bioinformatics eScience Gateway

1.4 MC-WND tissue image classification service

The close collaboration between John von Neumann Faculty of Informatics, Obuda University and 2nd Department of Internal Medicine, Semmelweis University enabled us to define 75 tissue parameters of the human colonic region. Identified main tissue parameter groups:

- morphological and morphometrical properties of surface epithelium,
- morphological and morphometrical properties of gland structures,
- morphological and morphometrical properties of lamina muscularis,
- morphological and morphometrical properties submucosa
- morphometrical properties of the cells

We have extended our generic MorphCheck medical (tissue) image analysis framework to accurately measure these tissue image parameters. The objective numerical values of the pre-defined tissue parameters calculated by MorphCheck enables us to integrate and adapt a generic image classifier software solution, which can do effective tissue image classification automatically based on our parameter set. The two software solutions (MorphCheck and WND-CHARM) have been loosely coupled together to realize a single tissue image classification service (MC-WND). Data exchange between the two software solutions is realized with simple file exchange mechanisms. The MC-WND (Figure 3.) tissue image classification service allows researchers to process and categorize medical high resolutions tissue images using HPC infrastructure in a fast and easy way.

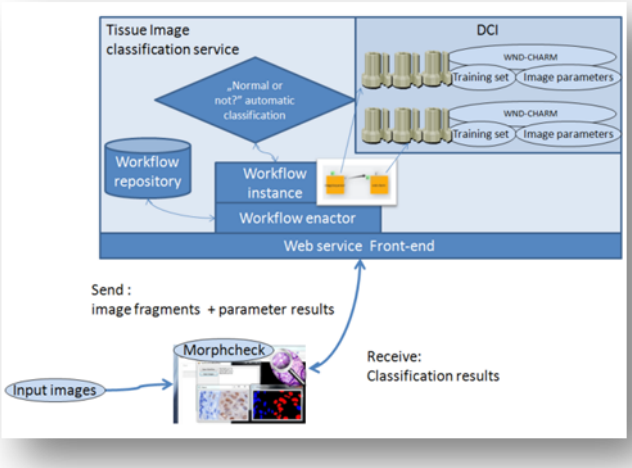

Fig. 3. MC-WND tissue image classification service schematic overview

1.5 MC-WND tissue image classification workflow

We have defined the image classification tasks within a single workflow. Inputs of the MC-WND workflow are:

- Fragment of the high resolution tissue image:
  - size: (512x512),
  - resolution/zoom level is the same which was defined during (automatic/manual) ROI definition/annotation,
  - in TIFF format.
- MC calculated parameter results exported into a csv file.

Output of the MC-WND workflow are:

- WND-CHARM calculated image parameter results stored in a csv file classification process results, decision with certainty values

1.6 MC-WND Training set

The modified WND-CHARM application is using our tissue image specific parameter as classifier parameters both during the training phase and during the image classification phase. The training phase is using the 2873 internal (WND-CHARM) parameters plus 75 tissue parameters measured by MorphCheck (MC). The results of the training phase are dumped into a single file (with WND-CHARM's internal coding format), and can be reused as an offline file for all the classification processes. Our tissue image training set contains more than 90 annotated HE (Hematoxilin-Eozin) colon tissue image samples with the following main categories: healthy, malignant (adenoma and carcinoma). All tissue image annotations are done by pathologist experts at 2nd Department of Internal Medicine, Semmelweis University. The training phase should be re-launched each time the annotated training image set or the parameter set extended. Luckily this is a rare event, because a single training phase lasts about 10 hours normally. In our recent implementation of the service generates a report in html format, which contains all the calculated statistical results of the classification process (accuracy, prediction, interpolation). We are using the stdout to monitor the process and receive status information.

