A FRAMEWORK FOR AGENT-BASED ANALYSIS OF BIG COLLECTIONS OF MEDICAL IMAGE DATA

V.A. Kovalev
Заведующий лабораторией анализа биомедицинских изображений ОИПИ НАН Беларуси, кандидат технических наук

Biomedical Image Analysis Department, United Institute of Informatics Problems National Academy of Sciences of Belarus, vassili.kovalev@gmail.com

This paper introduces an agent-based framework for the analysis of very large collections of medical image data. Towards this end, a scalable and highly flexible architecture of distributed image computing is suggested which utilizes the agent-based approach and some solutions borrowed from recent Big Data technology. The medical image analysis power stems from the idea of dynamic *N*-tuples of varying lengths and different ways of sampling of image elements. As of now, the framework remains on its early stage.

Introduction. Recently, there can be a strong trend observed towards exponential growth of the amount of natively-digital images used in medical diagnosis and treatment [1]. Along with the growth of the number of image items being stored in medical image archives, there is also a rapid progress in the increase of image resolution as well as clear tendency of shifting from the use of conventional 2D techniques to 3D tomography imaging. These two increase the volume of image data even further. For instance, one CT chest tomography image may consist of more than 300 separate 2D slices. Next, one single high-resolution histological RGB image acquired with the help of recent scanning optical microscopy system may easily achieve the resolution as high as 100 000 x 100 000 pixels what is amounted for 10 Gigapixels of image size. As a result, we may state that modern medical imaging field which is covering the medical image acquisition, store, retrieval, analysis, and the image mining domain already reached the entrance of Big Data door. Some examples of large collections of CT image data (e.g., an online service on content-based retrieval of 1.4 million axial CT slices) can be found on the experimental web site [2], which is expected to be incorporated into the tuberculosis portal [3] in future.

The conceptual problems of the analysis of Big Medical Data collections. We are capitalizing on the statement that in case the amount of medical image items, which are often associated with separate patients, is limited by hundreds or few thousands, there are no reasons to leave traditional technologies accumulated in medical image analysis. Such

a need comes into the way only in case of the analysis of hundred thousand and millions of entities. It is important to note that this is not just because of the volume of image data, which does not fit the memory of recent computers any more. There are two more actual reasons for that.

The first cause is the fact that we are not able to create a study group of images, which fulfils the existing statistical requirements in order to perform the image analysis task in a conventional manner (e.g., keep the age/gender balance, provide an equivalent representation of patients' sub-groups under comparison, involve the same number of patients and controls, etc.). Instead, we need to deal with all the data, which represent various groups of patients and diseases in the proportions, quantities and qualities that reflect their real incidence in real life and actual input of patients in hospitals.

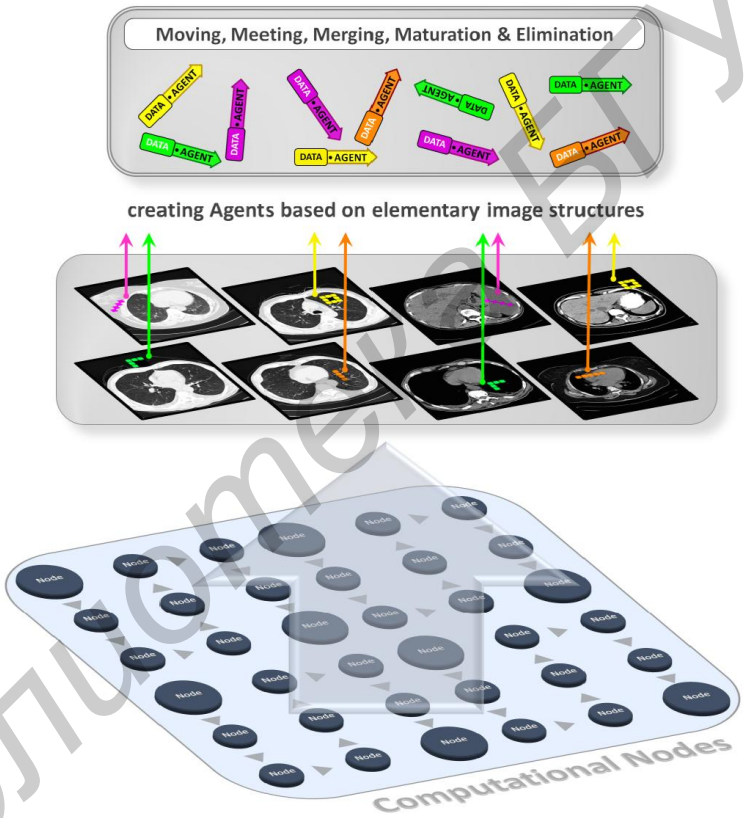

Fig. 1. The main components of the suggested framework
of agent-based analysis of big collections of medical image data.

The second reason is related to the problem of over-fitting. Indeed, dealing with the huge collection of medical data, we facing the problem of Gaussianing the quantitative image features of the groups of patients under study. As a result, we facing the problem of blurring and therefore missing (at least partially) the actually existing regularities as early as on the preliminary stage of assessing the significance of differences between the mean values of features characterizing two groups of patients with the help of Students' *t*-test. This in turn affects all the other subsequent analyses carried out by the existing sta-

tistical methods. Same holds true for pattern recognition methods, which can be applied afterwards.