1.7 Workflow implementation

As shown in Figure 4., the defined workflow consists of two consecutive jobs implemented in WS-PGRADE workflow language. The first job is a preprocessor, the second job utilize WND-CHARM in a parametric manner. The second job contains the WND-CHARM execution and it is launched in parallel as many times the service receives tissue images from outside. WND-CHARM is installed and launched in the so called user space, which was a hard task to realize. We are using LibTIFF [7] and FFTW [8] as external software packages inside our service. We are collecting the results from all the WND-CHARM instances both

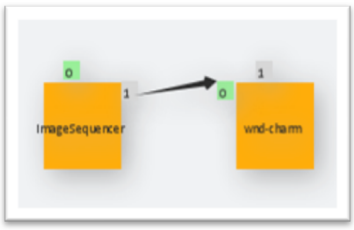

Fig. 4. MC-WND workflow in WS-PGrade/gUSE

from the stdout (as a file) and the generated html files. We have done performance evaluation to see how the WND-CHARM can run on HPC infrastructure. The workflow has been executed on one of the HPC centers operated by NIIF called "Budapest", which is an HP fat-node cluster using CP4000BL blades, consisting of 32 nodes with 24 Magny Cours CPU cores each (i.e. total number of CPU cores: 768). It has a mesh like topology with an Infiniband internal network. It has 1.96 TB memory and the total performance of the system is about 5,48 TFlops. Each measurement was executed 10 times, the average of the 10 executions was taken as the final result.

2 Performance Measurement Results

A single run of the image classification process is about 10 minutes for a 512x512 px tissue image size.

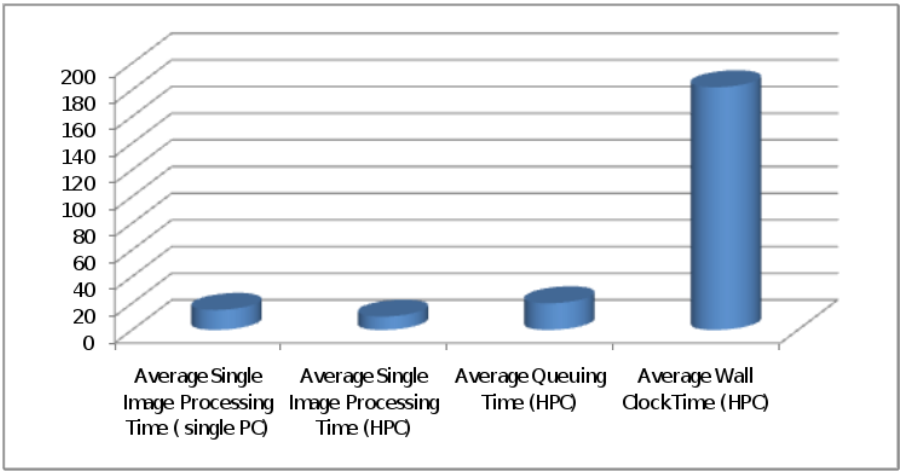

Fig. 5. Average MC-WND service execution time using HPC infrastructure (in min)

Nowadays a normal high resolution tissue image size (whole size) is about 4096x4096. Which is about 64 times larger than our 512x512 unit size. We have

launched 991 tissue image units against the HPC infrastructure. The following graphs show the result of the multiple executions of WND-CHARM on the HPC infrastructure. Figure 5. shows the processing time result. Average queuing overhead was 20 minutes. Average execution time was 10 minutes per image. The average total Wall Clock time was 3.02 hours for the whole image set (991 images), which means 11,503 sec was the average processing time for the whole image set. Figure 6. shows the WND-CHARM wall clock (execution) time compared to a single CPU. We gain significant speedup with parallel execution of the WND-CHARM, even if the average queuing time and result collection took us some negligible time.

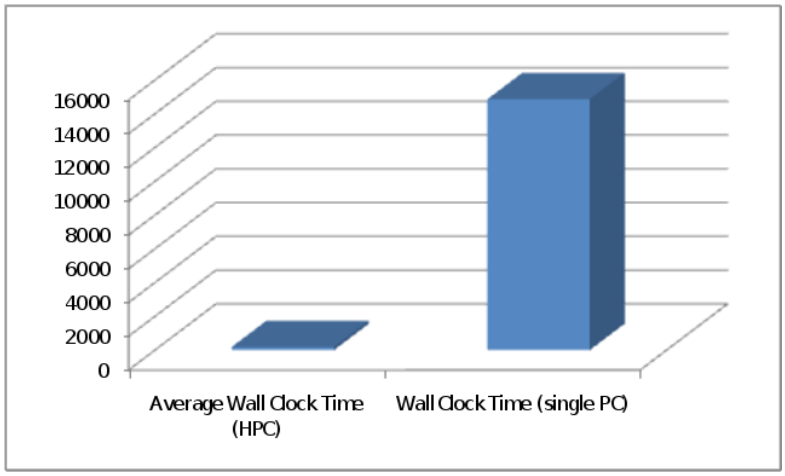

Fig. 6. MC-WND service execution time (HPC vs. single PC) (for 991 images in min)

2.1 Classification accuracy

To create a usable classification procedure for colonic tissue images based on automatic pre-filtering solution we have introduced a cut-off number which defines a tolerance level of the classification certainty. Below this certain numerical value the image marked automatically as malignant and forwarded to manual evaluation. In our large-scale tissue image classification tests with more than 200 tissue images and with the cut-off value below 60% we are able distinguish between healthy and malignant colonic tissue images with 100% accuracy at healthy category (that means healthy category strictly contains only healthy patient's images).

3 Conclusion

In this paper we have described how we have ported the modified WND-CHARM image classification software to work on distributed computing infrastructure (on

HP-SEE supercomputing infrastructure) as a service. The parameter field of the WND-CHARM application is extended with our high resolution tissue image analysis parameter field (+ 75 tissue image parameters). We have created the workflow structure for the MC-WND image classification service. The service is tested with HE stained tissue images and capable to separate healthy and malignant tissue images automatically with a high accuracy. The service and the internal workflow was developed at Obuda University and hosted on the HP-SEE Life Science/Bioinformatics eScience Gateway. The service can be used to do tissue image classification of the colonic region against our large tissue training databases in a short time using the HP-SEE supercomputing infrastructure at NIIF, Hungary. We also describe the performance analysis has been done on the applications. As a future work we are trying to include other image classifier solutions into our MC-WND service portfolio in a plug-in like manner. We are planning to open up our image classification service for the wider research community. So far the tissue image classification service can only be executed from the MorphCheck software. We are planning to create a portlet-based web user interface to let pathologists manually upload and evaluate tissue images.

Acknowledgments. The authors would like to thank two projects for their financial support. This work was supported by the HP-SEE (High-Performance Computing Infrastructure for South East Europe's Research Communities, under contract no. RI-261499) project and by the 3DHistech08 project, the Hungarian National Technology Programme, A1, Life sciences, the "Development of integrated virtual microscopy technologies and reagents for diagnosing, therapeutical prediction and preventive screening of colon cancer. Authors would like to thank Semmelweis University and Major & Co. to provide us annotated tissue samples for processing and classification. Authors would like to thank for the technical support of the WND-CHARM developer team as well.

References

1. P. Kacsuk,Z. Farkas,G. Sipos,G. Hermann,T. Kiss: Supporting Workflow-level PS Applications by the P-GRADE Grid portal,Towards Next Generation Grids Proceedings of the CoreGRID Symposium 2007
2. N. Orlov,L. Shamir,T. Macura,J. Johnston,D. Mark Eckley,Ilya G. Goldberg;WND-CHARM: Multi-purpose image classification using compound image transforms, Pattern Recognit Lett. 2008 January; 29(11): 1684–1693., PMCID: PMC2573471, NIHMSID: NIHMS53719
3. S. Szénási,Z. Vámossy,M. Kozlovszky;"Evaluation and comparison of cell nuclei detection algorithms", 16th International Conference on Intelligent Engineering Systems (INES), Lisbon, July. 2012, pp. 469 - 475, ISBN 978-1-4673-2694-0
4. `http://biotechweb.nik.uni-obuda.hu/web/en/research/projects/3dhist08/morphcheck2` [acc. 09.02.2013]
5. `http://www.grc.nia.nih.gov/branches/lg/iicbu/iicbu.htm` [acc. 09.02.2013]
6. `http://www.hp-see.eu` [acc. 09.02.2013]
7. `http://www.libtiff.org` [acc. 09.02.2013]
8. `http://www.fftw.org/download.html` [acc. 09.02.2013]