Some illustrations of the appearance of image data blurring problem when approaching the threshold of tens of thousands images in image mining task can be found in [4]. Non-medical examples can be seen in [5].

Why the agent-based paradigm. We are capitalizing here on the agent-based approach by the two following reasons.

(a) The agent-based paradigm provides a natural way of scaling the extremely computationally-extensive methods which are used in the image analysis and image mining domain. In the same time, these methods may be relatively easily parallelized based on the natural splitting of image data on the wide range starting from parallel processing of separate image groups and going up to the parallel analysis of different sub-regions and pixel/voxel rows of one single image in case of algorithms of sliding-window type.

(b) Contrary to the conventional computational schemes, the agent-based computational mechanism with its characteristic evolutionary, stage-free, steady and uniform convergence towards the final solution provides possibilities of interruption/finishing the process of analysis of large image collection as soon as we already got the intermediate results and observe the clear trend of process convergence or divergence. Besides, this is also the way of overcoming the above mentioned problem of over-fitting, which is implemented by the hidden sub-sampling of image data due to the dynamic process of permanent creation of new and new agents.

How does this work. A simplified scheme of the framework which illustrates its major components provided in Fig. 1. A brief description of basic concepts and related notes on implementation are given below. For the purposes of brevity and clarity, an itemized style is preferred.

(a) It is naturally supposed that original image files and their raw data derivatives are spread over a multi-node computational environment supported by such basic software as Apache Hadoop [6] or by an original software solutions in case of the use of a simplified local research multi-core hardware platform.

(b) The major point of the suggested framework is as follows. Agents are created independently at random time points. Their types and the small portions of data they carry are based on randomly selected local neighborhoods of images situated in computational nodes. Image pixels for a new agent may be sampled using a set of different geometrical configurations. These configurations, which are illustrated in Fig. 1 by different colors, may include:

- pixels surrounding the given seed image point and located at certain distance from the seed, like it is typically happens with so-called Local Binary Patterns (LBP),

- simple configurations formed by a short segments of a straight lines passing over the randomly taken seed point,

- an arbitrary, multi-point configurations constituted by pixels of a curved line passing through maximal or minimal intensity values of local neighborhood defined by the seed point, etc.

Note that despite some basic geometric configurations (elementary image structures) are employed, there is always certain stochasticity in the pixel sampling process, which affects (distort) in an unpredictable manner the spatial 2D/3D locations of sampled pixels/voxels, their original intensity/color values as well as their derivatives such as the sign of intensity difference, the suitably binned gradient magnitude, the local gradient directionality or the related local anisotropy value. The agent's type is defined by the type of corresponding elementary image structure and/or the method of extracting derivatives from basic intensity/color values. Agents are also inheriting the image labels (e.g., "sick-healthy" class) and associated key characteristics (e.g., patient age). Thus, from the pattern recognition point of view, the data samples carried by agents are nothing but dynamic stochastic *N*-tuples [7] of varying lengths.

(c) After an agent is launched, it starts to move across the agents living area (see the top panel of Fig. 1). Once the agent meets another agent of the same type, they merge their data, create a child-agent, and die by themselves. In the simplest case the data merging procedure consists of summing up the values of counters of previous occurrences which are initialized by 1 at the time of agent creation. If an agent did not meet any of its potential "spouse-agent" for considerably long time, it dies from loneliness. It is important to note that along with discovering of image class features, the above agents' dynamics may ultimately lead to selection of elementary image structures (descriptors) that most suited for the images under exploration.

Acknowledgements. This work was partly funded by the National Institute of Allergy and Infectious Diseases, National Institutes of Health, U.S. Department of Health and Human Services, USA through the CRDF project OISE-14-60497-1.

References

1. Gao X., Müller H., Deserno T.M. Integration of Medical Images into the Digital Hospital, *The Open Medical Informatics J,* vol. 2011, No 5, pp. 17–18.

2. http://imlab.grid.by/ last visited 15 May 2015.

3. http://tuberculosis.by/getretroimages last visited 15 May 2015.

4ю Kovalev V., Prus A., Vankevich P. Mining lung shape from X-ray images. In: *International Conf. on Machine Learning and Data Mining (MLDM-2009),* Leipzig, Germany, P.Perner (Ed.), LNAI, vol. 5632, Springer Verlag, 2009, pp. 554–568.

5. Kovalev V.A. Mining dichromatic colours from video. In: *Advances in Data Mining. Applications in Medicine, Web Mining, Marketing, Image and Signal Mining: 6th Industrial Conference on Data Mining (ICDM-06),* Leipzig, Germany, July 14-15, P. Perner (Ed.): LNAI, vol. 4065, Springer Verlag, 2006, pp. 431–443.

https://hadoop.apache.org/ last visited 15 May 2015.

6. Kovalev V. and Petrou M. Multidimensional co-occurrence matrices for object recognition and matching, *Graphical Models and Image Processing,* vol. 58, No. 3, 1996, pp. 187–197.